Bounds on inference

Bounds on inference

Flávio P. Calmon, Mayank Varia, Muriel Médard,
Mark M. Christiansen, Ken R. Duffy, Stefano Tessaro

***Abstract*—Lower bounds for the average probability of error of estimating a hidden variable X given an observation of a correlated random variable Y, and Fano's inequality in particular, play a central role in information theory. In this paper, we present a lower bound for the average estimation error based on the marginal distribution of X and the principal inertias of the joint distribution matrix of X and Y. Furthermore, we discuss an information measure based on the sum of the largest principal inertias, called k-correlation, which generalizes maximal correlation. We show that k-correlation satisfies the Data Processing Inequality and is convex in the conditional distribution of Y given X. Finally, we investigate how to answer a fundamental question in inference and privacy: given an observation Y, can we estimate a function $f(X)$ of the hidden random variable X with an average error below a certain threshold? We provide a general method for answering this question using an approach based on rate-distortion theory.**

I. Introduction

Consider the standard problem in estimation theory: Given an observation of a random variable Y, what can we learn about a correlated, hidden variable X? For example, in security systems, X can be the plaintext message, and Y the ciphertext or any additional side information available to an adversary. Throughout the paper, we assume that X and Y are discrete random variables with finite support.

If the joint distribution between X and Y is known, the probability of error of estimating X given an observation of Y can be calculated exactly. However, in most practical settings, this joint distribution is unknown. Nevertheless, it might be possible to estimate certain correlation measures of X and Y reliably, such as maximal correlation, χ^2-statistic or mutual information.

Given estimates of such correlation measures, is it possible to determine a lower bound for the average error probability of estimating X from Y over all possible estimators? We answer this question in the affirmative. In the context of security, this bound might characterize the best estimation of the plaintext that a (computationally unbounded) adversary can make given an observation of the output of the system.

Furthermore, owing to the nature of the joint distribution, it may be infeasible to estimate X from Y with small error probability. However, it is possible that a non-trivial function $f(X)$ exists that is of interest to a learner and can be estimated reliably. If f is the identity function, this reduces to the standard problem of estimating X from Y. Determining if such a function exists is relevant to several applications in inference, privacy and security [1].

In this paper, we establish lower bounds for the average estimation error of X and $f(X)$ given an observation of Y. These bounds depend only on certain measures of information between X and Y and the marginal distribution of X. The results hold for any estimator, and they shed light on the fundamental limits of what can be inferred about a hidden variable from a noisy measurement. The bounds derived here are similar in nature to Fano's inequality [2], and can be characterized as the solution of a convex program which, in turn, is closely related to the rate-distortion optimization problem.

Our work has two main contributions. First, we analyze properties of a measure of information (correlation) between X and Y based on the principal inertias of the joint distribution of X and Y. The estimation of principal inertias is widely studied in the field of correspondence analysis, and is used in practice to analyze categorical data. The metric we propose, called *k-correlation*, is defined as the sum of the k largest principal inertias, which, in turn, are the singular values of a particular decomposition of the joint distribution matrix of X and Y. We show that k-correlation generalizes both the maximal correlation and the χ^2 measures of correlation. We also prove that k-correlation satisfies two key properties for information measures: (i) the Data Processing Inequality and (ii) convexity in the conditional probabilities $p_{Y|X}$. Furthermore, we derive a family of lower bounds for the average error probability of estimating X given Y based on the principal inertias between X and Y and the marginal distribution of X.

The second contribution is a general procedure for bounding the average estimation error of a deterministic function of X given an observation of Y. These bounds are non-trivial and help characterize the fundamental limits of what can be learned about X given an observation of Y. For example, given $I(X;Y) \leq \theta$, a positive integer M and the marginal distribution of X, this procedure allows us to compute a lower bound for the average estimation error of any surjective function that maps the support of X onto $\{1, \ldots, M\}$.

The rest of the paper is organized as follows. Section II presents an overview of the main results and discusses related work. Section III introduces the k-correlation metric of infor-

This work is sponsored by the Intelligence Advanced Research Projects Activity under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. M.C. and K.D. are supported by Science Foundation Ireland Grant No. 11/PI/1177.

F. P. Calmon and M. Médard are with the Research Laboratory of Electronics at the Massachusetts Institute of Technology, Cambridge, MA (e-mail: flavio@mit.edu; medard@mit.edu).

M. Varia is with MIT Lincoln Laboratory, Lexington, MA (e-mail: mayank.varia@ll.mit.edu).

M. M. Christiansen and K. R. Duffy are with the Hamilton Institute at the National University of Ireland, Maynooth, Ireland (e-mail: mark.christiansen@nuim.ie; ken.duffy@nuim.ie).

S. Tessaro is with the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology, Cambridge, MA (e-mail: tessaro@csail.mit.edu).

mation, and proves that it is convex in the transition probability $p_{Y|X}$ and satisfies the Data Processing Inequality. Section IV presents a Fano-like inequality based on the principal inertias and the marginal distribution p_X. Section V presents a general method for deriving bounds for the average estimation error of deterministic surjective functions of X from an observation of Y. Finally, concluding remarks are presented in section VI.

II. Overview of main results and related work

A. Notation

We assume throughout this paper that, for a given sample space Ω, $X : \Omega \to \mathcal{X}$ is the hidden random variable and $Y : \Omega \to \mathcal{Y}$ is the observed random variable, where $\mathcal{X} = \{1, \ldots, m\}$ and $\mathcal{Y} = \{1, \ldots, n\}$ are the respective support sets. We denote by $P_{X,Y}$ and $P_{Y|X}$ the $m \times n$ matrices with entries $[P_{X,Y}]_{i,j} \triangleq p_{X,Y}(i,j)$ and $\left[P_{Y|X}\right]_{i,j} \triangleq p_{Y|X}(j|i)$, respectively. Furthermore, we denote by $\mathbf{p}_X \in \mathbb{R}^m$, $\mathbf{p}_Y \in \mathbb{R}^n$ and $\mathbf{p}_{Y|X=j} \in \mathbb{R}^n$ the column vectors with entries

$$[\mathbf{p}_X]_i \triangleq p_X(i), \quad [\mathbf{p}_Y]_i \triangleq p_Y(i) \text{ and } \left[\mathbf{p}_{Y|X=j}\right]_i \triangleq p_{Y|X}(i|j),$$

respectively. The diagonal matrices with entries p_X and p_Y are represented as $D_X = \operatorname{diag}(\mathbf{p}_X)$ and $D_Y = \operatorname{diag}(\mathbf{p}_Y)$. For a discrete random variable Z, we denote by $X \to Y \to Z$ the fact that $p_{X,Y,Z}(x,y,z) = p_X(x)p