HHS Public Access

Author manuscript

IEEE Trans Inf Theory. Author manuscript; available in PMC 2020 March 09.

Published in final edited form as:

IEEE Trans Inf Theory. 2008 January ; 54(1): 299–320. doi:10.1109/tit.2007.911296.

Achievable Rates for Pattern Recognition

M. Brandon Westover,

Department of Neurology, Massachusetts General Hospital, Boston, MA 02114-2622 USA

Joseph A. O'Sullivan [Fellow, IEEE]

Department of Electrical engineering, Washington University, St. Louis, MO 63130 USA

Abstract

Biological and machine pattern recognition systems face a common challenge: Given sensory data about an unknown pattern, classify the pattern by searching for the best match within a library of representations stored in memory. In many cases, the number of patterns to be discriminated and the richness of the raw data force recognition systems to internally represent memory and sensory information in a compressed format. However, these representations must preserve enough information to accommodate the variability and complexity of the environment, otherwise recognition will be unreliable. Thus, there is an intrinsic tradeoff between the amount of resources devoted to data representation and the complexity of the environment in which a recognition system may reliably operate.

In this paper, we describe a mathematical model for pattern recognition systems subject to resource constraints, and show how the aforementioned resource–complexity tradeoff can be characterized in terms of three rates related to the number of bits available for representing memory and sensory data, and the number of patterns populating a given statistical environment. We prove single-letter information-theoretic bounds governing the achievable rates, and investigate in detail two illustrative cases where the pattern data is either binary or Gaussian.

Index Terms—

Distributed source coding; multiterminal information theory; pattern recognition

I. Introduction

PATTERN recognition is the problem of inferring the state of an environment from incoming and previously stored data. In real-world operating environments, the volume of raw data available often exceeds a recognition system's resources for data storage and representation. Consequently, data stored in memory only partially summarizes the properties of patterns, and internal representations of incoming sensory data are likewise imperfect approximations. In other words, pattern recognition is frequently a problem of *inference from compressed data*. However, excessive compression precludes reliable

mb.westover@gmail.com.

Communicated by P. L. Bartlett, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

recognition. This apparent tradeoff raises a fundamental question: In a given environment, what are the least amounts of memory data and sensory data consistent with reliable pattern recognition?

The paper is organized as follows. In Section II, we introduce the general problem informally. Relationships between the present work and other pattern recognition research is briefly described in Section III. In Section IV, we formalize our problem as that of determining which combinations of three key rates are achievable, that is, determining which rate combinations allow the theoretical possibility of reliable pattern recognition. These rates quantify the information available for representing memory and sensory data, and the number of distinct patterns which the recognition system can discriminate. Our main results are single-letter formulas providing inner and outer bounds on the set of achievable rates, presented in Section V. In Section VI, we consider some instructive special cases of the main results, and compare our results to those for the related problem of distributed source coding. In Section VII, we explore explicit formulas for the bounds in two special binary and Gaussian cases. Section VIII contains concluding remarks. Proofs for most of the results are placed in the Appendices. The entire discussion is organized around the block diagram in Fig. 1.

II. Informal Problem Statement

In this section, we use an imagined example to motivate the mathematical model studied in the later technical sections. Suppose that our pattern recognition system consists of a homunculus living inside the head of some animal. The homunculus has access to a video monitor which displays data captured by the animal's retinas, and a set of index cards for storing information about the patterns in the environment relevant to survival, constituting a "memory." The homunculus must identify each pattern by comparing viewed images with information stored in memory. These identifications are then used to guide the animal's behavior. Let us consider which factors govern the difficulty of our homunculus' task.

A. Pattern Rate

First, the number of patterns that must be discriminated, M_c, obviously cannot exceed the number of images registerable on the animal's retinas, which depends in turn on the number of retinal photoreceptors and the number of distinct signaling states of each photoreceptor. Denoting the state of the retinas as $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$, where each Y_i takes values in a finite alphabet $\mathcal{Y}$, the number of possible retinal images is $|\mathcal{Y}|^n$. In a very simple animal with $n = 8$ photoreceptors, each able only to distinguish "bright" ($Y = 1$) from "dark" ($Y = 0$) (so $\mathcal{Y} \in \{0, 1\}$), the absolute upper limit on M_c would be $|\mathcal{Y}|^n = 2^8 = 256$.

With higher resolution eyes (larger n) an exponential explosion in the number of possible images rapidly overwhelms memory and computational resources. In humans, for whom $|\mathcal{Y}| \approx 256$, and $n \approx 2 \times 10^6$ [1], [2], $|\mathcal{Y}|^n \sim 10^{2,408,200}$, far exceeding estimates of the number of particles in the universe [3]. Fortunately, two features of real-world pattern recognition intervene: First, sensory data exhibits strong statistical structure, $p(\boldsymbol{y})$, so that the vast majority of the $|\mathcal{Y}|^n$ possible images are never experienced.[1] Second, much of the animal's

visual experience can be filtered out as irrelevant to survival. Thus, we express the number of patterns our homunculus must discriminate as $M_c = |\mathcal{Y}|^{nR'_c}$, where R'_c is called the *pattern rate*, $0 < R'_c < 1$, and generally $M_c \ll |\mathcal{Y}|^n$. Equivalently, we can express M_c in binary units, as $M_c = 2^{nR_c}$, in which case $R_c = R'_c \log_2|\mathcal{Y}|$.

B. Sensory Data Compression Rate

Our homunculus accesses sensory data indirectly through a video monitor that has limited display capacity. That is, whereas the retinas can be in up to $|\mathcal{Y}|^n$ distinct signaling states, the homunculus' internal monitor can display at most $M_y = 2^{nR_y} \ll |\mathcal{Y}|^n$, where R_y is thus the sensory data compression rate. Analogous data reductions arise in real recognition problems for various reasons, both computational (e.g., dimensionality reduction, sparsification, regularization, or other "feature extraction" operations), and economic (e.g., energy constraints, processing time constraints, storage limitations). We represent the transformation from retinal data $\boldsymbol{Y}$ to video data J by the action of an encoder ϕ, resulting in the displayed data $J = \phi(\boldsymbol{Y})$.

This sensory data compression step places another restriction on the number of discriminable patterns, so that, in general, $M_c \ll M_y \ll |\mathcal{Y}|^n$; or $R_c \leq R_y \leq \log_2|\mathcal{Y}|$.

C. Memory Data Compression Rate

The job of our homunculus is to recognize patterns. More formally, the homunculus must assign to each viewed image $\boldsymbol{Y}$ one of M_c class labels, which we take to be integers $\mathcal{M}_c = \{1, 2, \ldots, M_c\}$. As pre-job training, we imagine the homunculus studies a set of labeled class prototypes or "templates," $\boldsymbol{T}(w) = (\boldsymbol{X}(w), w)$, $w = \{1, 2, \ldots, M_c\} = \mathcal{M}_c$, each drawn from a distribution $p(\boldsymbol{x})$, where each template has dimensionality identical to that of the sensory data, $\boldsymbol{X}(w) = (X_1(w), \ldots, X_n(w))$.

The homunculus creates index cards, on which it writes the class labels and descriptive information about each class template. However, the number of cards and the amount of information per card are limited, allowing only a compressed summary of the available data. We represent the information memorized about a class $M(w)$ as the output of an encoder f, i.e., $M(w) = (I(w), w) = f(\boldsymbol{T}(w))$, where $I(w)$ is the compressed description of $\boldsymbol{X}(w)$ and w is the memorized class label. The degree of compression is quantified by specifying either the number of index cards comprising the memory, M_x, or by a compression rate (given in bits) $R_x = \frac{1}{n}\log M_x$.

As above, memory data compression restricts the number of discriminable patterns M_c, so that $M_c \ll M_x \ll |\mathcal{X}|^n$; or, in terms of rates, $R_c \leq R_x \leq \log_2|\mathcal{X}|$.

[1] If this were not the case, visual experience would be like watching television white noise.

D. Image Formation and Testing

The "testing" phase of our homunculus-driven pattern recognition system involves two processes:, *image formation* and *recognition*.

Image formation proceeds as follows. Nature selects a pattern class $W \in \mathcal{M}_c$ at random, then generates an image $\boldsymbol{Y}$ which is registered on the animal's retinas. (The class label W is not observable by the homunculus.) We model the image formation process as the transmission of the class template $\boldsymbol{X}(w)$ through a random channel $p(\boldsymbol{y}|\boldsymbol{x})$. The retinal image $\boldsymbol{Y}$ thus represents a "*signature*" of the underlying pattern $\boldsymbol{X}(w)$, and the channel $p(\boldsymbol{y}|\boldsymbol{x})$ represents two types of difficulties intrinsic to most real-world pattern recognition problems: *signature variation* (differences in the sensory data generated on repeated viewings of the same underlying pattern); and *signature ambiguities* (distinct patterns may produce similar signatures).[2]

The homunculus receives the compressed sensory data $J = \phi(\boldsymbol{Y})$, compares it with the memory data $\mathcal{C}_u = \{M(1), \ldots, M(M_c)\}$, and finally reports the class label of the best match, $\hat{W}$. The inference procedure used to make these comparisons can reflect knowledge of the pattern source $p(\boldsymbol{x})$ and image formation process $p(\boldsymbol{y}|\boldsymbol{x})$, but at the time of testing, it must be specified so as to depend only on the available data, i.e., g must be function only of $\mathcal{C}_u$ and J, $\widehat{W} = g(J, \mathcal{C}_u)$. We judge the homunculus' performance by the probability of error P_e. We will consider the system reliable if for some acceptable ϵ it achieves $P_e \leq \epsilon$.

E. Interpretations of the Problem Formulation

We have now introduced the basic elements of our problem, which is to determine the rate combinations (R_x, R_y, R_c) compatible with the possibility of reliable pattern recognition systems, where a "reliable" system is one for which the probability of recognition error can be made arbitrarily small. To summarize, these basic elements are 1) a model for the underlying patterns, consisting of the number of patterns $M_c = 2^{nR_c}$, a set of class labels $w \in \{1, \ldots, M_c\}$, and the class prototypes $\boldsymbol{X}(w)$ together with their generative model $p(\mathbf{x})$; 2) a model of the channel connecting class prototypes to the sensory data $p(\boldsymbol{y}|\boldsymbol{x})$; and 3) budgets specifying the number of bits allowed for representing sensory data R_y and memory data R_x inside the system.

We pause here to consider a few different possible perspectives on the problem under study.

Optimization views.—From an optimization point of view, we can ask our central question in two different but equivalent ways: Given the pattern rate R_c, what are the least amounts of sensory and memory data R_y and R_x, needed for reliable pattern recognition? Alternatively, given fixed information budgets for memory and sensory data representation R_x and R_y, what is the maximum achievable pattern recognition rate R_c?

[2]Grenander [4] and Mumford [5] have argued that four "universal transformations" (*noise and blur, superposition, domain warping*, and *interruptions*) account for most of the ambiguity and variability in naturally occurring signals.

Regarding "n."—Second, the problem has a different "feel" depending on whether one views the data dimensionality n either as a fixed or an increasing parameter. In the preceding discussion, we have primarily taken the static view, in which there are a fixed number $M_c = 2^{nR_c}$ of patterns, or "states of nature" of interest, and the problem is to investigate how many memory states M_x and sensory states M_y are needed to recognize them reliably. Alternatively, we may regard n as a dynamic, increasing parameter. Biologically, allowing n to increase might correspond to studying a series of animals with increasingly better eyes and memory organs. In engineering applications, the increase might correspond to building a sequence of machines with progressively higher camera resolution and data storage capacities [6]. Obviously, if while increasing n we hold the bit-budgets R_x and R_y fixed, then the number of memory and sensory states available for data representation grows exponentially, $M_x = 2^{nR_x}$, $M_y = 2^{nR_y}$. Less obviously, the maximum number of discriminable patterns also grows exponentially,[3] with a constant rate R_c, i.e., $M_c = 2^{nR_c}$. The "fixed n" and "increasing n" perspectives correspond to the familiar, complementary mathematical methods of proving a given inequality, respectively, either by the "adversarial" approach (given any ϵ, choose n large enough…); or the "asymptotic approach" (take the limit as $n \to \infty$…).

An important final point regarding n is that, like many results in information theory, our results rely on asymptotic arguments. Thus, we only prove the results valid only for "sufficiently large n," depending in turn on an ϵ corresponding to the tolerable error rate. The needed magnitude of n for a given ϵ (i.e., the issue of error exponents) will depend on the application, and is an important open problem.

III. Related Work

A. Machine Learning Approaches

Pattern recognition is a central topic in machine learning [7]–[10]. The machine learning approach to pattern recognition centers around the following problem: Given a set of labeled sensory data $\mathcal{D} = \{\boldsymbol{Y}(i), w(i), i = 1 \ldots N\}$, we wish to find a rule g that predicts the labels for future sensory data, i.e., if $\boldsymbol{Y}$ is in fact a signature of pattern class $w \in \{1, 2, \ldots, M_c\}$, we want $\Pr(g(\boldsymbol{Y}) \neq W) \leq \epsilon$ for some acceptable $\epsilon > 0$. Broadly speaking, two competing approaches dominate the literature. In the "generative modeling" approach, one attempts to estimate the distribution underlying the data $p(w, \boldsymbol{y})$, and then to use the conditional distribution $p(w, \boldsymbol{y})$ to infer w from $\boldsymbol{Y}$, i.e., $\widehat{w} = g(\boldsymbol{Y}) = \arg\max_{w \in M_c} p(w \mid \boldsymbol{Y})$. Alternatively, in the "discriminative" approach, one attempts to learn the optimal decision region boundaries directly, *without* estimating $p(w, \boldsymbol{y})$.

Our problem formulation resonates with the "generative modeling" approach, in that we allow the homunculus access to $p(w, \boldsymbol{y})$.[4] Informally, such knowledge might come from

[3]One should probably beware of the strange (and unnecessary) interpretation that, as we upgrade our camera (i.e., as we increase n), the number of patterns in the world consequently increases. More naturally, we may view the world as always presenting a practically unlimited number of patterns, while the number of patterns that can be taken advantage of by a system grows with increasing information processing resources.

allowing a very large volume of training data. Nevertheless, the distinction between generative and discriminative approaches then may become practically unimportant, as in many instances either approach can achieve asymptotically optimal performance.

In any case, in the present work we are not directly concerned with the problem of classifier *learning*. Rather, we investigate the conditions under which reliable classifiers can exist at all, regardless of how they are deigned or learned; we describe performance bounds to which *all* pattern recognition systems are subject.

It is also worth pointing out the distinction between the machine learning concept of "Vapnik–Chervonenkis (VC) dimension" and M_c in the present work. Informally, the VC dimension is the number of distinct patterns that can be shattered by a given family of classifiers (see [10], [12] for a detailed description). As such, VC dimension is a measure of the complexity of the decision boundaries that can be fit with a given family of classifiers. In contrast, M_c in our work is the number of patterns or pattern classes that can be distinguished, with no constraints on the family of classifiers.

B. Related Work in Combined Data Compression and Inference

Neuroscientist Horace Barlow has argued for more than four decades that data compression is an essential principle underlying learning and intelligent behavior in animal brains (see, e.g., [13]–[17]). Barlow and many others have amassed substantial experimental evidence showing efficient data coding mechanisms at work in the sensory systems of diverse animals, including monkeys, cats, frogs, crickets, and flies [18]. More recently, data compression is gaining appreciation as a mechanism for managing metabolic energy costs in neural systems [19].

In the engineering pattern recognition literature, data compression usually arises indirectly in the context of *feature extraction*, i.e., techniques for transforming raw data such that "irrelevant" data is discarded and the residual data is rendered into some advantageous format which facilitates storage and comparison, and is robust ("invariant") with respect to signature variations [20], [21]. In the information theory literature, probably the first direct investigation of the interplay between data compression and statistical inference is due to Ahlswede and Csiszár [22]. In [23] Han and Amari reviewed work up through 1998 on rate-constrained inference problems, including hypothesis testing, pattern recognition, and parameter estimation. Recently, Ishwar *et al.* have studied the problem of joint classification and reconstruction of sensory data subject to a fidelity constraint, in the context of video coding [24], [25]. In contrast to the problem studied in this paper, in that work there is no data compression constraint on memory data. Work on practical algorithms for joint classification and data compression includes [26]–[28].

IV. Problem Statement

We now proceed to the formal presentation of the main results.

[4]In particular, our formulation is consistent with the "General Pattern Theory" framework of Grenander and colleagues, which has provided a basis for much of the generative modeling work in pattern recognition research [11].

A. Notation

We adopt the following notational conventions. Random variables are denoted by capital letters (e.g., U), their values by lowercase letters (e.g., u), their alphabets by script capital letters (e.g., $\mathscr{U}$). Sequences of symbols are denoted either by boldface letters or with a superscript, e.g., $\boldsymbol{u} = u^n = (u_1, u_2, \ldots, u_n)$. The probability distribution for a random variable $U \in \mathscr{U}$ is denoted by $pU(u)$, or $p(u)$ simply when the implied subscript is clear from the context. Entropy, mutual information, and conditional mutual information are denoted in the usual ways, e.g., for random variables U, V, W, we write $H(U)$, $I(U; V)$, and $I(U; V|W)$, respectively. All logarithms are understood to be base two, i.e., $\log = \log_2$. Finally, to express statements such as "X and Z are conditionally independent given Y," i.e., $p(\mathrm{x}, \mathrm{y}, \mathrm{z}) = p(y)\, p(\mathrm{x} \mid \mathrm{y}) p(z \mid y)$, we write "$X - Y - Z$ form a Markov chain," or simply $X - . Y - \mathrm{Z}$.

B. Definitions and Assumptions

Definition 1: The *environment* $\mathscr{E}$ for a pattern recognition system is a set of eight objects

$$\mathscr{E} = \{\mathscr{M}_c, \mathscr{X}, \mathscr{Y}, p(\boldsymbol{x}), p(\boldsymbol{y}|\boldsymbol{x}), p(w), \mathscr{C}_x, \Phi\}$$

where

- $\mathscr{M}_c$, $\mathscr{X}$, $\mathscr{Y}$ are finite alphabets;
- $p(w)$, $p(\boldsymbol{x}\boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})$, are probability distributions over $\mathscr{M}_c$, and $\mathscr{X}^n \times \mathscr{Y}^n$, respectively;
- $\mathscr{C}_x = \{\boldsymbol{T}(1), \ldots \boldsymbol{T}(M_c)\}$ is a set of pairs $\boldsymbol{T}(w) = (\boldsymbol{X}(w), \mathrm{w})$ of random vectors $\boldsymbol{X}(w)$ drawn independent and identically distributed (i.i.d.) $\sim p(\boldsymbol{x})$, labeled by $w \in \{1, 2, \ldots, M_c\} = \mathscr{M}_c$;
- Φ is a mapping from labels $\mathscr{M}_c$ to vectors in $\mathscr{C}_x$, $\Phi : \mathscr{M}_c \to \mathscr{C}_x$, $\Phi(w) = \boldsymbol{X}(w)$.

We make the following simplifications:

- the distribution over class labels is uniform, $p(w) = 1/|\mathscr{M}_c|$ for all $w \in \mathscr{M}_c$;
- the pattern components are i.i.d., $p(\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i)$;
- the observation channel is memoryless, $p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} p(y_i|x_i)$.

Definition 2: An (M_c, M_x, M_y, n) pattern recognition code for an environment $\mathscr{E}$ consists of three sets of integers

$$\mathscr{M}_c = \{1 \ldots M_c\}, \mathscr{M}_x = \{1 \ldots M_x\}, \mathscr{M}_y = \{1 \ldots M_y\}$$

and three mappings

$$f: \mathcal{X}^n \times \mathcal{M}_c \to \mathcal{M}_x \times \mathcal{M}_c,\ f(\boldsymbol{t}(w)) = f(\boldsymbol{x}, w) = (i, w) \triangleq m(w)$$
$$\phi: \mathcal{Y}^n \to \mathcal{M}_y, \phi(\boldsymbol{y}) = j$$
$$g: \mathcal{M}_y \times (\mathcal{M}_x)^{M_c} \to \mathcal{M}_c, g(j, \mathscr{C}_u) = \widehat{w}$$

where $\mathscr{C}_u$ denotes the result of applying f to the entries of $\mathscr{C}_x$

$$\mathscr{C}_u = \{f(\boldsymbol{T}(1)), \ldots, f(\boldsymbol{T}(M_c))\} = \{m(1), \ldots, m(M_c))\,.$$

We call $\mathscr{C}_x$ the *pattern templates*; f, the *memory encoder*; $\mathscr{C}_u$, the *memorized data*; $\boldsymbol{\phi}$, the *sensory encoder*; and g, the *recognition function* or *classifier*.

Definition 3: The operation of a pattern recognition system ("agent") implementing a given (M_c, M_x, M_y, n) pattern recognition code $(f, \boldsymbol{\phi}, g)$ for an environment $\mathscr{E}$ is defined in terms of the following events.

Memorization phase:

- The agent observes $\mathscr{C}_x$, and uses f to compute the memory data $\mathscr{C}_u$.
- Access to $\mathscr{C}_x$ is taken away, and thereafter the agent knows of $\mathscr{C}_x$ only what is retained in $\mathscr{C}_u$.

Testing phase

- Nature selects an index $W \sim p(w)$.
- Nature encodes the pattern according to $\boldsymbol{X}(W) = \Phi(W)$.
- The pattern $\boldsymbol{X}(W)$passes through the channel $p(\boldsymbol{y}|\boldsymbol{x})$, giving rise to an observable signal $\boldsymbol{Y}$.
- The agent computes $J = \boldsymbol{\phi}(\boldsymbol{Y})$.
- The agent infers W by computing $\widehat{W} = g(J, \mathscr{C}_u)$.

With respect to the events just described, the probability of error for a code $(f, \boldsymbol{\phi}, g)$ in $\mathscr{E}$ is

$$P_e^n(w) = \Pr(\widehat{W} \neq w \,|\, W = w)$$

and the *average probability of error* of the code is

$$P_e^n = \sum_{w \in \mathcal{M}_c} p(w) \Pr(\widehat{W} \neq w \,|\, W = w)$$
$$= \frac{1}{M_c} \sum_{w \in \mathcal{M}_c} P_e^n(w)\,.$$

Definition 4: The rate $\boldsymbol{R} = (R_c, R_x, R_y)$ of an (M_c, M_x, M_y, n) code is

$$R_c = \frac{1}{n}\log_2 M_c, \quad R_x = \frac{1}{n}\log_2 M_x, \quad R_y = \frac{1}{n}\log_2 M_y$$

where the units are bits-per-symbol.

Definition 5: A rate $\boldsymbol{R} = (R_c, R_x, R_y)$ is *achievable* in a recognition environment $\mathscr{E}$ if for any $\epsilon > 0$ and for n sufficiently large, there exists an (M_c, M_x, M_y, n) code (f, ϕ, g) with rates

$$R'_c = \frac{1}{n}\log M_c, R'_x = \frac{1}{n}\log M_x, R'_y = \frac{1}{n}\log M_y$$

such that $R'_c \geq R_c$, $R'_x \leq R_x$, $R'_y \leq R_y$ and $P_e^n < \epsilon$.

Definition 6: The *achievable rate region* $\mathscr{R}$ for a recognition environment $\mathscr{E}$ is the set of all achievable rates $\boldsymbol{R} = (R_c, R_x, R_y)$.

Our ultimate goal in an information-theoretic analysis of this problem is to characterize the achievable rate region $\mathscr{R}$ in a way that does not involve the unbounded parameter n, that is, to exhibit a single-letter characterization of $\mathscr{R}$.

V. Main results

In this section, we present inner and outer bounds on the achievable rate region $\mathscr{R}$. The bounds are expressed in terms of sets of "auxiliary" random variable pairs UV, defined below. In these definitions, U and V are assumed to take values in finite alphabets $\mathscr{U}$ and $\mathscr{V}$ and have a well-defined joint distribution with the "given" random variables XY. To each such pair of auxiliary random variables UV we associate a set of rates

$$\mathscr{R}_{UV} = \left\{\boldsymbol{R}: R_x \geq I(U;X)\, R_y \geq I(V;Y)\, R_c \leq R_x + R_y - I(XY;UV)\right\}.$$

Next, define two sets of random variable pairs

$$\mathscr{P}_{\text{in}} = \{UV: U - X - Y,\ X - Y - V,\ U - (X,Y) - V\},$$

and

$$\mathscr{P}_{\text{out}} = \{UV: U - X - Y,\ X - Y - V\}.$$

We will also sometimes summarize the independence constraints in $\mathscr{P}_{\text{in}}$ as a single "long" Markov chain $U - X - Y - V$.

Next, define two additional sets of rates

$$\mathscr{R}_{\text{in}} = \{\boldsymbol{R}: \boldsymbol{R} \in \mathscr{R}_{UV} \text{ for some } UV \in \mathscr{P}_{\text{in}}\}$$
$$\mathscr{R}_{\text{out}} = \{\boldsymbol{R}: \boldsymbol{R} \in \mathscr{R}_{UV} \text{ for some } UV \in \mathscr{P}_{\text{out}}\}$$

and denote the convex hull of $\mathscr{R}_{\text{in}}$ by $\overline{\mathscr{R}}_{\text{in}}$.

Our main results are the following.

Theorem 1 (Inner Bound):

$$\mathscr{R}_{\text{in}} \subseteq \mathscr{R}.$$

That is, every rate $\boldsymbol{R} \in \mathscr{R}_{\text{in}}$ is achievable.

Theorem 2 (Better Inner Bound):

$$\overline{\mathscr{R}}_{\text{in}} \subseteq \mathscr{R}$$

That is, every rate $\boldsymbol{R} \in \overline{\mathscr{R}}_{\text{in}}$ is achievable.

Theorem 3 (Outer Bound):

$$\mathscr{R}_{\text{out}} \supseteq \mathscr{R}.$$

That is, no rate $\boldsymbol{R} \notin \mathscr{R}_{\text{out}}$ is achievable.

Finally, to ensure computability, we include a cardinality bound.

Theorem 4: Regions $\mathscr{R}_{\text{in}}$ and $\mathscr{R}_{\text{out}}$ are unchanged if we restrict the cardinality of $\mathcal{U}\mathcal{V}$ to

$$|\mathcal{U}\|\mathcal{V}| \leq |\mathcal{X}\|\mathcal{Y}| + 2.$$

Theorem 4 is a simple consequence of the Support Lemma [29 (p. 310)]: we must have $|\mathcal{X}\|\mathcal{Y}| - 1$ letters to ensure preservation of $p(xy|uv)$, and three additional letters to satisfy the constraints on $I(X; U)$, $I(Y; V)$ and $I(XY; UV)$.

Remark 1: If either $X = U$ or $Y = V$, or both, then the outer bound collapses to the inner bound, since in this case the extra Markov condition $U - (X, Y) - V$ in the definition of $\mathscr{P}_{\text{in}}$ is extraneous. For example, if $U = X$, then the condition is equivalent to $I(U; V|XY) = I(X; V|XY) = 0$, which is obviously true. Similar comments apply if U and V are any deterministic functions of X and Y, e.g., if $V = \gamma(Y)$, then $I(U; V|XY) = I(U; \gamma(Y)|XY) = 0$.

Remark 2: The bounds $\mathscr{R}_{\text{in}}$ and $\mathscr{R}_{\text{out}}$ can be expressed in various ways. For example, it is not difficult to show that the following replacements for $\mathscr{R}_{UV}$ lead to the same sets of rates $\mathscr{R}_{\text{in}}$ and $\mathscr{R}_{\text{out}}$:

$$\mathscr{R}'_{UV} = \left\{\boldsymbol{R}: R_x \geq I(U;X)\, R_y \geq I(V;Y)\, R_c \leq I(U;V) - I(U;V\,|\,XY)\right\} \quad (1)$$

$$\mathscr{R}''_{UV} = \left\{ \boldsymbol{R} : R_c \le I(U;V) - I(U;V|XY)\, R_x \ge I(XY;U|V) + R_c\, R_y \ge I(XY;V|U) + R_c\, R_x + R_y \ge I(XY;UV) + R_c \right\}. \quad (2)$$

That is, if we define

$$\mathscr{R}'_* = \left\{ \boldsymbol{R} : \boldsymbol{R} \in \mathscr{R}'_{UV} \text{ for some } UV \in \mathscr{P}_* \right\}$$
$$\mathscr{R}''_* = \left\{ \boldsymbol{R} : \boldsymbol{R} \in \mathscr{R}''_{UV} \text{ for some } UV \in \mathscr{P}_* \right\}$$

where * stands for either "in" or "out," then $\mathscr{R}_{\text{in}} = \mathscr{R}'_{\text{in}} = \mathscr{R}''_{\text{in}}$, and $\mathscr{R}_{\text{out}} = \mathscr{R}'_{\text{out}} = \mathscr{R}''_{\text{out}}$. These equivalencies are proved in Appendix E, and are used in Sections VI-B and VII.

Remark 3: In general, $\mathscr{R}_{\text{in}}$ is not a convex set, as evidenced by the examples studied in Section VIII. Thus, $\overline{\mathscr{R}}_{\text{in}}$ is in fact an improvement on $\mathscr{R}_{\text{in}}$. $\mathscr{R}_{\text{out}}$ is a convex set, as shown in Appendix C.

The proofs for Theorems 2 and 3 appear in Appendices A and B. Theorem 1 follows immediately from Theorem 2. In sketch-form, the method we use to prove achievability (the inner bound), based on $\mathscr{R}'UV$ (1), is as follows. We represent the memory and sensory data using codewords $\boldsymbol{U}(i)$, $i \in \left\{1 \ldots 2^{nR_x}\right\}$ and $\boldsymbol{V}(i)$, $j \in \left\{1 \ldots 2^{nR_y}\right\}$ that are typical according to $p(u)$ and $p(v)$, respectively, and the recognition system stores a list of these codewords. Making $R_x \ge I(X;\ U)$ and $R_y \ge I(Y;\ V)$ provides enough $\boldsymbol{U}$'s and $\boldsymbol{V}$'s to "cover" $\mathscr{X}^n$ and $\mathscr{Y}^n$. During pretesting, the system matches each of the labeled template patterns $(\boldsymbol{X}(w),\ w)$, $w = 1, \ldots, M_c$ presented to it with a unique memory codeword, and attaches to this codeword the corresponding class label (with matching defined in the sense of joint typicality according to $p(xu)$). The resulting set of M_c "active," labeled codewords constitutes the system's memory. During subsequent testing, suppose Nature selects class w, generating sensory data $\boldsymbol{Y} \sim p(\boldsymbol{y}|\boldsymbol{X}(w))$. The system receives the index J of the codeword for $\boldsymbol{Y}$, and uses it to retrieve the sensory codeword $\boldsymbol{V}(J)$. The system can then narrow down the list of M_c active memory codewords by a factor of $2^{-nI(U;\ V)}$ using knowledge of $p(uv)$.[5] Thus, the correct memory vector $\boldsymbol{U}(w)$ can be uniquely identified so long as $M_c = 2^{nR_c} \le 2^{nI(U;V)}$, i.e., if $R_c \le I(U;\ V)$.

It is also possible to prove the achievability result using a binning argument, which induces the set $\mathscr{R}''_{UV}$ (2): Generate $2^{nI(X;\ V)}$ $\boldsymbol{U}$'s and $2^{nI(Y;\ V)}\boldsymbol{V}$'s, and divide these equally among roughly 2^{nR_x} and 2^{nR_y} bins each, respectively. A pattern $\boldsymbol{X}(w)$ is encoded in memory by searching for a bin containing a matching (jointly typical) codeword $\boldsymbol{U}(\boldsymbol{X})$, and the $M_c = 2^{nR_c}$ bins thus selected are each assigned the class label w of the pattern stored therein. Sensory data $\boldsymbol{Y}$ is encoded as the bin index of a matching codeword $\boldsymbol{V}(\boldsymbol{Y})$. This number of $\boldsymbol{U}$'s and $\boldsymbol{V}$'s is sufficient to ensure that any given pair $\boldsymbol{X}$ and $\boldsymbol{Y}$ will have a matching (jointly typical) $\boldsymbol{U}(\boldsymbol{X})$, and $\boldsymbol{V}(\boldsymbol{Y})$, and the Markov lemma ensures joint typicality of the quadruple $(\boldsymbol{U}(\boldsymbol{X}), \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{V}(\boldsymbol{Y}))$. Given encoded sensory data $J = \phi(\boldsymbol{Y})$, recognition is done by comparing

[5] This step relies on the long Markov chain $U - X - Y - V$ and the Markov lemma.

the roughly $2^{nI(Y;V)}/2^{nR_y}$ sensory codewords in bin J with the $2^{nI(X;U)}/2^{nR_x}$ memory codewords in each of the 2^{nR_c} memory bins, then reporting the class label $\hat{w}$ assigned to the bin containing the matching memory codeword. No matches other than the correct one, $(\boldsymbol{U}(\boldsymbol{X}), \boldsymbol{V}(\boldsymbol{Y}))$, will be found provided the number of $(\boldsymbol{U}, \boldsymbol{V})$ comparisons grows exponentially with n at a rate less than $I(U; V)$, that is, provided $I(Y; V) - R_y + I(X; U) - R_x + R_c \leq I(U; V)$, which simplifies to the rate–sum constraint $R_x + R_y \geq I(XY; UV) + R_c$. The "side" constraints $R_x \geq I(XY; U|V) + R_c$ and $R_y \geq I(XY; V|U) + R_c$ then follow from requiring that each bin contain at least one codeword. The final inequality $R_c \leq I(U; V)$ follows from the first three.

VI. Discussion of the Main Results

A. The Gap Between Bounds

In general, there is a gap between $\mathscr{R}_{\text{in}}$ and $\mathscr{R}_{\text{out}}$, so that $\mathscr{R}_{\text{in}} \subsetneq \mathscr{R} \subsetneq \mathscr{R}_{\text{out}}$. This gap is due to the different constraints in the definitions of $\mathscr{P}_{\text{in}}$ and $\mathscr{P}_{\text{out}}$: Whereas distributions in satisfy three independence constraints $U - X - Y$, $X - Y - V$, and $U - (X, Y) - V$ (equivalently, the single "long chain" constraint $U - X - Y - V$), distributions in $\mathscr{P}_{\text{out}}$ only need satisfy the first two "short chain" constraints.

Further insight into the nature of the gap can be gained by attempting to *construct* $\mathscr{P}_{\text{out}}$ by combining distributions from $\mathscr{P}_{\text{in}}$ in various ways, and then considering whether the resulting distributions can be used to expand the achievable rate region.[6] We consider two such constructions. In both, let $Q \in \mathcal{Q}$ be a finite random variable, independent of X and Y. Holding $p(xy)$ fixed, to describe a pair of auxiliary random variables $U \in \mathcal{U}$, $V \in \mathcal{V}$ with joint distribution $p(xyuv)$, we need only to specify the marginal distribution $p(uv|xy)$. Consider the following two sets:

$$\mathscr{P}_{\text{mix}} = \left\{ UV : p(uv|xy) = \sum_{q \in \mathcal{Q}} p(q)p(u|xq)p(v|yq) \right\} \tag{3}$$

$$\mathscr{P}_{\text{conv}} = \left\{ UV : U = (U_Q, Q), V = (V_Q, Q), U_q V_q \in \mathscr{P}_{\text{in}} \ \forall q \in \mathcal{Q} \right\}. \tag{4}$$

In words, $\mathscr{P}_{\text{mix}}$ is the set of UV whose distributions $p(uv|xy)$ can be constructed as "mixtures" of product marginals; and $\mathscr{P}_{\text{conv}}$ is the set of "convexifying" random variables (this terminology is explained below).[7] In both of these sets it is possible to have dependencies between U and V given XY; i.e., in general $p(uv|xy) \neq p(u|x)\, p(v|y)$, hence, in general, $\mathscr{P}_{\text{mix}}, \mathscr{P}_{\text{conv}} \supsetneq \mathscr{P}_{\text{in}}$.

There is a gap similar to the one under discussion between the best known bounds for the distributed source coding problem (DSC), established by Berger and Tung (see Section VI-

[6] Alternatively, one can search for ways to tighten the outer bound.

[7] The random variables in these sets behave differently in mutual information computations. For example, compare $I(XY; UV)$ for UV in either set, using the same set of distributions $p(u|xq)$, $p(v|yq)$ as ingredients. For $UV \in \mathscr{P}_{\text{mix}}$

B). In both problems, the bounds are given in terms of sets with independence (Markov) constraints identical to those in $\mathscr{P}_{\text{in}}$ and $\mathscr{P}_{\text{out}}$.[8] Berger has suggested that in the DSC problem the gap is due to the fact that $\mathscr{P}_{\text{out}}$ admits convex mixtures of product marginal distributions, whereas $\mathscr{P}_{\text{in}}$ does not; i.e., in our notation, $\mathscr{P}_{\text{in}} \subsetneq \mathscr{P}_{\text{mix}} \subseteq \mathscr{P}_{\text{out}}$ [31]. The inclusion $\mathscr{P}_{\text{mix}} \subseteq \mathscr{P}_{\text{out}}$ is verified by checking $U - X - Y$ and $X - Y - V$: Write

$$\begin{aligned} p(u|xy) &= \sum_v p(uv|xy) \\ &= \sum_q p(q)p(u|xq)\sum_v p(v|yq) \\ &= \sum_q p(q)p(u|xq) \\ &\triangleq p(u|x) \end{aligned}$$

hence, $U - X - Y$; and a symmetric calculation shows $X - Y - V$. While clearly $\mathscr{P}_{\text{mix}}$ is a larger set than $\mathscr{P}_{\text{in}}$, it is unclear whether the admission of mixtures can account for all of the gap between $\mathscr{P}_{\text{in}}$ and $\mathscr{P}_{\text{out}}$. That is, we know of no proof that $\mathscr{P}_{\text{mix}} = \mathscr{P}_{\text{out}}$. Moreover, we know of no way to use auxiliary random variables from $\mathscr{P}_{\text{mix}}$ in achievability arguments.

It is also straightforward to verify that the second set $\mathscr{P}_{\text{conv}}$ is contained in $\mathscr{P}_{\text{out}}$

$$\begin{aligned} I(U;Y|X) &= I(U_Q, Q; Y|X) \\ &= I(Q;Y|X) + I(U_Q;Y|X,Q) \\ &\overset{(a)}{=} 0 + \sum_q p(q) I(U_q;Y|X,Q=q) \\ &\overset{(b)}{=} 0 \end{aligned}$$

(where the reasons are: (a) Q is independent of X and Y, and (b) $U_q V_q \in \mathscr{P}_{\text{in}}$; hence, $U - X - Y$; and a symmetric calculation shows $X - Y - V$. $\mathscr{P}_{\text{conv}}$ has a form sometimes introduced in time-sharing arguments, as a means to convexify a given rate region. For example, for $UV \in \mathscr{P}_{\text{conv}}$, we have

$$I_{\text{mix}} = I(XY;UV) = \sum_{xyuv} p(xy)\left(\sum_q p(q)p(u|xq)p(v|yq)\right) \times \log \frac{\left(\sum_q p(q)p(u|xq)p(v|yq)\right)}{\left(\sum_q p(q)p(u|q)p(v|q)\right)}$$

whereas for $UV \in \mathscr{P}_{\text{conv}}$

$$\begin{aligned} I_{\text{conv}} &= I(XY;UV) \\ &= \sum_q p(q) \sum_{xyuv} p(xy)p(u|xq)p(v|yq)\log\frac{p(u|xq)p(v|yq)}{p(u|q)p(v|q)}. \end{aligned}$$

It follows from the log–sum inequality [30, p. 29] that $I_{\text{conv}} \geq I_{\text{mix}}$

[8] The notation is ours.

$$\begin{aligned} I(XY;UV) &= I\left(XY;U_Q,V_Q,Q\right) \\ &= I(XY;Q) + I\left(XY;U_Q;V_Q\middle|Q\right) \\ &\overset{(a)}{=} \sum_q p(q) I\left(XY;U_qV_q\middle|Q=q\right) \\ &\triangleq \sum_q p(q) I\left(XY;U_qV_q\right) \end{aligned}$$

(where (a) is because Q is independent of X and Y) and similarly $I(X;\ U) = \Sigma_q p(q) I(X;\ U_q)$, and $I(Y;\ V) = \Sigma_q p(q) I(Y;\ V_q)$. It follows that the convex hull of $\mathscr{R}_{\text{in}}$ may be represented as

$$\overline{\mathscr{R}}_{\text{in}} = \left\{ \boldsymbol{R} : \boldsymbol{R} \in \mathscr{R}_{UV} \text{ for some } UV \in \mathscr{P}_{\text{conv}} \right\}. \tag{5}$$

In contrast to $\mathscr{P}_{\text{mix}}$, auxiliary variables from $\mathscr{P}_{\text{conv}}$ *can* be used as the basis for standard achievability arguments, as we have done in the proof of Theorem 2 (see Appendix A). From this, we have the following logical statement:

$$\begin{aligned} &\text{If } \mathscr{P}_{\text{out}} = \mathscr{P}_{\text{conv}} \\ &\text{then } \mathscr{R}_{\text{out}} = \overline{\mathscr{R}}_{\text{in}}. \end{aligned} \tag{6}$$

Unfortunately, we have no proof that $\mathscr{P}_{\text{out}} = \mathscr{P}_{\text{conv}}$. Notwithstanding, in Subsection VII-A, we examine one case where it appears that $\mathscr{R}_{\text{out}} = \overline{\mathscr{R}}_{\text{in}}$ does hold, giving grounds to conjecture that this equality may hold at least under special conditions.

B. Relationship With Distributed Source Coding

There are interesting connections between the results of Tung and Berger [32], [33] for the DSC problem and our results in Theorems 1 and 3. Briefly, the situation treated in the DSC problem is as follows (see Fig. 2). Two correlated sequences, $\boldsymbol{X}$ and $\boldsymbol{Y}$, are encoded separately as $i = f(\boldsymbol{X})$, $j = \phi(\boldsymbol{Y})$ and the decoder g must reproduce the original sequences subject to a fidelity constraint $\left(Ed_x(\widehat{\boldsymbol{X}}, \boldsymbol{X}), Ed_y(\widehat{\boldsymbol{Y}}, \boldsymbol{Y})\right) \leq \boldsymbol{D}$, where $\boldsymbol{D} = (D_x, D_y)$. The problem is to characterize, for any given distortion $\boldsymbol{D}$, the set of achievable rates $\widetilde{\mathscr{R}}(\boldsymbol{D})$.

The best known inner and outer bounds for the DSC problem can be expressed as follows.[9] Let $\mathscr{P}_{\text{in}}$ and $\mathscr{P}_{\text{out}}$ be defined as above, and define two new sets incorporating the distortion constraint

$$\begin{aligned} \mathscr{P}_{\text{in}}(\boldsymbol{D}) &= \mathscr{P}_{\text{in}} \cap \mathscr{P}_{UV}(\boldsymbol{D}) \\ \mathscr{P}_{\text{out}}(\boldsymbol{D}) &= \mathscr{P}_{\text{out}} \cap \mathscr{P}_{UV}(\boldsymbol{D}) \end{aligned}$$

where

$$\mathscr{P}_{UV}(\boldsymbol{D}) = \left\{ UV : \exists \hat{X}(U,V), \hat{Y}(U,V) \text{ s.t.} \left(Ed_x(\hat{X}, X), Ed_y(\hat{Y}, Y)\right) \leq \boldsymbol{D} \right\}.$$

[9] But see the footnote at the end Section VIII.

Parallelling (1), also define the sets of rates

$$\widetilde{\mathscr{R}}_{UV} = \left\{ \boldsymbol{R} : R_x \geq I(XY;U|V)\, R_y \geq I(XY;V|U)\, R_x + R_y \geq I(XY;UV) \right\} \tag{7}$$

and

$$\widetilde{\mathscr{R}}_{\text{in}}(\boldsymbol{D}) = \left\{ \boldsymbol{R} : \boldsymbol{R} \in \widetilde{\mathscr{R}}_{UV} \text{ for some } UV \in \mathscr{P}_{\text{in}}(\boldsymbol{D}) \right\}$$
$$\widetilde{\mathscr{R}}_{\text{out}}(\boldsymbol{D}) = \left\{ \boldsymbol{R} : \boldsymbol{R} \in \widetilde{\mathscr{R}}_{UV} \text{ for some } UV \in \mathscr{P}_{\text{out}}(\boldsymbol{D}) \right\}.$$

Then the Berger–Tung bounds for the DSC problem are $\widetilde{\mathscr{R}}_{\text{in}}(\boldsymbol{D}) \subseteq \widetilde{\mathscr{R}}(\boldsymbol{D})$ and $\widetilde{\mathscr{R}}_{\text{out}}(\boldsymbol{D}) \supseteq \widetilde{\mathscr{R}}(\boldsymbol{D})$.

There are strong formal similarities between our bounds and the DSC bounds. Most importantly, the gap between bounds for both problems is due to the difference between the length-four constraint $U - X - Y - V$ and the less stringent length-three constraints $U - X - Y$, $X - Y - V$. Further, note the formal similarity between the sets $\widetilde{\mathscr{R}}_{UV}$ (7) and $\mathscr{R}''_{UV}$. To carry this comparison further, suppose in the problem under study that, in addition to recognizing patterns, we also wish to reproduce an estimate of the original signals subject to a fidelity constraint, as in the DSC problem.[10] Denote the achievable rate region for this "joint recognition and recovery" problem by $\mathscr{R}(\boldsymbol{D})$. Making this addition in fact adds little technical difficulty, and the resulting bounds can be expressed, not surprisingly, as $\mathscr{R}_{\text{in}}(\boldsymbol{D}) \subseteq \mathscr{R}(\boldsymbol{D})$ and $\mathscr{R}_{\text{out}}(\boldsymbol{D}) \supseteq \mathscr{R}(\boldsymbol{D})$, where

$$\mathscr{R}_{\text{in}}(\boldsymbol{D}) = \left\{ \boldsymbol{R} : \boldsymbol{R} \in \mathscr{R}_{UV} \text{ for some } UV \in \mathscr{P}_{\text{in}}(\boldsymbol{D}) \right\}$$
$$\mathscr{R}_{\text{out}}(\boldsymbol{D}) = \left\{ \boldsymbol{R} : \boldsymbol{R} \in \mathscr{R}_{UV} \text{ for some } UV \in \mathscr{P}_{\text{out}}(\boldsymbol{D}) \right\}.$$

Apparently, the pattern recognition problem can be construed as a kind of generalization of the DSC problem, with the added complication that the "decoder" receives with $\boldsymbol{Y}$ not one sequence $\boldsymbol{X}$ but $M_c = 2^{nR_c}$ such sequences $\boldsymbol{X}(1),\ldots,\boldsymbol{X}(M_c)$ and must first determine which is the appropriate one with which to jointly decode $\boldsymbol{Y}$. This extra discrimination evidently requires that extra information be included at the encoders. This "rate excess" is the difference between the minimum encoding rates required for the DSC and pattern recognition problems.[11] Comparing $\mathscr{R}''_{UV}$ (7) with $\widetilde{\mathscr{R}}_{UV}$ (7), this rate excess is the same at both decoders, and is equal to R_c. Thus, R_c can be interpreted as the number of extra bits needed at both encoders to decide which of the possible $M_c = 2^{nR_c}$ patterns $\boldsymbol{X}(w)$ the sensory data $\boldsymbol{Y}$ represents, beyond the information required to simply reproduce the pair $\boldsymbol{Y}$, $\boldsymbol{X}(w)$ within the allowed distortion limits.

[10] This is related to the problem addressed in [24], [25], except that in that work there is no requirement that the memory data be compressed.

[11] Similar comments are made in [24].

C. Degenerate Cases

We now briefly examine the degenerate cases where either $X = U$, or $Y = V$, or both. In these cases, $I(U; V|XY) = 0$. Hence, using (1), we see that both inner and outer bounds on $\mathcal{R}$ both reduce to the three inequalities $R_X \geq I(U; X)$, $R_Y \geq I(V; Y)$, $R_C \leq I(U; V)\}$. Clearly, in these cases the bounds are tight, in that the inner and outer bounds are *equal*; there is no gap (see Remark 1). These degenerate cases have simple interpretations and are thus useful for building intuition about Theorems 1–3.

Sharp memory, sharp eyesight.—First, consider a system in which the budgets for memory and sensory representations are unrestricted, i.e., no compression is required. In this case, we can effectively treat the memories and sensory representations as veridical; i.e., we can set $U = X$ and $V = Y$. The theorem constraints then become $R_X \geq I(X; X) = H(X)$, $R_Y \geq I(Y; Y) = H(Y)$, and

$$R_c \leq I(U;V) = I(X;Y)\,. \tag{8}$$

This result indicates that, in the absence of compression, the recognition problem is formally equivalent to the following classical communication problem: Transmit one of $M_c = 2^{nR_c}$ possible messages (patterns) to a receiver (the recognition module) [6]. In this case, the patterns can be thought of as random codewords stored without compression and available to the decoder; Shannon's random coding for communication [30], [34] applies, yielding the mutual information $I(X; Y)$ (see (8)) as the bound on R_c.

Sharp memory, poor eyesight.—Next, suppose that memory is effectively unlimited, so that we can put $U = X$, but sensory data may be compressed. In this case, we can readily rewrite the condition on R_c as

$$R_c \leq I(X;Y) - I(X;Y|V) = I(X;V)\,. \tag{9}$$

We check the extreme cases: If V is fully informative about Y, $Y = \phi^{-1}(V)$, then $I(X; Y|V) = H(Y|V) - H(Y|X, V) = 0$, and we recover the case discussed above, $R_C \leq I(X; Y)$. For intermediate cases, where V is partially informative, the effect of V is to degrade the achievable performance of the system below that possible with "perfect senses," and the reduction incurred is $I(X; Y|V)$. In the extreme case that V is utterly uninformative (e.g., a constant $V = 0$, or otherwise independent of Y), $I(X; Y|V) = I(X; Y)$, and we get $R_c = 0$, or $M_c \leq 2^{nR_c} = 1$; hence, the system is useless.

Poor memory, sharp eyesight.—In the case of limited memory but unrestricted resources for sensory data representation ($V = Y$), we get an expression symmetric with the previous case

$$R_c \leq I(X;Y) - I(X;Y|U) = I(U;Y)\,. \tag{10}$$

As before, if the memory is perfect ($U = X$), we get $I(X;\ Y|\ U) = I(X;\ Y|\ X) = 0$, recovering the channel coding constraint $R_c \leq I(X;\ Y)$; assuming useless memories ($U = 0$) yields $R_c \leq I(X;\ Y) - I(X;\ Y) = 0$; and intermediate cases place the system between these extremes.

VII. Examples

In this section, we investigate the achievable rate regions for binary and Gaussian versions of our problem. For this purpose, it will be convenient to characterize the sets $\mathscr{R}$, $\mathscr{R}_{\text{in}}$ and $\mathscr{R}_{\text{out}}$ by their surfaces in the positive orthant $\mathbb{R}_+^3$. The surface of $\mathscr{R}$ can be expressed as

$$r(r_x, r_y) = \max_{R \in \mathscr{C}(r_x, r_y)} R_c$$
$$\mathscr{C}(r_x, r_y) = \left\{\boldsymbol{R} : \boldsymbol{R} \in \mathscr{R}, R_x = r_x, R_y = r_y\right\}.$$

Similarly, by direct extension of Theorems 1 and 3, and using the representation of $\mathscr{R}_{\text{in}}$ and $\mathscr{R}_{\text{out}}$ based on $\mathscr{R}'_{UV}$ (1), the surfaces of $\mathscr{R}_{\text{in}}$ and $\mathscr{R}_{\text{out}}$ are

$$r_{\text{in}}(r_x, r_y) = \max_{UV \in \mathscr{C}_{\text{in}}(r_x, r_y)} I(U;V) - I(U;V|XY) \tag{11}$$

$$r_{\text{out}}(r_x, r_y) = \max_{UV \in \mathscr{C}_{\text{out}}(r_x, r_y)} I(U;V) - I(U;V|XY) \tag{12}$$

where

$$\mathscr{C}_{\text{in}}(r_x, r_y) = \left\{UV \in \mathscr{P}_{\text{in}} : I(U;X) = r_x, I(V;Y) = r_y\right\}$$
$$\mathscr{C}_{\text{out}}(r_x, r_y) = \left\{UV \in \mathscr{P}_{\text{out}} : I(U;X) = r_x, I(V;Y) = r_y\right\}$$

The expression for the inner bound surface (11) reduces to

$$r_{\text{in}} = \max_{UV \in \mathscr{C}_{\text{in}}(r_x, r_y)} I(U;V).$$

An alternative expression for the outer bound surface which will be used in Subsection VII-B, based on $R_{UV}(1)$, is

$$r_{\text{in}} = r_x + r_y - \min_{UV \in \mathscr{C}_{\text{in}}(r_x, r_y)} I(XY;UV). \tag{13}$$

Finally, denote the convex hull of the inner and outer bound surfaces $\overline{r_{\text{in}}}(r_x, r_y)$, $\overline{r_{\text{out}}}(r_x, r_y)$.

In the specific cases studied in the following examples we seek to convert these implicit characterizations into explicit formulas not involving the optimization over $\mathscr{C}_{\text{in}}(r_x, r_y)$ and $\mathscr{C}_{\text{out}}(r_x, r_y)$.

A. Binary Case

We first investigate the inner and outer bound surfaces for a case in which the template patterns and sensory data alphabets are binary, $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. Let the template patterns $\boldsymbol{X} = (X_1,\ldots,X_n)$ consist of n independent drawings from a uniform Bernoulli distribution $X_i \sim B(1/2)$, $i = 1\ldots n$, and let the sensory data $\boldsymbol{Y} = (Y_1,\ldots,Y_n)$ be the output of a binary-symmetric channel with crossover probability q, $p(y|x) = q^{\bar{\delta}(x,y)}\bar{q}^{\delta(x,y)}$ where $\bar{q} = 1 - q$; $\bar{\delta}(x, y) = 1 - \delta(x, y)$; and $\delta(x, y) = 1$ if $x = y$, and otherwise $\delta(x, y) = 0$. Equivalently, we can represent Y as $Y = X \oplus W$, where $W \sim B(q)$ and is independent of X.

1) Numerical Results: We have taken two approaches to studying the surfaces of $\mathcal{R}_{\text{in}}$ and $\mathcal{R}_{\text{out}}$ for this binary case. First, we carried out the optimizations in (11) and (12) numerically. This calculation was via a Monte Carlo method which executed a dense random sampling of the set of probability distributions $p(uv|xy)$ associated with $\mathcal{P}_{\text{in}}$ and $\mathcal{P}_{\text{out}}$.[12] For each sample $p(uv|xy)$, we calculated $I(X; U)$, $I(Y; V)$ and $I(XY; UV)$; then, for each value of $r_x, r_y \in [0,1]$ the numerical estimate of $r_{\text{in}}(r_x, r_y)$ or $r_{\text{out}}(r_x, r_y)$ was the largest sample value found by the Monte Carlo search for $r_x + r_y - I(XY; UV)$. From here on, we denote the numerical surface estimates by $\widehat{r_{\text{in}}}(r_x, r_y)$ and $\widehat{r_{\text{out}}}(r_x, r_y)$.

The cardinality bound in Theorem 4 is not necessarily tight. Therefore, to assess the alphabet sizes required of $\mathcal{U}$ and $\mathcal{V}$ for the binary case, we performed our numerical experiments for increasing values of $|\mathcal{U}|$ and $|\mathcal{V}|$. For the inner bound surface, we found $|\mathcal{U}| = |\mathcal{V}| = 2$ was sufficient: no further increase in $\widehat{r_{\text{in}}}(r_x, r_y)$ was afforded by allowing $|\mathcal{U}|$, $|\mathcal{V}| = 3, 4$. For the outer bound surface, $|\mathcal{U}| = |\mathcal{V}| = 3$ was sufficient.

The surface plots from our numerical experiments are shown in Fig. 3. Fig. 4 shows representations of the distributions $p(uv|xy)$ underlying 25 different points (r_x, r_y) for Fig. 3 (a) the inner and Fig. 3 (b) the outer bounds, in which probabilities are represented by the area of white squares.[13], [14] The row–column format of the matrix $p(uv|xy)$ is $xy = 00, 01, 10, 11$ moving down rows; moving across columns, for the inner bound with $\mathcal{U}, \mathcal{V} = \{0, 1\}$ the format is $uv = 00, 01, 10, 11$, whereas for the outer bound with $\mathcal{U}, \mathcal{V} = \{0, 1, e\}$, the column format is $uv = 00, 01, 1e, 10, 1e, e0, e1, ee$. (The choice of for the third letter of $\mathcal{U}$ and $\mathcal{V}$ is explained below.)

2) Conjectured Formulas: Second, we guessed formulas for the inner and outer bound surfaces, which turned out to fit the numerical results just described. We first present the formulas, then discuss the motivations behind them.

Our formulas involve the following two functions. First, define

[12]The optimization over distributions $p(xyuv)$ reduces to a search over conditional distributions $p(uv|xy)$ because $p(xy)$ is fixed. Details of the optimization algorithm are given in [35].
[13]These are called Hinton diagrams in the machine learning literature, after their inventor Geoffrey Hinton.
[14]These distributions are not unique, as the mutual informations are unchanged under various reassignments of values of x, y, u, v, and consequent rearrangements of the entries of $p(uv|xy)$; the distributions shown have been accordingly rearranged into a common format to facilitate comparison.

$$s(r_x, r_y) = 1 - h(q_x * q * q_y),$$

where

$$q_x = h^{-1}(1 - r_x),$$
$$q_y = h^{-1}(1 - r_y);$$

$h(\cdot)$ is the binary entropy function

$$h(x) = -x\log(x) - (1 - x)\log(1 - x);$$

"*" denotes binary convolution

$$x * y = x(1 - y) + y(1 - x);$$

and $q_x, q_y \in [0, 1/2]$ to ensure that $h(\cdot)$ is invertible. Next, let $s^*(r_x, r_y)$ denote the *upper concave envelope* of $s(r_x, r_y)$

$$s^*(r_x, r_y) = \sup\theta s(r_{x1}, r_{y1}) + \bar{\theta} s(r_{x2}, r_{y2})$$

where $\bar{\theta} = 1 - \theta$; and the supremum is over all combinations $(\theta, r_{x1}, r_{y1}, r_{x2}, r_{y2})$ such that

$$(r_x, r_y) = \theta(r_{x1}, r_{y1}) + \bar{\theta}(r_{x2}, r_{y2})$$

and each variable in the optimization is restricted to the unit interval [0, 1]. As explained in Appendix F, in both this case and for the corresponding Gaussian formulas in the next section, the expression for this convex hull simplifies to

$$s^*(r_x, r_y) = \sup\theta s(r'_x, r'_y)$$

with the supremum over all combinations (θ, r'_x, r'_y) such that

$$(r_x, r_y) = \theta(r'_x, r'_y).$$

Conjecture 1: For the binary case the surfaces of $\mathscr{R}_{\text{in}}$ and $\mathscr{R}_{\text{out}}$ are

$$r_{\text{in}}(r_x, r_y) = s(r_x, r_y) \tag{14}$$

$$r_{\text{out}}(r_x, r_y) = s^*(r_x, r_y) \tag{15}$$

and the surface of the achievable rate region $\mathscr{R}$ is

$$r(r_x, r_y) = \overline{r_{\text{in}}}(r_x, r_y) = r_{\text{out}}(r_x, r_y). \tag{16}$$

3) Rationale for the Inner Bound (14): The surfaces $r_{\text{in}}(r_x, r_y)$, $r_{\text{out}}(r_x, r_y)$ are specified in terms of probability distributions $p(xyuv) = p(xy)p(uv \mid xy)$ that maximize (11) and (12). For $r_{\text{in}}(r_x, r_y)$, the distribution factorizes as $p(uv \mid xy) = p(u \mid x)p(v \mid y)$, and a natural guess is that in the maximizing distribution both $p(u \mid x)$ and $p(v \mid y)$ are binary symmetric channels

$$p(u|x) = q_x^{\bar{\delta}(x,u)} \bar{q}_x^{\delta(x,u)}, \quad p(v|y) = q_y^{\bar{\delta}(y,v)} \bar{q}_y^{\delta(y,v)},$$

or, equivalently, $U = X \oplus W_x$, $V = Y \oplus W_y$, where $W_x \sim B(q_x)$, $W_x \sim B(q_y)$, and q_x, $q_y \in \left[0, \frac{1}{2}\right]$; see Fig. 5(a). For this choice of U and V we calculate

$$\begin{aligned} r_x &= I(X; U) \\ &= H(X) - H(X|U) \\ &= 1 - H(U \oplus W_x|U) \\ &= 1 - H(W_x) \\ &= 1 - h(q_x) \end{aligned}$$

and likewise $r_y = 1 - h(q_y)$. Then

$$\begin{aligned} I(U; V) &= H(V) - H(V|U) \\ &= 1 - H\left(U \oplus W_x \oplus W \oplus W_y \middle| U\right) \\ &= 1 - h\left(q_x * q * q_y\right) \\ &= s\left(r_x, r_y\right). \end{aligned}$$

Clearly, $s(r_x, r_y)$ is a lower bound on $r_{\text{in}}(r_x, r_y)$, since: 1) $U - Y - V$, hence $UV \in \mathscr{P}_{\text{in}}$; 2) $UV \in \mathscr{C}\left(r_x, r_y\right)$; and 3)

$$\begin{aligned} r_{\text{in}}\left(r_x, r_y\right) &= \max_{UV \in \mathscr{C}\left(r_x, r_y\right)} I(U; V) \geq 1 - h\left(q * q_x * q_y\right) \\ &= s\left(r_x, r_y\right). \end{aligned}$$

The converse, $r_{\text{in}}(r_x, r_y), \leq s(r_x, r_y)$ is unproven, so the identification of $r_{\text{in}}(r_x, r_y)$ with $s(r_x, r_y)$ remains a conjecture. Nevertheless, in our numerical optimization we found no points outside of this region for any choice of (r_x, r_y), and the distributions which emerge from our computer experiments (Fig. 4(a)) closely resemble the long binary-symmetric channel in the calculation of $s(r_x, r_y)$. This provides strong experimental evidence supporting (14) in Conjecture 1.

4) Rationale for the Outer Bound (15): Clearly, $s^*(r_x, r_y)$ is a lower bound on r_{out}, since, for all $r_x, r_y \in [0, 1]$

- $r_{\text{out}}\left(r_x, r_y\right) \geq r_{\text{in}}\left(r_x, r_y\right) \Rightarrow \overline{r_{\text{out}}}\left(r_x, r_y\right) \geq \overline{r_{\text{in}}}\left(r_x, r_y\right)$;

- $\mathscr{R}_{\text{out}}$ is convex, $\Rightarrow \overline{r_{\text{out}}}(r_x, r_y) = r_{\text{out}}(r_x, r_y)$;
- $r_{\text{in}}(r_x, r_y) \geq s(r_x, r_y) \Rightarrow \overline{r_{\text{in}}}(r_x, r_y) \geq \bar{s}(r_x, r_y) = s^*(r_x, r_y)$;

and together these imply $r_{\text{out}}(r_x, r_y) \geq s^*(r_x, r_y)$. Unfortunately, we do not have a proof of the converse, $r_{\text{out}}(r_x, r_y) \leq s^*(r_x, r_y)$, so the identification of $r_{\text{out}}(r_x, r_y)$ with $s^*(r_x, r_y)$ remains a conjecture. Nevertheless, empirically (i.e., according to our numerical experiments) the outer bound surface is identical to the convex hull of the inner bound surface. Moreover, empirically, the cardinalities required to construct the outer bound are $|\mathscr{U}| = |\mathscr{V}| = 3$.

We can provide an explicit construction of the conjectured outer bound surface and the probability distributions that achieve it as follows. The distributions in this construction also agree with those found empirically, shown in Fig. 4(b). Let $\mathscr{U} = \mathscr{V} = \{0, 1, e\}$. Consider the channel diagrammed in Fig. 5(b), which could be called a "synchronous erasure channel." Here, U and V are generated by first passing X and Y through binary-symmetric channels, followed by an "erasure" $E \in \{0, 1\}$ event in which both channel outputs are preserved with probability $\theta = \Pr(E = 0)$, or both are erased ($UV = ee$) with probability $\bar{\theta} = 1 - \theta = \Pr(E = 1)$. An explicit formula for this channel is

$$\begin{aligned} &p(uv|xy) \\ &= \theta^{\bar{\delta}_e(u,v)} \bar{\theta}^{\delta_e(u,v)} \\ &\left(q_x^{\bar{\delta}(x,u)} \bar{q}_x^{\delta(x,u)} q_y^{\bar{\delta}(y,v)} \bar{q}_y^{\delta(y,v)}\right)^{\bar{\delta}_e(u,v)} \Delta_e(u,v) \end{aligned}$$

where

$$\begin{aligned} \delta(\alpha, \beta) &= \begin{cases} 1, & \text{if } \alpha = \beta \\ 0, & \text{if } \alpha \neq \beta \end{cases} \\ \delta_e(\alpha, \beta) &= \begin{cases} 1, & \text{if } (\alpha, \beta) = (e, e) \\ 0, & \text{if } (\alpha, \beta) \neq (e, e) \end{cases} \\ \Delta_e(\alpha, \beta) &= \begin{cases} 0, & \text{if } \alpha = e, \beta \neq e \\ 0, & \text{if } \alpha \neq e, \beta = e \\ 1, & \text{otherwise} \end{cases} \end{aligned}$$

and

$$\bar{\delta}(\alpha, \beta) = 1 - \delta(\alpha, \beta), \bar{\delta}_e(\alpha, \beta) = 1 - \delta_e(\alpha, \beta).$$

Equivalently, we can represent U and V as follows. Let $W \sim B(q)$, $W_x \sim B(q_x)$, $W_y \sim B(q_y)$, $E \sim B(\theta)$ be Bernoulli random variables that are independent of each other and independent of X and Y, and define

$$\begin{aligned} Y &= X \oplus W \\ U &= (X \oplus W_x) \otimes E \\ V &= (Y \oplus W_y) \otimes E \end{aligned}$$

where the multiplication by E is defined by

$$\alpha \otimes E = \begin{cases} \alpha, & \text{if } E = 0 \\ e, & \text{if } E = 1 . \end{cases}$$

It is straightforward to verify $UV \in \mathscr{P}_{\text{out}}$. To check $U - X - Y$, write

$$\begin{aligned} I(U;Y|X) &= I((X \oplus W_X) \otimes E; X \oplus W|X) \\ &= I(W_X \otimes E; W|X) \\ &= 0 \end{aligned}$$

where the last line follows from the independence of W, W_X and E from each other and X. A similar calculation shows $X - Y - V$. Finally, calculating the rate region surface associated with this choice of UV we get, first

$$\begin{aligned} r_X &= I(X;U) \\ &= I(X;(X \oplus W_X) \otimes E, E) \\ &= I(X;E) + I(X;(X \oplus W_X) \otimes E|E) \\ &= 0 + \bar{\theta} I(X;E|E=1) + \theta I(X; X \oplus W_X|E=0) \\ &= 0 + 0 + \theta(1 - h(q_X)) \end{aligned}$$

where the last step follows from the previous calculations for the inner bound; and a similar calculation shows $r_Y = \theta(1 - h(q_Y))$. Then, using $U - X - Y$, $X - Y - V$ to write

$$\begin{aligned} &I(U;V) - I(U;V|XY) \\ &= I(X;U) + I(Y;V) - I(XY;UV) \\ &= I(X;U) + I(Y;V) - [H(U) + H(U|V) - H(UV|XY)] \end{aligned}$$

we have (suppressing some detail)

$$\begin{aligned} H(U) &= H(U, E) \\ &= H(E) + H(U|E) \\ &= h(\theta) + \theta H(U|E=0) + \bar{\theta} H(U|E=1) \\ &= h(\theta) + \theta + 0 \end{aligned}$$

$$\begin{aligned} H(U|V) &= H(U|V, E) \\ &= \theta H(U|V, E=0) + \bar{\theta} H(U|V, E=1) \\ &= \theta h(q_X * q * q_Y) \end{aligned}$$

$$\begin{aligned} H(UV|XY) &= H(E|XY) + H(UV|XY, E) \\ &= h(\theta) + \theta[H(U|X, E=0) + H(V|Y, E=1)] \\ &= h(\theta) + \theta[h(q_X) + h(q_Y)] . \end{aligned}$$

Putting these together and canceling terms

$$\begin{aligned} &I(U;V) - I(U;V|XY) = \theta(1 - h(q_X * q * q_Y)) \\ &= s^*(r_X, r_Y) . \end{aligned}$$

Thus, we have constructed an explicit example which achieves s*(r_x, r_y) with $|\mathcal{U}| = |\mathcal{V}| = 3$.

B. Gaussian Case

We now consider a Gaussian version of our problem. Let X and Y be zero-mean Gaussian random variables with correlation coefficient ρ_{xy}. We propose explicit formulas for the surfaces of $\mathcal{R}_{\text{in}}$ and $\mathcal{R}_{\text{out}}$ for the Gaussian case, in terms of the following two functions. In both formulas, put

$$r_x = -\frac{1}{2}\log\left(1-\rho_{xu}^2\right)$$
$$r_y = -\frac{1}{2}\log\left(1-\rho_{yv}^2\right).$$

Note that these expressions determine the correlation coefficients ρ_{xu} and ρ_{yv}. Define

$$S(r_x, r_y) = -\frac{1}{2}\log\left(1-\rho_{xy}^2\rho_{yv}^2\rho_{xu}^2\right) \tag{17}$$

and

$$S^*(r_x, r_y) = r_x + r_y + \frac{1}{2}\log\left[1+\frac{2\rho\gamma-\beta}{1-\rho^2}\right] \tag{18}$$

where

$$\begin{aligned}\gamma &= \rho_{xy}\rho_{xu}\rho_{yv}\\ \beta &= \rho_{xu}^2+\rho_{yv}^2-\left(1-\rho_{xy}^2\right)\rho_{xu}^2\rho_{yv}^2,\\ \rho &= \frac{\beta}{2\gamma}-\sqrt{\left(\frac{\beta}{2\gamma}\right)^2-1}\,.\end{aligned} \tag{19}$$

Conjecture 2: In the Gaussian case, the surfaces of $\mathcal{R}_{\text{in}}$ and $\mathcal{R}_{\text{out}}$ are

$$r_{\text{in}}(r_x, r_y) = S(r_x, r_y) \tag{20}$$

$$r_{\text{out}}(r_x, r_y) = S^*(r_x, r_y)\,. \tag{21}$$

Fig. 6 shows plots of the inner and outer bounds and their difference, as well as the difference between the outer bound and the convex hull of the inner bound. Interestingly, unlike the binary case, for the Gaussian case the outer bound is not equal to the convex hull of the inner bound.

The following proof relies on some basic properties of the mutual information between Gaussian random variables, given as lemmas in Appendix G.

In the analysis that follows, we assume that the maximizing distributions are Gaussian. Under this assumption, we solve the inner and outer bounds. Except for this unproved assumption, the proof of the conjecture is complete.

Proof: (Conjecture 2, eq. (20)) As noted in Appendix G, mutual informations between jointly Gaussian random variables are completely determined by their correlation coefficients. For a length-4 Markov chain $U - X - Y - V$ of jointly Gaussian random variables $I(U; V \mid XY) = 0$ and, applying Lemma 9 from Appendix G we have $\rho_{uv} = \rho_{xu}\rho_{xy}\rho_{yv}$, hence

$$I(U;V) - I(U;V \mid XY) = -\frac{1}{2}\log\left(1 - \rho_{xu}^2\rho_{xy}^2\rho_{yv}^2\right).$$

This mutual information is maximized when the constraints $I(X; U) \le r_x$, $I(Y; V) \le r_y$ are satisfied with equality, hence when ρ_{xu} and ρ_{yv} satisfy $r_x = -\frac{1}{2}\log\left(1 - \rho_{xu}^2\right)$ and $r_y = -\frac{1}{2}\log\left(1 - \rho_{yv}^2\right)$.

The following proof for the surface of the outer bound region uses the form of $r_{\text{out}}(r_x, r_y)$ given by (13). In this case, the optimization problem reduces to minimizing $I(XY; UV)$ subject to the length-3 Markov constraints $U - X - Y$, $X - Y - V$.

Proof: (Conjecture 2, eq. (21)) Using Lemma 10 from Appendix G, we have

$$C_{xy,uv} = \begin{bmatrix} \rho_{xu} & \rho_{xv} \\ \rho_{yu} & \rho_{yv} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{bmatrix} \begin{bmatrix} \rho_{xu} & 0 \\ 0 & \rho_{yv} \end{bmatrix}.$$

The left-hand matrix in this decomposition is $C_{xy,xy}$, denoted hereafter simply as C, and we denote the right-hand matrix by D. Then applying Lemma 8 from Appendix G yields

$$\begin{aligned} &I(XY; UV) \\ &= \frac{1}{2}\log|C| - \frac{1}{2}\log\left|C - C_{xy,uv}C_{uv,uv}^{-1}C_{uv,yx}\right| \\ &= \frac{1}{2}\log|C| - \frac{1}{2}\log\left|C - CDC_{uv,uv}^{-1}DC\right| \\ &= -\frac{1}{2}\log|C| - \frac{1}{2}\log\left|C^{-1} - DC_{uv,uv}^{-1}D\right|. \end{aligned}$$

Substituting for the 2×2 matrices in this last expression and rearranging terms yields

$$I(XY;UV) = -\frac{1}{2}\log\left|1 + \frac{2\rho_{uv}\gamma - \beta}{1 - \rho_{uv}^2}\right|$$

where γ and β are defined in (19).

By assumption, ρ_{xu} and ρ_{yv} are fixed, so we optimize $I(XY; UV)$ only with respect to ρ_{uv}. Setting $\partial I(XY; UV)/\partial\rho_{uv} = 0$ and solving, we obtain that, if $\beta > 2\gamma > 0$, then the maximum is achieved at $\rho_{uv}^* = \rho$, where ρ is defined in (19).

To complete the proof we must show that $\beta > 2\gamma > 0$. Noting that β, $\gamma > 0$ and substituting, the desired inequality becomes

$$\rho_{xu}^2 + \rho_{yv}^2 - \rho_{xu}^2\rho_{yv}^2 > 2\rho_{xy}\rho_{xu}\rho_{yv} - \rho_{xy}^2\rho_{xu}^2\rho_{yv}^2.$$

Subtracting from each side and factoring yields the equivalent inequality

$$-\left(1-\rho_{xu}^2\right)\left(1-\rho_{yv}^2\right) > -\left(1-\rho_{xy}\rho_{xu}\rho_{yv}\right)^2.$$

To show that this holds for all ρ_{xy}, note that the maximum of the right-hand side is achieved by $\rho_{xy} = 1$, so that the inequality becomes

$$\left(1-\rho_{xu}^2\right)\left(1-\rho_{yv}^2\right) - \left(1-\rho_{xu}\rho_{yv}\right)^2 < 0.$$

This inequality holds, since

$$\begin{aligned}&\left(1-\rho_{xu}^2\right)\left(1-\rho_{yv}^2\right) - \left(1-\rho_{xu}\rho_{yv}\right)^2\\ &= -\rho_{xu}^2 - \rho_{yv}^2 + 2\rho_{xu}\rho_{yv}\\ &= \left(\rho_{xu}-\rho_{yv}\right)\left(\rho_{yv}-\rho_{xu}\right)\\ &= -\left(\rho_{xu}-\rho_{yv}\right)^2\\ &< 0.\end{aligned}$$

VIII. Conclusion

We have presented an information-theoretic analysis of pattern recognition systems subject to data compression constraints. Our main results consist of fundamental bounds characterizing the minimum sensory and memory information budgets required for reliable pattern recognition, or, equivalently, the maximum number of patterns that can be discriminated on given sensory and memory data budgets.

As a starting point, we have focused on the case of unstructured data, in which patterns are representable as vectors with i.i.d. components, and the sensory data observation channel is memoryless. In recent years, there has been much theoretical and experimental work aimed at developing methods to render data into a format with independent (or approximately independent) components (see, e.g., [36]–[39]). Such methods have been especially successful in the study of "natural" signals, e.g., sounds and imagery in naturally occurring environments. Nevertheless, a decomposition into independent components is often impossible or only approximate, and it will be important in future work to extend our results to cover the case of correlated components and channels with memory.

We have focused on "reliable" pattern recognition systems, in the sense that the recognition error rate is able to be made arbitrarily close to zero. Nevertheless, in some applications it is of interest (or unavoidable) to allow less-than-perfect accuracy. This can be partly addressed by recasting the recognition problem as a "coarse-to-fine" search, where the system is given information in several successive stages, and at each stage is required only to partially recognize the pattern, i.e., to identify the pattern as belonging to a particular subclass,

postponing definitive identification for the final stage. Extending our results to this successive refinement setting is relatively straightforward; see [40]. The more direct approach of explicitly allowing a strictly positive error rate is an open problem.

Much work remains to be done in designing practical pattern recognition systems that achieve the bounds described herein. One of the most challenging problems in this regard is the design of adequate statistical models of real-world signals. For examples of progress on this exciting front, see [11], [36], [41]–[48]. Another significant challenge is that of learning optimal classifiers from training data. In this connection, it will likely prove fruitful to explore connections between the present results and those established in machine learning theory; see, e.g., [7]–[10]. Another practical challenge is to build systems that make optimal use of time. Donald Geman and colleagues have been developing the theory of systems that reach their pattern recognition decisions with a minimum amount of computation [49]. It will be interesting to explore the relationship of this concept with our results concerning recognition using the minimum amount of information.

Open theoretical problems include the calculation of error exponents and, most importantly, the closing of the gap between our inner and outer bounds. As discussed in Section VI-B, the gap in our problem bears close resemblance to that in the distributed source coding problem. A solution to the distributed source coding problem would likely lead to a solution to ours, and *vice versa*.[15]

Acknowledgment

The authors gratefully acknowledge stimulating discussions with Michael DeVore, Naveen Singla, and Po-Hsiang Lai.

This work was supported by the Mathers Foundation and by the Office of Naval Research. The material in this paper was presented in part at an ONR PI meeting, Minneapolis, MN, May 2003; the Neural Information Processing Systems Workshop on Information Theory and Learning, Whistler, BC, Canada, December 2003; the IEEE International Symposium on Information Theory, Chicago, IL, June/July 2004; and the IEEE Information Theory Workshop, Punta del Este, Uruguay, March 2006.

Appendix A: Proof of the Inner Bound

In this section we prove the inner bound $\overline{\mathscr{R}}_{\text{in}} \subseteq \mathscr{R}$, Theorem 2. The proof relies on standard random coding arguments and properties of strongly jointly typical sets [30]. Given a joint distribution $p(xyuv)$, the strongly jointly δ-typical set is defined by

$$\mathscr{T}^{\delta}_{UVXY} = \left\{ \boldsymbol{xyuv}: \left| N(xyuv|\boldsymbol{xyuv})/n - p(xyuv) \right| \leq \delta \forall xyuv \in \mathscr{X}\mathscr{Y}\mathscr{U}\mathscr{V} \right\}$$

where $N(xyuv \mid \boldsymbol{xyuv})$ is the number of times the symbol combination $xyuv$ occurs in $\boldsymbol{xyuv}$. Likewise, we write. e.g., $\mathscr{T}^{\delta}_{X}$, $\mathscr{T}^{\delta}_{XY}$, $\mathscr{T}^{\delta}_{XYU}$ for singles, pairs, and triples. We will also use conditionally strongly jointly δ-typical sets, for example

[15]During the review process for this paper, Servetto indeed claimed a solution to the distributed source coding problem using a novel approach. Unfortunately, he suffered an untimely death on 7/24/2007, before finalizing his work. The most recent public draft of his paper on this topic is available on the arXiv (see [50]).

$$\mathcal{T}^{\delta}_{\boldsymbol{x}U} = \left\{\boldsymbol{u} : (\boldsymbol{x}\boldsymbol{u}) \in \mathcal{T}^{\delta}_{XU}\right\}.$$

The subscripts are omitted when context allows. We will also need the fact that for any positive numbers $\delta, \epsilon > 0$, fixed vector $\boldsymbol{x}$, and large enough n

$$2^{-n[I(X;Y)+\epsilon]} \leq \Pr\left(\boldsymbol{x}\boldsymbol{Y} \in \mathcal{T}^{\delta}_{\boldsymbol{x}Y}\right) \leq 2^{-n[I(X;Y)-\epsilon]}. \tag{22}$$

Theorem 1: Suppose $\boldsymbol{R}$ is a point in the convex hull of $\mathcal{R}_{\text{in}}$, $\boldsymbol{R} \in \overline{\mathcal{R}}_{\text{in}}$. That is, $\boldsymbol{R} = \sum_{q \in \mathcal{Q}} p(q)\boldsymbol{R}_q$, where $p(q)$ is a probability distribution over some finite alphabet $\mathcal{Q}$, and for each $q \in \mathcal{Q}$, $\boldsymbol{R}_q = \left(R_{xq}, R_{yq}, R_{cq}\right) \in \mathcal{R}_{\text{in}}$.

We wish to show that for any $\epsilon > 0$ and large enough n, there exists an (M_c, M_x, M_y, n) code $(f, \boldsymbol{\phi}, g)$ with rates $R'_c = \frac{1}{n}\log M_c$, $R'_x = \frac{1}{n}\log M_x$, $R'_y = \frac{1}{n}\log M_y$ such that $R'_c \geq R_c$, $R'_x \leq R_x$, $R'_y \leq R_y$, and $P_e^n \leq \epsilon$.

By definition, $\boldsymbol{R}_q \in \mathcal{R}_{\text{in}}$ implies that for each $q \in \mathcal{Q}$ there exist random variables U_q, V_q such that

$$p_q(xyuv) = p(xy)p_q(u|x)p_q(v|y)$$

and

$$\begin{aligned} R_{xq} &= I\left(X; U_q\right) + \alpha_{xq} \\ R_{yq} &= I\left(Y; V_q\right) + \alpha_{yq} \\ R_{cq} &= R_{xq} + R_{yq} - I\left(XY; U_q V_q\right) - \gamma_q \end{aligned}$$

for some values $\alpha_{xq}, \alpha_{yq}, \gamma_q > 0$ such that $\gamma_q \geq \alpha_{xq} + \alpha_{yq}$.[16] Now let

$$\begin{aligned} R'_{xq} &= I\left(X; U_q\right) + \alpha_{xq}/4, \quad R'_x = \sum_{q \in \mathcal{Q}} p(q)R'_{xq} \\ R'_{yq} &= I\left(Y; V_q\right) + \alpha_{yq}/4, \quad R'_y = \sum_{q \in \mathcal{Q}} p(q)R'_{yq} \end{aligned}$$

and

$$\begin{aligned} R'_{cq} &= R'_{xq} + R'_{yq} - I\left(XY; U_q V_q\right) - \gamma_q/4 \\ R'_c &= \sum_{q \in \mathcal{Q}} p(q)R'_{cq}. \end{aligned}$$

With these choices, we have $R'_x \leq R_x$, $R'_y \leq R_y$, $R'_c \geq R_c$.

[16] This last condition ensures $R_{cq} \leq R_{xq} + R_{yq} - I(XY; U_q V_q)$.

Given

$$X^n \sim \prod_{i=1}^{n} p(x_i), \quad Y^n \sim \prod_{i=1}^{n} p(y_i)$$

divide the sequences into $|\mathcal{Q}|$ segments with lengths $n_q = np(q)$, denoted X^{n_q}, Y^{n_q}, i.e.,

$$X^n = \left[X^{n_1} X^{n_2} \ldots X^{n_{|\mathcal{Q}|}}\right], \quad Y^n = \left[Y^{n_1} Y^{n_2} \ldots Y^{n_{|\mathcal{Q}|}}\right].$$

Finally, we will use the additional notation: $M_{xq} = 2^{n_q R'_{xq}}$, $M_{yq} = 2^{n_q R'_{yq}}$ and $\mathcal{M}_{xq} = \{1, \ldots, M_{xq}\}$, $\mathcal{M}_{yq} = \{1, \ldots, M_{yq}\}$.

We will construct the desired overall code (f, ϕ, g) with rate (R'_c, R'_x, R'_y) by first constructing encoders f_q, ϕ_q with rates R'_{xq}, R'_{yq} for the $|\mathcal{Q}|$ component sequences X^{n_q}, Y^{n_q}, then constructing a classifier g which acts on the combined outputs of the encoders.

Please refer to Fig. 7 for a summary of the notation introduced below.

1. *Codebooks*: For each $q \in \mathcal{Q}$, from $p_q(xyuv)$ compute the marginal distributions $p_q(u)$,$p_q(v)$. To serve as memory codewords, select M_{xq} length-n_q vectors by sampling with replacement from a uniform distribution over the set $\mathcal{T}^{\delta}_{U}$. Assign each codeword a unique index $i_q \in \mathcal{M}_{xq} = \{1, 2, \ldots, M_{xq}\}$. To serve as sensory codewords, similarly select M_{yq} length-n_q vectors by from $\mathcal{T}^{\delta}_{V}$, and assign each an index $j_q \in \mathcal{M}_{yq} = \{1, 2, \ldots, M_{yq}\}$. Denote the codebooks

$$\mathcal{B}_u(q) = \left\{u^{n_q}(1), \ldots, u^{n_q}(M_{xq})\right\}$$
$$\mathcal{B}_v(q) = \left\{v^{n_q}(1), \ldots, v^{n_q}(M_{yq})\right\}.$$

2. *Encoders*: We define encoders f_q and ϕ_q in terms of maps

$$\varphi_{xq}: \mathcal{X}^{n_q} \to \mathcal{U}^{n_q} \quad b_{xq}: \mathcal{U}^{n_q} \to \mathcal{M}_{xq}$$
$$\varphi_{yq}: \mathcal{Y}^{n_q} \to \mathcal{V}^{n_q} \quad b_{yq}: \mathcal{V}^{n_q} \to \mathcal{M}_{yq}$$

as follows. Given any $x^{n_q} \in \mathcal{X}^{n_q}$, search the codebook $\mathcal{B}_u(q)$ for a codeword u^{n_q} such that $(x^{n_q}, u^{n_q}) \in \mathcal{T}^{\delta}_{XU}$. If this search is successful, set $\varphi_{xq}(x^{n_q}) = u^{n_q}$, $b_{xq}(u^{n_q}) = i_q$, $b^{-1}_{xq}(i_q) = u^{n_q}$, where i_q is the index of u^{n_q} in $\mathcal{B}_u(q)$. If the search fails, (arbitrarily) set $u^{n_q} = u^{n_q}(1)$ so that φ_{xq} and b_{xq} are defined for all of $\mathcal{X}^{n_q}$. In the same way, given any $y^{n_q} \in \mathcal{Y}^{n_q}$, search $\mathcal{B}_v(q)$ for a codeword v^{n_q} such that $(y^{n_q}, v^{n_q}) \in \mathcal{T}^{\delta}_{YV}$. If successful, denote the index of the found codeword j_q, and

set $\varphi_{yq}(y^{nq}) = v^{nq}$, $b_{yq}(v^{nq}) = j_q$, $b_{yq}^{-1}(j_q) = v^{nq}$. For search failure, set $j_q = 1$, $v^{nq} = v^{nq}(1)$ so that φ_{yq} and b_{yq} are defined for all of $\mathcal{Y}^{nq}$.

Finally, define

$$f_q : \mathcal{X}^{nq} \to \mathcal{M}_{xq}, \quad f_q(x^{nq}) = i_q \triangleq b_{xq}(\varphi_{xq}(x^{nq}))$$
$$\phi_q : \mathcal{Y}^{nq} \to \mathcal{M}_{yq}, \quad \phi_q(y^{nq}) = j_q \triangleq b_{yq}(\varphi_{yq}(y^{nq})).$$

Now, given vectors $x^n = (x^{n1} \ldots x^{n|\mathcal{Q}|})$ and $y^n = (y^{n1} \ldots y^{n|\mathcal{Q}|})$, the encoders above each produce $|\mathcal{Q}|$ vectors $\varphi_{xq}(x^{nq}) = u^{nq}$, $\varphi_{yq}(y^{nq}) = v^{nq}$, and $|\mathcal{Q}|$ indices $b_{xq}(u^{nq}) = i_q$, $b_{yq}(y^{nq}) = j_q$. Denote the concatenations of these

$$\begin{aligned} u^n &= \varphi_x(x^n) \triangleq [u^{n_1} \ldots u^{n_{|\mathcal{Q}|}}], \boldsymbol{i} = b_x(u^n) \\ &\triangleq [i_1 \ldots i_{|\mathcal{Q}|}] v^n = \varphi_y(y^n) \triangleq [v^{n_1} \ldots v^{n_{|\mathcal{Q}|}}], \boldsymbol{j} = b_y(v^n) \\ &\triangleq [j_1 \ldots j_{|\mathcal{Q}|}]. \end{aligned}$$

Note that the vector of integers $\boldsymbol{i}$ ranges over $M_x = \prod_{q \in \mathcal{Q}} M_{xq}$ different values $\boldsymbol{i}(i), i = 1 \ldots M_x$. Let the map between the vectors $\boldsymbol{i}$ and the corresponding integers i be ℓ_x, i.e., if $\boldsymbol{i} = \boldsymbol{i}(i)$, let $\ell_x(\boldsymbol{i}) = i$, and $\ell_x^{-1}(i) = \boldsymbol{i}$. Similarly, $\boldsymbol{j}$ ranges over $M_y = \prod_{q \in \mathcal{Q}} M_{yq}$ values, $\boldsymbol{j}(j), j = 1 \ldots M_y$ and we define ℓ_y such that if $\boldsymbol{j} = \boldsymbol{j}(j)$, then $\ell_y(\boldsymbol{j}) = j$ and $\ell_y^{-1}(j) = \boldsymbol{j}$. Then we can specify encoders $f' : \mathcal{X}^n \to \mathcal{M}_x$ and $\phi : \mathcal{Y}^n \to \mathcal{M}_y$ for full length-n vectors x^n and y^n by

$$f'(x^n) = i \triangleq \ell_x(b_x(\varphi_x(x^n)))$$
$$\phi(y^n) = j \triangleq \ell_y(b_y(\varphi_y(y^n))).$$

To finalize the construction of the memory encoder, for any given labeled template pattern $\boldsymbol{t}(w) = (x^n, w)$, let $f : \mathcal{X}^n \times \mathcal{M}_c \to \mathcal{M}_x \times \mathcal{M}_c$ be defined by

$$f(\boldsymbol{t}(w)) = m(w) = (i, w) \triangleq (f'(x^n), w).$$

The rates of the encoders are (R'_x, R'_y), as verified by calculating

$$\begin{aligned} \frac{1}{n} \log M_x &= \frac{1}{n} \log \prod_{q \in \mathcal{Q}} M_{xq} = \frac{1}{n} \sum_{q \in \mathcal{Q}} \log 2^{nq} R'_{xq} \\ &= \sum_{q \in \mathcal{Q}} p(q) R'_{xq} = R'_x \\ \frac{1}{n} \log M_y &= \frac{1}{n} \log \prod_{q \in \mathcal{Q}} M_{yq} = \frac{1}{n} \sum_{q \in \mathcal{Q}} \log 2^{nq} R'_{yq} \\ &= \sum_{q \in \mathcal{Q}} p(q) R'_{yq} = R'_y. \end{aligned}$$

3. *Memorization*: Given a realization of the template patterns $\mathcal{C}_x = (\boldsymbol{t}(1) \ldots \boldsymbol{t}(M_c))$, $\boldsymbol{t}(w) = (x^n(w), w)$ and the encoders defined above f and ϕ, compute the memory data $\mathcal{C}_u = f(C_x) \triangleq \{f(\boldsymbol{t}(1)), \ldots, f(\boldsymbol{t}(M_c))\} = \{m(1), \ldots, m(M_c)\}$.

4. *Recognition Function*: Given the stored memory data $\mathscr{C}_u$, we proceed to construct the classifier g as follows.

For each $q \in \mathcal{Q}$, given any pair of length n_q vectors u^{n_q}, v^{n_q}, define a function that tests the pair for strong joint typicality

$$\rho_q\left(u^{n_q}, v^{n_q}\right) = \mathbb{T}\left[\left(u^{n_q}, v^{n_q}\right) \in \mathscr{T}^{\delta}_{UV}\right]$$

where $\mathbb{T}[\cdot]$ is the truth-indicator function $\mathbb{T}[A] = 1$ if A is true, and $\mathbb{T}[A] = 0$ if A is false.

Now, given the sensory data j, compute its vector representation $\boldsymbol{j} = \left[j_1 \ldots j_{|\mathcal{Q}|}\right] = \ell_y^{-1}(j)$. For each $q \in \mathcal{Q}$, retrieve from $\mathscr{B}_v(q)$ the corresponding codeword $v^{n_q} = b_{yq}^{-1}(j_q)$. Similarly, for each memory $m(w) = (i, w) \in \mathscr{C}_u$, compute the corresponding vector $\boldsymbol{i} = \left[i_q \ldots i_{|\mathcal{Q}|}\right] = \ell_x^{-1}(i)$, and from the memory codebook $\mathscr{B}_u(q)$ retrieve $u^{n_q} = b_{xq}^{-1}(i_q)$. Next, define a function $r_w : \mathscr{M}_y \to \{0, 1\}$ that tests each v^{n_q} in $v^n = \left[v^{n_1} \ldots v^{n_{|\mathcal{Q}|}}\right]$ against the corresponding u^{n_q} in $u^n = \left[u^{n_1} \ldots u^{n_{|\mathcal{Q}|}}\right]$, reporting a 1 if all compared pairs are jointly typical and zero otherwise

$$r_w(j) = \mathbb{T}\left[\rho_1\left(u^{n_1}, v^{n_1}\right) \ldots \rho_{|\mathcal{Q}|}\left(u^{n_{|\mathcal{Q}|}}, v^{n_{|\mathcal{Q}|}}\right) = \mathbf{1}^{|\mathcal{Q}|}\right]$$

where $\mathbf{1}^{|\mathcal{Q}|}$ denotes the length-$|\mathcal{Q}|$ all-ones vector.

We can now specify the recognition function g as follows. Given the encoded sensory data j, the recognition module searches for a unique $w' \in \mathscr{M}_c$ such that $r_{w'}(j) = 1$. If this search is successful, set $\widehat{w} = w'$. Otherwise, if there is none or more than one such value, declare an error and (arbitrarily) set $\widehat{w} = 1$. Thus, we have defined $g : \mathscr{M}_y \times (\mathscr{M}_x)^{M_c} \to \mathscr{M}_c$, $g(j, \mathscr{C}_u) = \widehat{w}$, as desired.

A. Performance Analysis

1) Error Events:

We analyze the probability of error for a given $W = w$, $\boldsymbol{T}(w) = (X^n(w), w)$, Y^n. Denote the results of processing these with the components of the code (f, ϕ, g) above by $M(w) = (I(w), w) = f(\boldsymbol{T}(w))$, $J(w) = \phi(Y^n)$; $U^{n_q}(w) = \varphi_{xq}\left(X^{n_q}\right)$, $V^{n_q}(w) = \varphi_{yq}\left(Y^{n_q}\right)$ for each $q \in \mathcal{Q}$, and $R_w = r_w(J(w))$. The following is an exhaustive list of possible errors.

First, in words, the possible errors are as follows:

- the sensory data and pattern template are not jointly typical;
- the pattern template is unencodable;
- the sensory data is unencodable;
- the codewords for the memory and sensory data are not jointly typical;

- the sensory data is jointly typical with more than one memory codeword;
- two different patterns are assigned the same memory codeword.

More formally, we express the error events thus: For events E_i, let $\bar{E}_k = \left(\cup_{i=1}^{k} E_i\right)^c$, where A^c denotes the complement of A. For each $q \in \mathcal{Q}$

- $E_1(q) = \left\{\left(X^{nq}, Y^{nq}\right) \notin \mathcal{T}_{XY}^{\delta}\right\}$;
- $E_2(q) = \bar{E}_1(q) \cap \left\{\forall U^{nq} \in \mathcal{B}_u(q) : \left(X^{nq}, U^{nq}\right) \notin \mathcal{T}_{XU}^{\delta}\right\}$;
- $E_3(q) = \bar{E}_1(q) \cap \left\{\forall V^{nq} \in \mathcal{B}_v(q) : \left(Y^{nq}, V^{nq}\right) \notin \mathcal{T}_{YV}^{\delta}\right\}$;
- $E_4(q) = \bar{E}_3(q) \cap \left\{\left(U^{nq}(w), V^{nq}(w)\right) \notin \mathcal{T}_{UV}^{\delta}\right\}$; and letting $E_i \triangleq \cup_{q \in \mathcal{Q}} E_i(q)$, $i = 1 \ldots 4$
- $E_5 = \bar{E}_4 \cap \{\exists w' \in \mathcal{M}_c : w' \neq w, r_{w'}(J(w)) = 1\}$;
- $E_6 = \bar{E}_5 \cap \{\exists w' \in \mathcal{M}_c : w' \neq w, I(w') = I(w)\}$.

Each of these vanishes as $n \to \infty$, for the following reasons:

- $P(E_1(q)) \to 0$, by the strong asymptotic equipartition property (AEP);
- $P(E_2(q)) \to 0$, because $R'_{xq} \geq I(X; U_q)$;
- $P(E_3(q)) \to 0$, because $R'_{yq} \geq I(Y; V_q)$;
- $P(E_4(q)) \to 0$, because of the factorization $p_q(xyuv) = p(xy)p_q(u \mid x)p_q(v \mid y)$ and the Markov lemma.
- Regarding $P(E_5)$ Rewrite E_5 as

$$\begin{aligned} E_5 &= \bar{E}_4 \cap \{\exists w' \in \mathcal{M}_c : w' \neq w, r_{w'}(J(w)) = 1\} \\ &= \bar{E}_4 \cap \left\{ \bigcup_{w' \in \{\mathcal{M}_c \setminus w\}} \left\{\forall q \in \mathcal{Q}, \left(U^{nq}(w'), V^{nq}(w)\right) \in \mathcal{T}_{UV}^{\delta}\right\}\right\}. \end{aligned}$$

Then

$$\begin{aligned} P(E_5) &\leq |\mathcal{C}_u| \prod_{q \in \mathcal{Q}} 2^{-nq\left[I(U_q; V_q) - \epsilon_q\right]} \\ &\leq M_c \prod_{q \in \mathcal{Q}} 2^{-nq\left[I(U_q; V_q) - \epsilon_q\right]} \\ &= 2^{nR'_c 2^{-n\sum_{q \in \mathcal{Q}} p(q)\left[I(U_q; V_q) + \epsilon_q\right]}} \\ &= 2^{n\left[R'_c - \sum_{q \in \mathcal{Q}} p(q) I(U_q; V_q) - \epsilon\right]}. \end{aligned}$$

So, $P(E_5) \to 0$ as $n \to \infty$ if. $R'_c \leq \sum_{q \in \mathcal{Q}} p(q) I(U_q; V_q) + \epsilon$. This is indeed the case, since

$$
\begin{aligned}
R'_c &= \sum_{q \in \mathcal{Q}} p(q) R'_{cq} \\
&= \sum_{q \in \mathcal{Q}} p(q)\left[R'_{xq} + R'_{yq} - I\left(XY; U_q V_q\right) - \gamma_q\right] \\
&= \sum_{q \in \mathcal{Q}} p(q)\left[I\left(X; U_q\right) + I\left(Y; V_q\right) - I\left(XY; U_q V_q\right) + \alpha_{xq} + \alpha_{yq} - \gamma_q\right] \\
&\overset{(a)}{\leq} \sum_{q \in \mathcal{Q}} p(q)\left[I\left(X; U_q\right) + I\left(Y; V_q\right) - I\left(XY; U_q V_q\right)\right] + \epsilon \\
&\overset{(b)}{\leq} \sum_{q \in \mathcal{Q}} p(q) I\left(U_q; V_q\right) + \epsilon
\end{aligned}
$$

where (a) is because $(\gamma - \alpha_{xq} + \alpha_{yq})$; and (b) follows from elementary properties of mutual information and from the factorization. $p(q)p(xy)p(u \mid xq)\, p(v \mid yq)$.

- Regarding $P(E_6)$: Denoting the components of the codeword for each memory $M(w) \in \mathscr{C}_u$ by $U^{n_q}(w)$, rewrite event E_6 as

$$
\begin{aligned}
E_6 &= \bar{E}_5 \cap \{\exists w' \in \mathscr{M}_c : w' \neq w,\, I(w') = I(w)\} \\
&= \bar{E}_5 \cap \left\{\exists w' \in \mathscr{M}_c : w' \neq w, \forall q \in \mathcal{Q} : \left(X^{n_q}(w'), U^{n_q}(w)\right) \in \mathscr{T}^{\delta}_{XU}\right\}.
\end{aligned}
$$

Then

$$
\begin{aligned}
P(E_6) &\leq M_c \prod_{q \in \mathcal{Q}} 2^{-n_q\left[I\left(X; U_q\right) - \epsilon_q\right]} \\
&= 2^{n\left[R'_c - \sum_{q \in \mathcal{Q}} p(q) I\left(X; U_q\right) - \epsilon\right]}.
\end{aligned}
$$

So as if $P(E_6) \to 0$ as $n \to \infty$ if $R'_c \leq \sum_{q \in \mathcal{Q}} p(q) I\left(X; U_q\right) + \epsilon]$. This is indeed the case: From our preceding calculation for $P(E_5)$, we have $R'_c \leq \sum_{q \in \mathcal{Q}} p(q) I\left(U_q; V_q\right) + \epsilon$; and the assumed factorization of $p_q(xyuv)$ implies that the following is a Markov chain: $U_q - X - Y - V_q$. Hence, by the data processing inequality

$$
\begin{aligned}
R_c &\leq \sum_{q \in \mathcal{Q}} p(q) I\left(U_q; V_q\right) + \epsilon \\
&\leq \sum_{q \in \mathcal{Q}} p(q) I\left(U_q; X\right) + \epsilon.
\end{aligned}
$$

This concludes the proof of the inner bound.

Appendix B: Proof of the Outer Bound

In this section we prove Theorem 1, which states the outer bound $\mathscr{R} \subseteq \mathscr{R}_{\text{out}}$. In the proof let W be the test index, selected from a uniform distribution $p(w)$ over the pattern indices $\mathscr{M}_c$; let $\boldsymbol{T} = \boldsymbol{T}(W) = (\boldsymbol{X}, W)$ be the selected test pattern from the set of template patterns $\mathscr{C}_x$; let $M = M(W) = (I, W) = f(\boldsymbol{T})$ be the compressed, memorized form of $\boldsymbol{T}$; let $\mathscr{C}_u = f(\mathscr{C}_x)$ be the memorized data; let $\boldsymbol{Y}$ be the sensory data; let $J = J(W) = \phi(\boldsymbol{Y})$ be the encoded sensory data,

and let $\widehat{W} = g(J, \mathscr{C}_u)$ be the inferred value of W. Note that M, $\widehat{W}$ are random variables through their dependence on $\boldsymbol{X}$, W, $\boldsymbol{Y}$, and $\mathscr{C}_x$. The mutual informations in the proof are calculated with respect to the joint distribution (and its marginals) over (W, $\mathscr{C}_x$, $\mathscr{C}_u$, $\boldsymbol{X}$, $\boldsymbol{Y}$, M, I, J, $\widehat{W}$). We can verify that this distribution is well defined by writing it out explicitly. Let $\mathbb{T}[\cdot]$ be the truth-indicator function $\mathbb{T}[A] = 1$ if A is true, and $\mathbb{T}[A] = 0$ if A is false. Then

$$\begin{aligned} &p(w, \mathscr{C}_x, \mathscr{C}_u, \boldsymbol{x}, \boldsymbol{y}, m, j) \\ &= p(w)p(\mathscr{C}_x)p(\mathscr{C}_u|\mathscr{C}_x)p(\boldsymbol{x}|w, \mathscr{C}_x) \\ &p(\boldsymbol{y}|\boldsymbol{x})p(m|\boldsymbol{x}, w)p(j|\boldsymbol{y})p(\hat{w}|j, \mathscr{C}_u) \end{aligned}$$

where

$$\begin{aligned} &p(w) = \frac{1}{M_c}\mathbb{T}[w \in \mathscr{M}_c] \\ &p(\mathscr{C}_x) = \prod_{w=1}^{M_c}\prod_{i=1}^{n} p(x_i(w)) \\ &p(\mathscr{C}_u|\mathscr{C}_x) = \mathbb{T}[\mathscr{C}_u = f(\mathscr{C}_x)] \\ &p(\boldsymbol{x}|w, \mathscr{C}_x) = \mathbb{T}[\boldsymbol{x} = \boldsymbol{x}(w) \in \mathscr{C}_x] \\ &p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} p(y_i|x_i) \\ &p(m|\boldsymbol{x}, w) = \mathbb{T}[m = f(\boldsymbol{x}, w)] \\ &p(j|\boldsymbol{y}) = \mathbb{T}[j = \phi(\boldsymbol{y})] \\ &p(\hat{w}|j, \mathscr{C}_u) = \mathbb{T}[\hat{w} = g(j, \mathscr{C}_u)]. \end{aligned}$$

The independence relationships underlying the structure of this distribution are evident from the block diagram of Fig. 1.

Proof. (Theorem 3) Assume $\boldsymbol{R} = (R_x, R_y, R_c) \in \mathscr{R}$. Then there exists a sequence of (Mx, M_y, M_c, n) codes (f, $\boldsymbol{\phi}$, g), such that for any $\epsilon > 0$

$$\begin{aligned} M_c &\geq 2^{nR_c} \\ M_x &\leq 2^{nR_x} \\ M_y &\leq 2^{nR_y} \end{aligned}$$

and $P_e^n = \Pr(\widehat{W} \neq W) \leq \epsilon$. To show that $\boldsymbol{R} \in \mathscr{R}_{\text{out}}$, we must construct a pair of auxiliary random variables UV such that $UV \in \mathscr{P}_{\text{out}}$ and $\boldsymbol{R} \in \mathscr{R}_{UV}$.

We construct the desired pair UV in three steps: 1) We introduce a set of intermediate random variable pairs U_iV_i, $i = 1,2,\ldots,n$, individually contained in $\mathscr{P}_{\text{out}}$; 2) we derive mutual information inequalities for R_x, R_y, and R_c involving sums of the intermediate variables; 3) we convert the sum-inequalities into inequalities in the desired pair UV.

Step 1:

Let the intermediate auxiliary random variables be

$$U_i = \left(M, X^{i-1}\right)$$
$$V_i = \left(J, Y^{i-1}\right)$$

for $i = 1,2,\ldots,n$. Each pair is in $\mathscr{P}_{\text{out}}$. This is verified for the U_i by calculating

$$\begin{aligned} I(U_i; Y_i|X_i) &= H(Y_i|X_i) - H\left(Y_i \middle| M, X^{i-1}, X_i\right) \\ &= H(Y_i|X_i) - H\left(Y_i \middle| M, X^i\right) \\ &\overset{(a)}{\leq} H(Y_i|X_i) - H\left(Y_i \middle| M, X^n\right) \\ &\overset{(b)}{=} H(Y_i|X_i) - H\left(Y_i \middle| X^n\right) \\ &\overset{(c)}{=} H(Y_i|X_i) - H(Y_i|X_i) \\ &= 0 \end{aligned}$$

where the reasons for the lettered steps are (a) conditioning does not increase entropy, (b) the Y_i are independent of all other variables given X^n, and (c) the pairs X_iY_i are i.i.d. Hence, $U_i - X_i - Y_i$ is a Markov chain. By a similar argument, $X_i - Y_i - V_i$ is also a Markov chain. Hence, $U_iV_i \in \mathscr{P}_{\text{out}}$ for each $i = 1,2,\ldots,n$.

Step 2:

First, for the sensory encoder rate

$$\begin{aligned} nR_y &\geq H(J) \\ &\overset{(a)}{=} H(J) - H\left(J \middle| Y^n\right) \\ &= \sum_{i=1}^{n} H(Y_i) - H\left(Y_i \middle| Y^{i-1}J\right) \\ &= \sum_{i=1}^{n} H(X_i) - H(X_i|V_i) \\ &= \sum_{i=1}^{n} I(X_i; V_i) \end{aligned}$$

where (a) follows from $J = \phi(Y^n)$.

Next, taking account of all M_c memorized patterns, for the memory encoder rate we have

$$
\begin{aligned}
M_c(nR_X) &= M_c \log M_X \\
&\geq H(\mathscr{C}_u) \\
&= H(\mathscr{C}_u) - H(\mathscr{C}_u|\mathscr{C}_X) \\
&= H(\mathscr{C}_X) - H(\mathscr{C}_X|\mathscr{C}_u) \\
&= \sum_{w=1}^{M_c} H(T(w)) - H(T(w)|M(w)) \\
&\overset{(a)}{=} \sum_{w=1}^{M_c} H\left(X^n(W)|W=w\right) - H\left(X^n(W)|I, W=w\right) \\
&\overset{(b)}{=} \sum_{w=1}^{M_c} H\left(X^n\right) - H\left(X^n|I, W=w\right) \\
&\overset{(c)}{=} M_c \sum_{w=1}^{M_c} p(w)\left[H\left(X^n\right) - H\left(X^n|I, W=w\right)\right] \\
&\overset{(d)}{=} M_c\left[H\left(X^n\right) - H\left(X^n|I, W\right)\right] \\
&\overset{(e)}{=} M_c \sum_{i=1}^{n}\left[H(X_i) - H\left(X_i|X^{i-1}, I, W\right)\right] \\
&\overset{(f)}{=} M_c \sum_{i=1}^{n}\left[H(X_i) - H(X_i|U_i)\right] \\
&= M_c \sum_{i=1}^{n} I(X_i; U_i)
\end{aligned}
$$

where (a) is simply a matter of variable definitions and notation, (b) follows from the assumption that the $X^n(W)$'s all have the same distribution and are drawn independently of W, (c) follows from the definition of $p(w)$, (d) follows from the definition of conditional entropy, (e) follows from $p(x^n) = \prod_{i=1}^{n} p(x_i)$ and the telescoping property, and (f) follows from the definition of U_i. Hence, $nR_X \geq \sum_{i=1}^{n} I(X_i; U_i)$.

$$
\begin{aligned}
&\text{Finally } nR_c \leq \log M_c \\
&= H(W) \\
&= I(W; \mathscr{C}_u, J) + H(W|\mathscr{C}_u, J) \\
&\overset{(a)}{\leq} I(W; \mathscr{C}_u, J) + n\epsilon_n \\
&= I(W; \mathscr{C}_u) + I(W; J|\mathscr{C}_u) + n\epsilon_n \\
&\overset{(b)}{=} 0 + I(W; J|\mathscr{C}_u) + n\epsilon_n \\
&= I(W, \mathscr{C}_u; J) - I(J; \mathscr{C}_u) + n\epsilon_n \\
&\leq I(W, \mathscr{C}_u; J) + n\epsilon_n \\
&\overset{(c)}{=} I(M; J) + n\epsilon_n \\
&\overset{(d)}{=} \sum_{i=1}^{n} I(X_i; U_i) + I(Y_i; V_i) - I(X_iY_i; U_iV_i) + n\epsilon_n .
\end{aligned}
$$

The lettered steps are justified as follows.

a. By assumption, $\Pr(\widehat{w} \neq W) = P_e^n \to 0$, where $\widehat{W} = g(J, \mathscr{C}_u)$. Thus, applying Fano's inequality yields

$$H(W|\mathscr{C}_u, J) \leq H\left(P_e^n\right) + P_e^n \log(M_c - 1) \leq n\epsilon_n$$

where $\epsilon_n \to 0$.

b. The test index W and patterns $\mathscr{C}_x$ are drawn independently, hence, $I(W; \mathscr{C}_u) = I(W; f(\mathscr{C}_x)) = 0$.

c. Writing $\mathscr{C}_u = \mathscr{C}_u^* \cup M$, $\mathscr{C}_u^* = \mathscr{C}_u \backslash M$, we have

$$\begin{aligned} I(W, \mathscr{C}_u; J) &= I(W, M, \mathscr{C}_u^*; J) \\ &= I(M; J) + I(W, \mathscr{C}_u^*; J \mid M) \\ &= I(M; J) + I(W, \mathscr{C}_u^*; J \mid I, W) \\ &= I(M; J) + I(\mathscr{C}_u^*; J \mid I, W) \\ &= I(M; J) + 0, \end{aligned}$$

since the $M(i) = (I(i), i)$ are independent of J for $i \neq W$.

d. To justify this step, we invoke the following two results, proved in Appendix D. Let A, α, B, β and γ be arbitrary discrete random variables.

Then we get the following.

Lemma 6:

$$I(\alpha; \beta) \geq I(A; \alpha) + I(B; \beta) - I(AB; \alpha\beta),$$

with equality if and only if $I(A\alpha; B\beta) = I(A; B)$.

Lemma 7: Let $Z_i = (\gamma, A^{i-1})$, $i = 1,2,\ldots,n$ where the A_i are i.i.d. Then

$$\sum_{i=1}^{n} I(A_i; Z_i) = I(A^n; \gamma).$$

To apply Lemma 6, make the substitution $(\alpha, \beta, A, B) \to (M, J, X^n, Y^n)$. Then the condition for equality is satisfied

$$\begin{aligned} & I(X^n, M; Y^n, J) \\ &= I(X^n, I, W; Y^n, J) \\ &= I(X^n, W; Y^n, J) + I(I; Y^n, J \mid X^n, W) \\ &= I(X^n, W; Y^n, J) + I(I, W; Y^n, J \mid X^n, W) \\ &\overset{(a)}{=} I(X^n, W; Y^n, J) + 0 \\ &= I(X^n, W; Y^n) + I(X^n, W; J \mid Y^n) \\ &\overset{(b)}{=} I(X^n, W; Y^n) + 0 \\ &= I(X^n; Y^n) + I(W; Y^n \mid X^n) \\ &\overset{(c)}{=} I(X^n; Y^n) + 0 \end{aligned}$$

since (a) $M = (I, W) = f(X^n, W)$, (b) $J = \phi(Y^n)$, and (c) Y^n only depends on W through $X^n = X^n(W)$, so that $H(Y^n \mid X^n, W) = H(Y^n \mid X^n)$. Thus, we have

$$I(M;J) = I(X^n;M) + I(Y^n;J) - I(X^n,Y^n;M,J). \tag{23}$$

Next, apply Lemma 7 three times with the substitutions

$$\begin{aligned}&(Z_i,\gamma,A^{i-1}) \to (U_i,M,X^{i-1}),\\ &\to (V_i,J,Y^{i-1}),\\ &\to (U_iV_i,MJ,X^{i-1}Y^{i-1})\end{aligned}$$

to obtain

$$\begin{aligned}\sum_{i=1}^{n} I(X_i;U_i) &= I(X^n;M)\\ \sum_{i=1}^{n} I(Y_i;V_i) &= I(Y^n;J)\\ \sum_{i=1}^{n} I(X_iY_i;U_iV_i) &= I(X^nY^n;MJ).\end{aligned}$$

Adding the first two expressions and subtracting the third yields

$$\begin{aligned}&\sum_{i=1}^{n} [I(\boldsymbol{X}_i,\boldsymbol{U}_i) + I(\boldsymbol{Y}_i;\boldsymbol{V}_i) - I(\boldsymbol{X}_i,\boldsymbol{Y}_i;\boldsymbol{U}_i,\boldsymbol{V}_i)]\\ &= [I(X^n;M) + I(Y^n;J) - I(X^nY^n;MJ)].\end{aligned} \tag{24}$$

Combining (23) and (24) yields

$$I(M;J) = \sum_{i=1}^{n} I(X_i;U_i) + I(Y_i;V_i) - I(X_i,Y_i;U_i,V_i)$$

as claimed.

Step 3:

For this step, we use the following lemma, proved in Appendix C as part of the demonstration that $\mathscr{R}_{\text{out}}$ is convex.

Lemma 1: Suppose $U_iV_i \in \mathscr{P}_{\text{out}}$, $i = 1,2,\ldots,n$. Then there exists $UV \in \mathscr{P}_{\text{out}}$ such that

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i;U_i) = I(X;U)$$
$$\frac{1}{n}\sum_{i=1}^{n} I(Y_i;V_i) = I(Y;V)$$
$$\frac{1}{n}\sum_{i=1}^{n} I(X_iY_i;U_iV_i) = I(XY;UV).$$

Applying Lemma 1 to the results of Steps 1 and 2, we obtain

$$R_x \geq \frac{1}{n}\sum_{i=1}^{n} I(X_i;U_i) = I(X;U)$$
$$R_y \geq \frac{1}{n}\sum_{i=1}^{n} I(Y_i;V_i) = I(Y;V)$$
$$R_c \leq R_x + R_y - \frac{1}{n}\sum_{i=1}^{n} I(X_iY_i;U_iV_i)$$
$$= R_x + R_y - I(XY;UV)$$

where $UV \in \mathscr{P}_{\text{out}}$. With respect to this UV, by definition we have $\boldsymbol{R} \in \mathscr{R}_{UV}$. Hence, $\boldsymbol{R} \in \mathscr{R}_{\text{out}}$, and the proof is complete. □

Appendix C: Convexity of the Outer Bound

In this appendix, we prove a slightly more general version of Lemma 1 from Appendix B, and use this result to show that the outer bound rate region $\mathscr{R}_{\text{out}}$ is convex.

In the following, let $\mathcal{Q}$ be any finite alphabet, and assume that we have pairs X_qY_q for all $q \in \mathcal{Q}$ which are i.i.d. ~ $p(xy)$.

Lemma 2: Suppose $U_qV_q \in \mathscr{P}_{\text{out}}$ for all $q \in \mathcal{Q}$, and let Q~$p(q)$, $q \in \mathcal{Q}$ be any discrete random variable independent of the pairs $\{X_qY_q\}$. Then there exists a pair of discrete random variables $UV \in \mathscr{P}_{\text{out}}$ such that

$$\sum_{q\in\mathcal{Q}} p(q) I\left(X_q;U_q\right) = I(X;U)$$
$$\sum_{q\in\mathcal{Q}} p(q) I\left(Y_q;V_q\right) = I(Y;V)$$
$$\sum_{q\in\mathcal{Q}} p(q)\left[I\left(X_q;U_q\right) + I\left(Y_q;V_q\right) - I\left(X_qY_q;U_qV_q\right)\right]$$
$$= I(X;U) + I(Y;V) - I(XY;UV).$$

Remark 4: Lemma 1 in Appendix B follows immediately from the above Lemma, by choosing $\mathcal{Q} = \{1, 2, \ldots, n\}$ and $p(q) = 1/n$ for all $q \in \mathcal{Q}$.

Proof. As a candidate for the pair UV in the lemma, consider $UV \in \mathscr{P}_{\text{conv}}$ for the given Q (see (4)), i.e., $U = (U_Q, Q)$ and $V = (V_Q, Q)$. To verify that $UV \in \mathscr{P}_{\text{out}}$, we proceed to check that $U - X - Y$ and $X - Y - V$ are Markov chains.

By the assumption $U_q V_q \in \mathscr{P}_{\text{out}}$ for each $q \in \mathcal{Q}$, we have $I(U_q; Y_q | X_q) = 0$ and $I(V_q; X_q | Y_q) = 0$. Hence

$$\begin{aligned} 0 &= \sum_{q \in \mathcal{Q}} p(q) I\left(U_q; Y_q | X_q\right) \\ &= \sum_{q \in \mathcal{Q}} p(q) I\left(U_q; Y_q | X_q, Q = q\right) \\ &= I\left(U_Q; Y_Q | X_Q Q\right) \\ &\overset{(a)}{=} I\left(U_Q; Y | X, Q\right) \\ &= I\left(U_Q Q; Y | X\right) - I(Q; Y | X) \\ &\overset{(b)}{=} I\left(U_Q Q; Y | X\right) \\ &= I(U; Y | X) \end{aligned}$$

where in (a) we are able to drop the subscript Q on X_Q and Y_Q because the X_q and Y_q are i.i.d. and independent of Q; and similarly (b) is because $I(Q; Y | X) = 0$, due to the independence of Q and Y. By an analogous calculation, we also find $I(V; X | Y) = 0$. Hence, $U - X - Y$ and $X - Y - V$, and $UV \in \mathscr{P}_{\text{out}}$ as desired.

It remains to demonstrate the three equalities in the lemma. For the first, write

$$\begin{aligned} I(X; U) &= I\left(X; U_Q Q\right) \\ &= I\left(X; U_Q \middle| Q\right) + I(X; Q) \\ &\overset{(a)}{=} I\left(X; U_Q \middle| Q\right) \\ &\overset{(b)}{=} I\left(X_Q; U_Q \middle| Q\right) \\ &= \sum_{q \in \mathcal{Q}} p(q) I\left(X_q; U_q\right) \end{aligned}$$

where, as above, (a) and (b) follow from the fact that the X_q are i.i.d. and independent of Q. Similar calculations yield

$$I(Y; V) = \sum_{q \in \mathcal{Q}} p(q) I\left(Y_q; V_q\right),$$

and

$$I(XY; UV) = \sum_{q \in \mathcal{Q}} p(q) I\left(X_q Y_q; U_q V_q\right).$$

The convexity of $\mathscr{R}_{\text{out}}$ follows readily from the preceding lemma.

Lemma 3: $\mathscr{R}_{\text{out}}$ is convex. That is, let $\boldsymbol{R}_q$ be any set of rates such that $\boldsymbol{R}_q \in \mathscr{R}_{\text{out}}$ for all $q \in \mathcal{Q}$, where $\mathcal{Q}$ is a finite alphabet, and let $p(q)$ be any probability distribution over $\mathcal{Q}$. Then $\boldsymbol{R} = \sum_{q \in \mathcal{Q}} p(q) \boldsymbol{R}_q \in \mathscr{R}_{\text{out}}$.

Proof: Fix an arbitrary distribution $p(q)$ and rates $\boldsymbol{R}_q \in \mathscr{R}_{\text{out}}$ for all $q \in \mathcal{Q}$. By the definition of $\mathscr{R}_{\text{out}}$, for each rate $\boldsymbol{R}_q$, there exists a pair $U_q V_q \in \mathscr{P}_{\text{out}}$ such that $\boldsymbol{R}_q \in \mathscr{R}_{U_q V_q}$. Consequently

$$\begin{aligned} R_x &= \sum_{q \in \mathcal{Q}} p(q) R_{x,q} \geq \sum_{q \in \mathcal{Q}} p(q) I\left(X_q; U_q\right) \\ R_y &= \sum_{q \in \mathcal{Q}} p(q) R_{y,q} \geq \sum_{q \in \mathcal{Q}} p(q) I\left(Y_q; V_q\right) \\ R_c &= \sum_{q \in \mathcal{Q}} p(q) R_{c,q} \\ &\leq \sum_{q \in \mathcal{Q}} p(q)\left[I\left(X_q; U_q\right) + I\left(Y_q; V_q\right) - I\left(X_q Y_q; U_q V_q\right)\right]. \end{aligned}$$

As in the proof of Lemma 2, use these pairs to construct a new pair UV, by defining U = (UQ, Q), V = (VQ, Q). From the proof of Lemma 2, we know 1) that $UV \in \mathscr{P}_{\text{out}}$, and 2) the sums on the right-hand sides of the inequalities above can be replaced with expressions in U and V, yielding

$$\begin{aligned} R_x &\geq I(X;U) \\ R_y &\geq I(Y;V) \\ R_c &\leq I(X;U) + I(Y;V) - I(XY;UV) \\ &\leq R_x + R_y - I(XY;UV) \end{aligned}$$

which means that $R \in \mathscr{R}_{UV}$ for the given UV. Hence, $\boldsymbol{R} = \sum_{q \in \mathcal{Q}} p(q) \boldsymbol{R}_q \in \mathscr{R}_{\text{out}}$. Since p(q) and $\boldsymbol{R}_q \in \mathscr{R}_{\text{out}}$ were arbitrary, we conclude that $\mathscr{R}_{\text{out}}$ is convex. □

Appendix D: Mixing Lemmas

In this appendix, we prove Lemmas 6 and 7, which are used in proving the outer bound.

Consider the elementary Shannon inequalities, stated in the following two lemmas. The variables $A, B, \alpha, \beta, \gamma, \delta$ appearing in the lemmas denote arbitrary discrete random variables.

Lemma 4:

$$I(A;\alpha) = I(A;\alpha,\gamma) - I(A,\alpha;\gamma) + I(\alpha;\gamma).$$

Proof:

$$\begin{aligned} I(A;\gamma|\alpha) &= I(A;\alpha,\gamma) - I(A;\alpha) \\ &= I(A,\alpha;\gamma) - I(\gamma;\alpha). \end{aligned}$$

Lemma 5:

$$I(A;\alpha) + I(B;\beta) = I(A;B) + I(\alpha;\beta) - I(A,\alpha;B,\beta) + I(A,B;\alpha,\beta).$$

Proof:

$$\begin{aligned} &I(A,\alpha;B,\beta)-I(A,B;\alpha,\beta)\\ &=H(A,\alpha)+H(B,\beta)-H(A,B)-H(\alpha,\beta)\\ &=-I(A;\alpha)-I(B;\beta)+I(A;B)+I(\alpha;\beta). \end{aligned}$$

Lemma 6 follows directly from the preceding lemmas.

Lemma 6:

$$I(\alpha;\beta)\geq I(A;\alpha)+I(B;\beta)-I(A,B;\alpha,\beta)$$

with equality if and only if $I(A, \alpha; B, \beta) = I(A; B)$.

Proof: Rearrange Lemma 5 to get

$$I(\alpha;\beta)=I(A;\alpha)+I(B;\beta)-I(A,B;\alpha,\beta)+[I(A,\alpha;B,\beta)-I(A;B)].$$

The lemma now follows readily from the preceding expression: We obtain equality in the lemma if (and only if) the term in brackets is zero. Otherwise, the bracketed term is nonnegative, since

$$\begin{aligned} &I(A,\alpha;B,\beta)-I(A;B)\\ &=H(\alpha|A)+H(\beta|B)-H(\alpha,\beta|A,B)\\ &=H(\alpha|A)-H(\alpha|A,B)+H(\beta|B)-H(\beta|A,B,\alpha)\\ &\geq 0, \end{aligned}$$

where the inequality is due to the fact that conditioning does not increase entropy. □

Lemma 7: If $U_i = (\gamma, A^{i-1})$, then

$$I(A^n;\gamma)=\sum_{i=1}^{n}I(A_i;U_i)-\sum_{i=2}^{n}I\left(A_i;A^{i-1}\right),$$

Proof: In Lemma 4, put $A = A_i$, $\alpha = A^{i-1}$ Note that $U_1 = \gamma$. Hence, substituting and summing from 2 to n yields

$$\begin{aligned} &\sum_{i=2}^{n}I\left(A_i;A^{i-1}\right)\\ &=\sum_{i=2}^{n}I(A_i;U_i)-I\left(A^n;\gamma\right)+I(A_1;\gamma)\\ &=\sum_{i=2}^{n}I(A_i;U_i)-I\left(A^n;\gamma\right)+I(A_1;U_1)\\ &=\sum_{i=1}^{n}I(A_i;U_i)-I\left(A^n;\gamma\right). \end{aligned}$$

Appendix E: Alternative Representations of Rin and Rout

Here we show that the alternative representations of the inner and outer bound surfaces introduced in Remark 2 are in fact equivalent, i.e., $\mathscr{R}_{\text{in}} = \mathscr{R}'_{\text{in}} = \mathscr{R}''_{\text{in}}$, $\mathscr{R}_{\text{out}} = \mathscr{R}'_{\text{out}} = \mathscr{R}''_{\text{out}}$.

We show first that $\mathscr{R}'_{\text{out}} = \mathscr{R}''_{\text{out}}$. Suppose $\boldsymbol{R} \in \mathscr{R}'_{\text{out}}$. Then, using $U - X - Y$, $X - Y - V$ we have

$$\begin{aligned} R_c &\le I(U;V) - I(U;V|XY) \\ &= I(X;U) + I(Y;V) - I(XY;UV) \end{aligned} \tag{25}$$

which implies

$$\begin{aligned} R_X &\ge I(X;U) \\ &\ge I(XY;UV) - I(Y;V) + R_c \\ &= I(XY;U|V) + R_c \end{aligned}$$

and similarly $R_Y \ge I(XY; V|U) + R_c$; and

$$\begin{aligned} R_X + R_Y &\ge I(X;U) + I(Y;V) \\ &\ge I(XY;UV) + R_c . \end{aligned}$$

We conclude that $\boldsymbol{R} \in \mathscr{R}''_{\text{out}}$, hence $\mathscr{R}'_{\text{out}} \subseteq \mathscr{R}''_{\text{out}}$. A symmetrical argument shows $\mathscr{R}''_{\text{out}} \subseteq \mathscr{R}'_{\text{out}}$, proving $\mathscr{R}'_{\text{out}} = \mathscr{R}''_{\text{out}}$.

Next, we show that $\mathscr{R}_{\text{out}}$ and $\mathscr{R}'_{\text{out}}$ are identical. To this end, note that these sets correspond to regions in the positive orthant $\mathbb{R}^3_+$, and that two such regions are identical if they have the same surfaces. Following the presentation in Section VII, the surfaces of $\mathscr{R}_{\text{out}}$ and $\mathscr{R}'_{\text{out}}$ are

$$\begin{aligned} r_{\text{out}}(r_X, r_Y) &= \max_{UV \in \mathscr{C}_{\text{out}}(r_X, r_Y)} r_X + r_Y - I(XY;UV) \\ r'_{\text{out}}(r_X, r_Y) &= \max_{UV \in C_{\text{out}}(r_X, r_Y)} I(U;V) - I(U;V|XY) \end{aligned}$$

where

$$\mathscr{C}_{\text{out}}(r_X, r_Y) = \left\{ UV \in \mathscr{P}_{\text{out}} : I(U;X) = r_X, I(V;Y)) = r_Y \right\}.$$

$$\begin{aligned} & r'_{\text{out}}(r_X, r_Y) \\ &= \max_{UV \in \mathscr{C}_{\text{out}}(r_X, r_Y)} I(X;U) + I(Y;V) - I(XY;UV) \\ &= \max_{UV \in \mathscr{C}_{\text{in}}(r_X, r_Y)} r_X + r_Y - I(XY;UV) \\ &= r_{\text{out}}(r_X, r_Y) . \end{aligned}$$

Thus, the desired equivalence $\mathscr{R}_{\text{out}} = \mathscr{R}'_{\text{out}}$ follows simply from the fact that at the surfaces, the inequalities defining each region become equalities.

The same line of argument as above of course also shows $\mathscr{R}_{\text{in}} = \mathscr{R}'_{\text{in}} = \mathscr{R}''_{\text{in}}$.

Appendix F: Simplification of Convex Hulls

In this appendix, we argue geometrically that the expressions for the convex hulls of the inner bound regions simplify to just one term in both the binary and Gaussian cases. To discuss both cases simultaneously, let us represent the surface of either inner bound by a positive-valued function $f : \mathscr{D} \to \mathbb{R}_+$. Here, $\mathscr{D}$ is a square region

$$\mathscr{D} = \left\{ r = (x, y) \in \mathbb{R}^2 : 0 \le x \le M, 0 \le y \le M \right\}$$

and M is a positive constant. In the binary case, $f(r) = s(r)$, and $\mathscr{D} = [0, 1] \times [0, 1]$; in the Gaussian case, $f(r) = S(r)$ and $\mathscr{D} = [0, \infty) \times [0, \infty)$. Some important properties shared by both cases are that for all $\mathscr{D} = [0, \infty) \times [0, \infty)$

$$\begin{gathered} f(x, y) \ge 0, f(0, y) = f(x, 0) = 0 \\ f_x(r), f_y(r) > 0, f_{xx}(r), f_{yy}(r) < 0 \end{gathered}$$

where the subscripts denote partial derivatives.

Denote the convex hull of $f(r)$ by $c(r)$. Generically, the boundary of the convex hull is

$$c(r) = \max \theta f(r_1) + \bar{\theta} f(r_2)$$

where the maximum is over all triples (θ, r_1, r_2) such that $r = \theta r_1 + \bar{\theta} r_2, \theta \in [0, 1]$, and $r_1, r_2 \in \mathscr{D}$. However, as argued next, for the cases under study this simplifies to

$$c(r) = \max \theta f(r')$$

where $r = \theta r'$.

The convex hull of a surface can be characterized in terms of its tangent planes. Given any point $r' = (x, y) \in \mathscr{D}$, if its tangent plane lies entirely above the surface, then $(r', f(r'))$ is on the convex hull. If the tangent plane cuts *through* the surface at one or more other points, or if the tangent plane lies below the surface, then $(r, f(r))$ is not on the convex hull. If the tangent plane intersects the surface at exactly two points, then both points are on the convex hull.

The tangent plane at an arbitrary point $r' = (x', y') \in \mathscr{D}$ is the set of points satisfying

$$z(x, y) = f_x(x - x') + f_y(y - y') + z'$$

where the partial derivatives are evaluated at r', i.e., $f_x = f_x(r')$, $f_y = f_y(r')$, and $z' = f(r')$. The tangent plane intersects the $z = 0$ plane in a line. Setting $z(r) = 0$ and solving

$$\begin{aligned} &y = mx + b, \text{ where} \\ &m = -\left(f_x / f_y\right) \\ &b = 1/f_y\left[x' f_x + y' f_y - z'\right]. \end{aligned}$$

Since $f_x, f_y > 0$, the slope $m = -(f_x/f_y)$ is negative. This line intersects the positive orthant whenever the intercept b ≥ 0, in which case the tangent plane cuts through the surface, since $f \geq 0$. Thus, the only points on the original surface $f(x,y)$ that can be on the convex hull are those for which b ≤ 0.

Next, consider any path through $\mathcal{D}$ along a line segment $y = ax$, $a > 0$, starting from one of the "outer edges" of $\mathcal{D}$, where $x = M$ or $y = M$, and consider what happens to the tangent plane's line of intersection ℓ with the $z = 0$ plane as we move in along the path toward the origin (0, 0). Initially, the tangent planes lie entirely above the surface, and the intercept of ℓ is negative, $b < 0$. This intercept increases along the path until $b = 0$, at which point ℓ intersects (0, 0). Here, the tangent plane contains a line segment attached on one end to the point of tangency, and at the other end to the point $(r, f(r)) = (0, 0, 0)$; everywhere else, the tangent plane is above the surface. Continuing toward the origin, all other points along the path have tangent planes such that ℓ has a positive intercept $b > 0$, hence, these points are excluded from the convex hull.

These considerations imply that the convex hull $c(r)$ is composed entirely of two kinds of points. First, points which coincide with the original surface, $c(r), = \theta f(r)$ with $\theta = 1$. These points occur at values of $r = (x, y)$ "up and to the right" of (0, 0). Second, points along line segments connecting surface points "up and to the right" $(r', f(r'))$ with the point $(r, f(r)) = (0, 0, 0)$, that is $c(r) = \theta f(r') + \bar{\theta} f(0,0) = \theta f(r')$, where $r = \theta r'$ and $\theta \in [0, 1]$. Hence, for all $r \in \mathcal{D}$, $c(r)$ has the desired form.

Two more examples of functions that behave in the same way just described are $f(x, y) = (1 - (1 - x)^2)(1 - (1 - y)^2)$ and $f(x, y) = xy$, with $\mathcal{D} = [0, 1] \times [0, 1]$.

Appendix G: Properties of Gaussian Mutual Information

Our analysis of the Gaussian pattern recognition problem relies on the following well-known results, stated below without proof.

Lemma 8: The mutual information between two Gaussian random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ depends only on the matrices of correlation coefficients. Specifically

$$I(\boldsymbol{X}; \boldsymbol{Y}) = \frac{1}{2}\log\left(\det \boldsymbol{C}_{x,x}\right) - \frac{1}{2}\log\left(\det \boldsymbol{C}_{x,x|y}\right)$$

where

$$\boldsymbol{C}_{x,x|y} = \boldsymbol{C}_{x,x} - \boldsymbol{C}_{x,y}\boldsymbol{C}_{y,y}^{-1}\boldsymbol{C}_{y,x}.$$

In the special case $Y = X + W$, where X and W are independent Gaussian random variables with variances P and N, respectively, we have

$$I(X;Y) = \frac{1}{2}\log\left(1 + \frac{P}{N}\right) = -\frac{1}{2}\log\left(1 - \rho_{x,y}^2\right)$$

where the correlation coefficient $\rho_{x,y} = \sqrt{P/(P+N)}$.

Lemma 9: If X, Y and Z are zero-mean Gaussian random vectors that form a Markov chain $X - Y - Z$, then

$$\boldsymbol{C}_{x,z} = \boldsymbol{C}_{x,y}\boldsymbol{C}_{y,y}^{-1}\boldsymbol{C}_{y,z}.$$

Note that for dimension one, $X \to Y \to Z$ $\rho_{x,z} = \rho_{x,y}\rho_{y,z}$ implies.

Lemma 10: Let X, Y, U, and V be jointly Gaussian random variables such that $U - X - Y$ and $X - Y - V$ are Markov chains. Then the matrix of correlation coefficients $C_{xy,uv}$ decomposes as

$$C_{xy,uv} = \begin{bmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{bmatrix} \begin{bmatrix} \rho_{xu} & 0 \\ 0 & \rho_{yv} \end{bmatrix}.$$

This lemma follows immediately by using Lemma 9 to obtain the substitutions $\boldsymbol{C}_{x,v} = \boldsymbol{C}_{x,y}\boldsymbol{C}_{y,y}^{-1}\boldsymbol{C}_{y,v} = \rho_{xy}\rho_{yv}$ and $\boldsymbol{C}_{x,y} = \boldsymbol{C}_{u,x}\boldsymbol{C}_{x,x}^{-1}\boldsymbol{C}_{x,y} = \rho_{ux}\rho_{xy}$.

References

[1]. Barlow HB, "Sensory mechanisms, the reduction of redundancy, and intelligence," in Proc. Symp. Mechanization of Thought Processes, 1959, pp. 537–574.

[2]. Van Essen DC and Anderson CH, "Information processing strategies and pathways in the primate retina and visual cortex," in An Introduction to Neural and Electronic Networks, Lau C, Zornetzer SF, and Davis JL, Eds. San Diego, CA: Academic, 1995, pp. 43–72.

[3]. Eddington AS, The Mathematical Theory of Relativity, 3rd ed Cambridge, U.K.: Cambridge Univ. Press, 1963.

[4]. Grenander U, Elements of Pattern Theory, ser Johns Hopkins Series in the Mathematical Sciences. Baltimore, MD: Johns Hopkins Univ. Press, 1996.

[5]. Mumford D, "Pattern theory: A unifying perspective," in Perception as Bayesian Inference, Knill DC and Richards W, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1996, ch. 1, pp. 25–62.

[6]. Schmid JA and O'Sullivan NA, "Performance prediction methodology for biometric systems using a large deviations approach," IEEE Trans. Signal Process, vol. 52, no. 10, pp. 3036–3045, Oct..

[7]. Bishop CM, Pattern Recognition and Machine Learning. New York: Springer, 2006.

[8]. MacKay DJC, Information Theory, Inference, and Learning Algorithms. Cambridge, U.K.: Cambridge Univ. Press, 2003 [Online]. Available: http://www.inference.phy.cam.ac.uk/mackay/itila/

[9]. Kearns MJ and Vazirani UV, An Introduction to Computational Learning Theory. Cambridge, MA: MIT Press, 1994.

[10]. Vapnik VN, Statistical Learning Theory. New York: Wiley, 1998.

[11]. Grenander U and Miller M, Pattern Theory: From Representation to Inference. New York: Oxford Univ. Press, 2006.

[12]. Burges CJC, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121–167, 1998.

[13]. Barlow HB, "The coding of sensory messages," in Current Problems in Animal Behavior, Thorpe WH and Zangwill OL, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1961.

[14]. Barlow HB, "Understanding natural vision," in Physical and Biological Processing of Images, ser. Springer series in Information Sceinces, Braddick OJ and Sleigh AC, Eds. Berlin, Germany: Springer-Verlag, 1983, vol. 11, pp. 2–14.

[15]. Barlow H, "What is the computational goal of the neocortex?," in Large Scale Neuronal Theories of the Brain, Koch C and Davis JL, Eds. Cambridge, MA: MIT Press, 1994, ch. 1, pp. 1–22.

[16]. Barlow H, "Redundancy reduction revisited," Network-Computa. Neural Syst, vol. 12, no. 3, pp. 241–53, 2001.

[17]. Gardner-Medwin AR and Barlow HB, "The limits of counting accuracy in distributed neural representations," Neural Comput, vol. 13, no. 3, pp. 477–504, 2001. [PubMed: 11244552]

[18]. Rieke F, Warland D, de Ruyter van Steveninck R, and Bialek W, Spikes: Exploring the Neural Code. Cambridge, MA: MIT Press, 1996.

[19]. Lennie P, "The cost of cortical computation," Curr. Biol, vol. 13, no. 6, pp. 493–497, 2003. [PubMed: 12646132]

[20]. LeCun Y, Bottou L, Bengio Y, and Haffner P, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[21]. Jain AK, Duin RPW, and Mao J, "Statistical pattern recognition: A review," IEEE Trans. Pattern Anal. Machine Intell, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[22]. Ahlswede RF and Csiszár I, "Hypothesis testing with communication constraints," IEEE Trans. Inf. Theory, vol. IT-32, no. 4, pp. 533–542, Jul. 1986.

[23]. Han TS and Amari SI, "Statistical inference under multiterminal data compression," IEEE Trans. Inf. Theory, vol. 44, no. 6, pp. 2300–2324, Oct. 1998.

[24]. Ishwar P, Prabhakaran VM, and Ramchandran K, "Toward a theory for video coding using distributed compression principles," in Proc. Int. Conf. Image Processing (ICIP 2), Barcelona, Spain, Sept. 2003, pp. 687–690.

[25]. Ishwar P, Prabhakaran VM, and Ramchandran K, "On joint classification and compression in a distributed source coding framework," in Proc. IEEE Workshop on Statistical Signal Processing, St. Louis, MO, 2003, pp. 25–28.

[26]. Oehler KL and Gray RM, "Combining image compression and classification using vector quantization," IEEE Trans. Pattern Anal. Machine Intell, vol. 17, no. 5, pp. 461–473, 5 1995.

[27]. Hinton GE and Salakhutdinov RR, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, Jul. 2006. [PubMed: 16873662]

[28]. Dong Y, Chang S, and Carin L, "Rate-distortion bound for joint compression and classification with application to multi-aspect sensing," IEEE Sensors J, vol. 5, no. 3, pp. 481–492, Jun. 2005.

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

[30]. Cover TM and Thomas JA, Elements of Information Theory. New York: Wiley, 1990.

[31]. Berger T, "Multiterminal source coding," in The Information Theory Approach to Communications, Longo G, Ed. New York: Springer-Verlag, 1977.

[32]. Tung SY, "Multiterminal Source Coding," Ph.D. dissertation, Cornell Univ., Ithaca, NY, 1978.

[33]. Berger T, The Information Theory Approach to Communications. New York: Springer-Verlag, 1977, ch. Multiterminal Source Coding.

[34]. Shannon CE, "A mathematical theory of communication," Bell Syst. Tech. J, vol. 27, pp. 379–423, 623–656, 1948.

[35]. Westover MB, "Image Representation and Pattern Recognition in Brains and Machines," Ph.D. dissertation, Washington Univ., St.Loouis, MO, 2004.

[36]. Olshausen BA and Field DJ, "Sparse coding with and overcomplete basis set: A strategy employed by {V}1?," Vision Res, vol. 37, pp. 3311–3325, 1997. [PubMed: 9425546]

[37]. Sejnowski TJ and Bell AJ, "The "independent components" of natural scenes are edge filters," Vision Res, vol. 37, pp. 3327–3338, 1997. [PubMed: 9425547]

[38]. Hyvärinen A, "Survey on independent component analysis," Neural Comput. Sur, vol. 2, pp. 94–128, 1999.

[39]. Schwartz O and Simoncelli EP, "Natural signal statistics and sensory gain control," Nature Neurosc, vol. 4, no. 8, pp. 819–825, Aug. 2001.

[40]. O'Sullivan JA, Singla N, and Westover MB, "Successive refinement for pattern recognition," in Proc. IEEE Information Theory Workshop, Punta del Este, Uruguay, 2006, pp. 141–145.

[41]. Srivastava A, Lee AB, Simoncelli EP, and Zhu SC, "On advances in statistical modeling of natural images," J. Math. Imaging and Vision, vol. 18, no. 1, pp. 17–33, Jan. 2003.

[42]. Lee AB, Mumford D, and Huang J, "Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model," Int.J. Comp. Vision, vol. 41, pp. 35–59, 2001.

[43]. Zhu SC, "Statistical modeling and conceptualization of visual patterns," IEEE Trans. Pattern Anal. Machine Intell, vol. 25, no. 6, pp. 691–712, Jun. 2003.

[44]. Heeger DJ and Bergen JR, "Pyramid-based texture analysis/synthesis," in Proc. 22nd Annu. Conf. Computer Graphics and Interactive Techniques, 1995, pp. 229–238, ACM Press.

[45]. Zhu SC, Wu YN, and Mumford D, "Filters, random fields and maximum entropy (frame): Toward a unified theory for texture modeling," Int. J. Comp. Vision, vol. 27, no. 2, pp. 107–126, 1998.

[46]. De Bonet JS and Viola PA, "A nonparametric multi-scale statistical model for natural images," in Advances in Neural Information Processing Systems, Jordan MI, Kearns MJ, and Solla SA, Eds. Cambridge, MA: MIT Press, 1998, vol. 10.

[47]. Portilla J and Simoncelli EP, "A parametric texture model based on joint statistics of complex wavelet coefficients," Int. J. Comp. Vision, vol. 40, no. 1, pp. 49–71, 2000.

[48]. Jelinek F, Statistical Methods for Speech Recognition. Cambridge, MA: MIT Press, 1999.

[49]. Blanchard G and Geman D, "Hierarchical testing designs for pattern recognition," Ann. Statist, vol. 33, pp. 1155–1202, 2005.

[50]. Servetto S, Multiterminal Source Coding With Two Encoders—I: A Computable Outer Bound Apr. 2006 [Online]. Available: http://www.arxiv.org/abs/cs.IT/0604005/

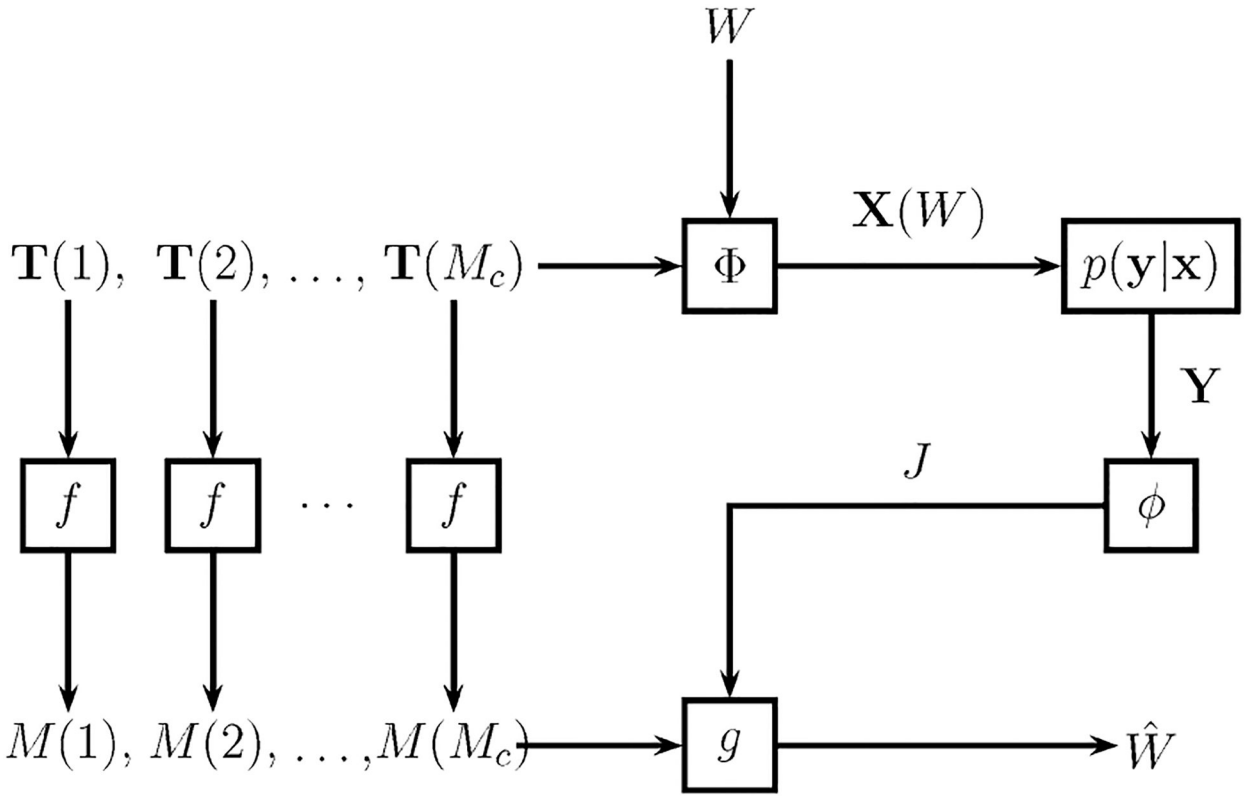

Fig. 1.
Pattern recognition subject to data compression.

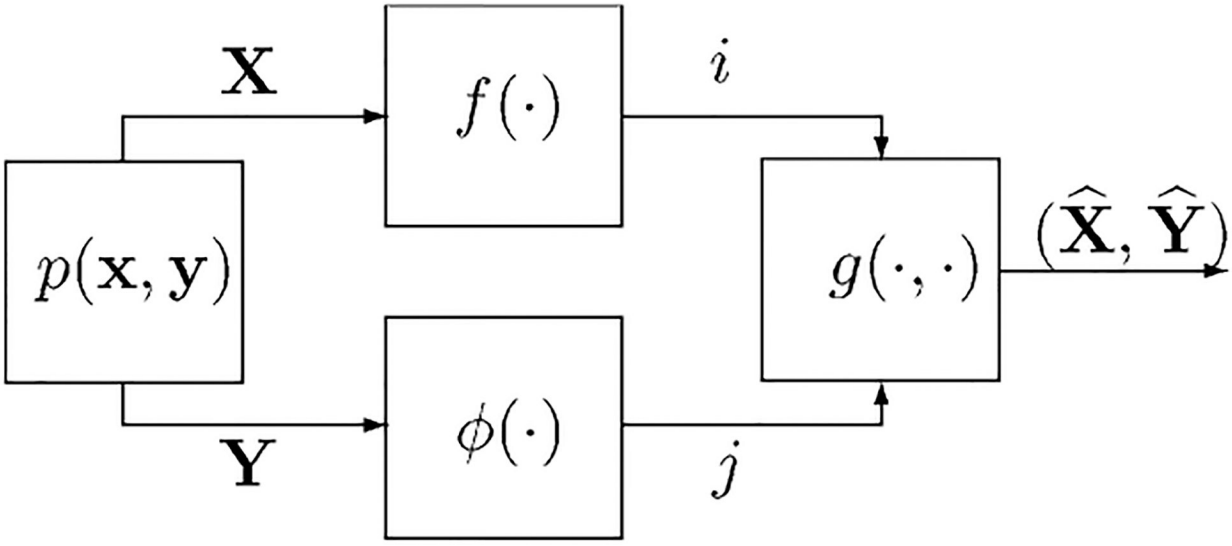

Fig. 2.
The distributed source coding problem.

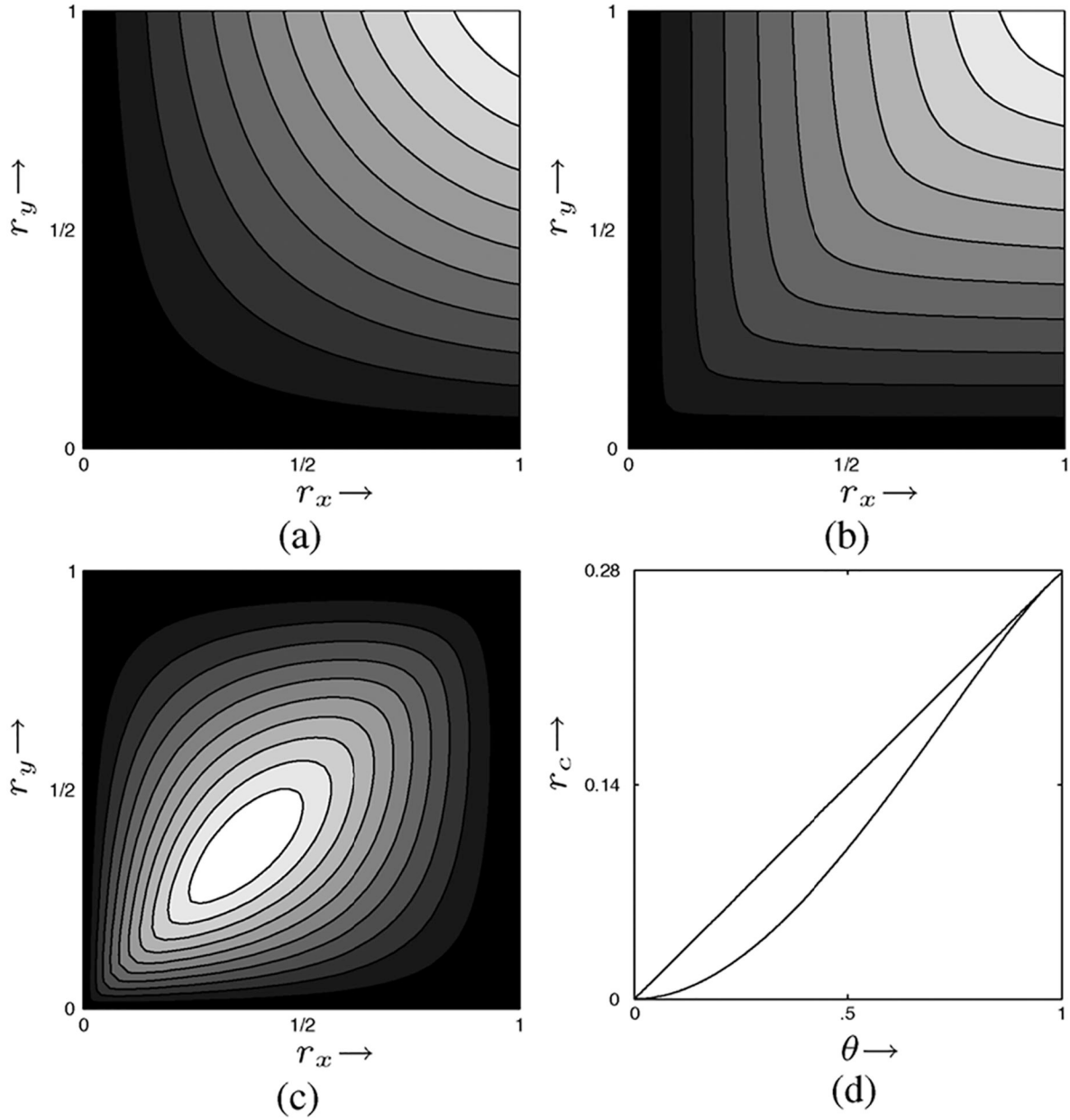

Fig. 3.

Contour plots of the binary inner bound surface (a); outer bound surface (b); differences between the outer bound and inner bounds (c); and plot of inner and outer bound surfaces along a diagonal cut, $(r_x(\theta), r_y(\theta)) = \theta(1, 1)$, $\theta \in [0, 1]$. In these plots $q = 0.2$.

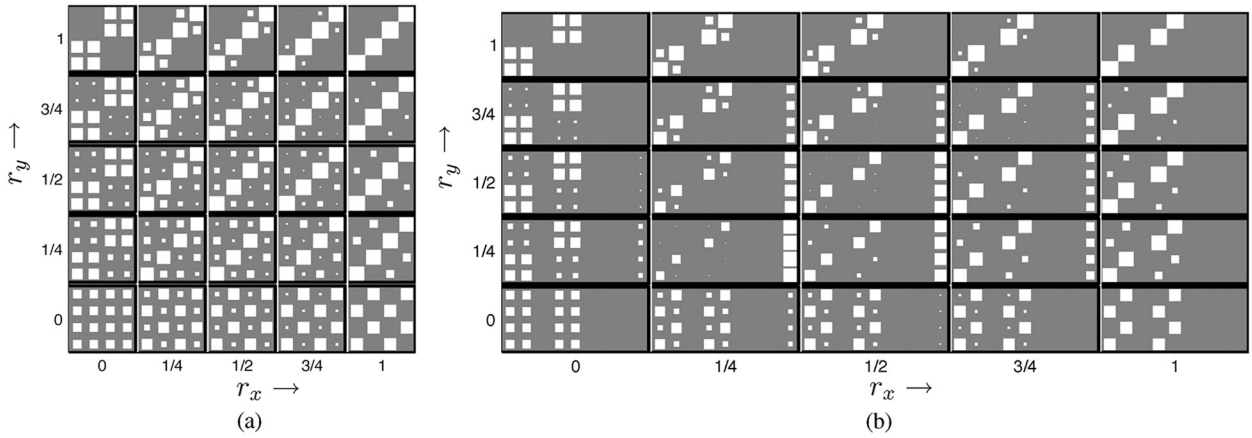

Fig. 4.

Hinton diagrams of the maximizing probability distributions $p(uv \mid xy)$ for 25 values of (r_x, r_y). (a) Distributions for $\mathscr{R}_{\text{in}}$. (b) Distributions for $\mathscr{R}_{\text{out}}$.

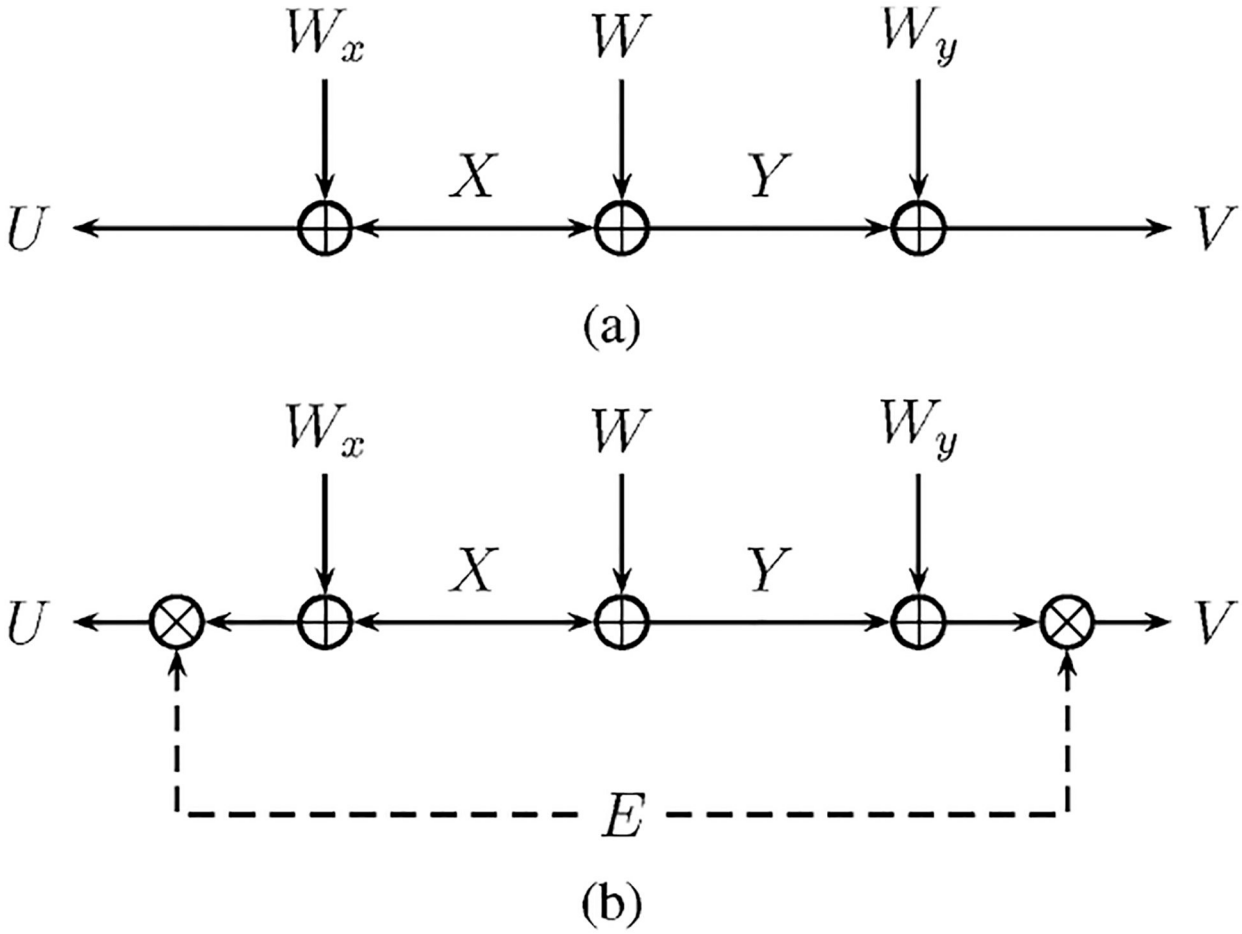

Fig. 5.

Binary-symmetric channel models for the inner and outer bounds. (a) Model for $r_{\text{in}}(r_x, r_y)$ with $|\mathcal{U}| = |\mathcal{V}| = 2$. (b) Model for $r_{out}(r_x, r_y)$ with $|\mathcal{U}| = |\mathcal{V}| = 3$. See text for explanation.

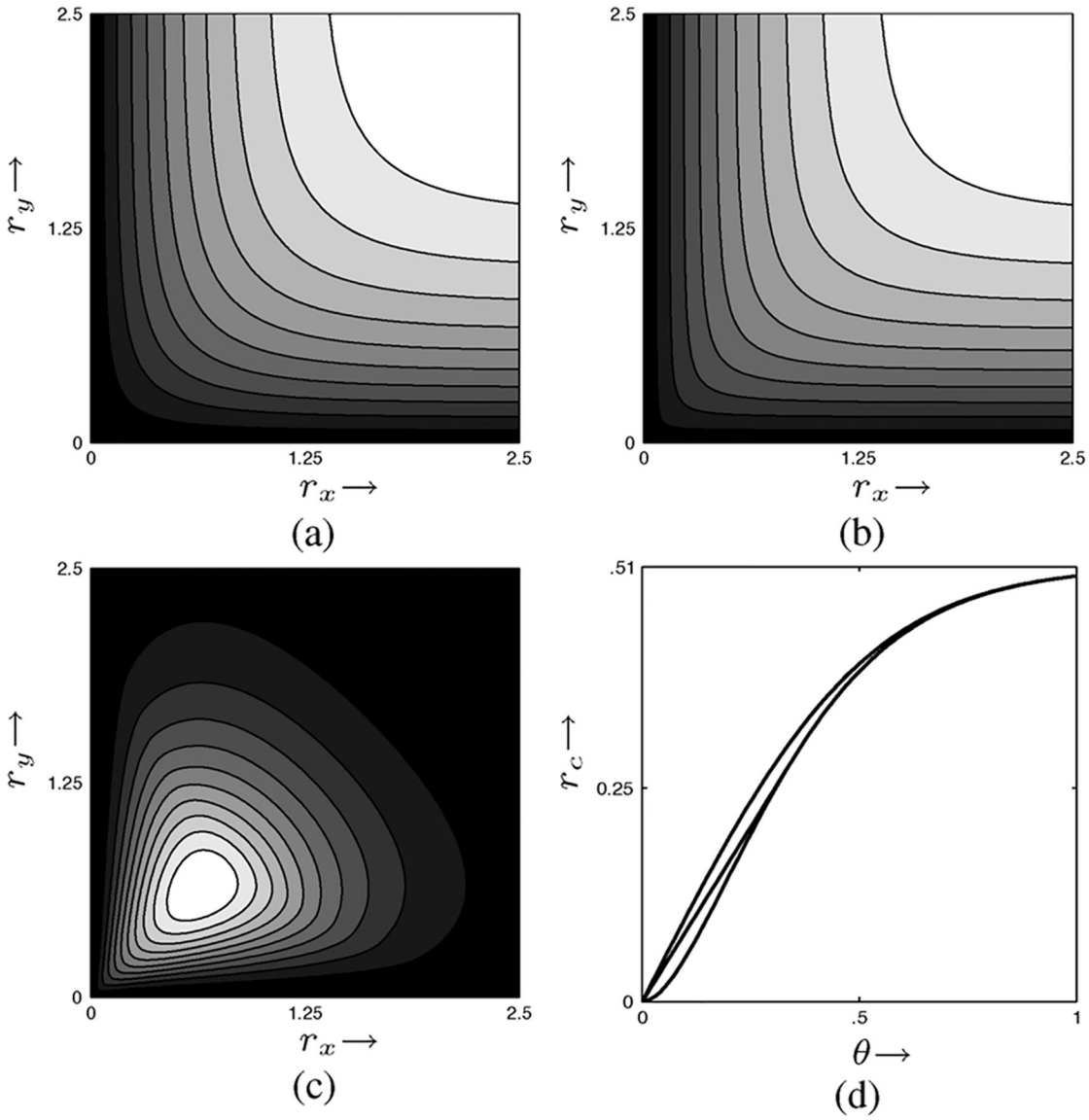

Fig. 6.

Contour plots of (a) the binary inner bound surface; (b) outer bound surface; (c) difference between the outer bound and inner bounds; and (d) plot of surfaces for the inner bound, its convex hull, and the outer bound along a diagonal cut, $(r_x(\theta), r_y(\theta)) = \theta(1, 1)$, $\theta \in [0, 1]$. In these plots $\rho_{xy} = 0.8$.

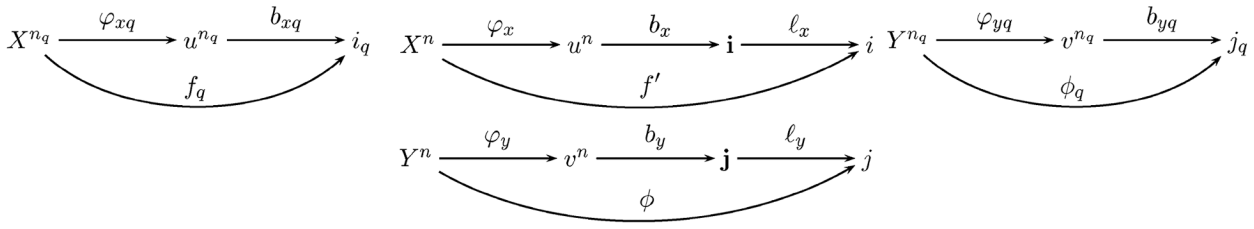

Fig. 7.

Mappings for sequences $\left(X^{n_q}, Y^{n_q}\right)$, $q \in \mathcal{Q}$ and concatenations (X^n, Y^n).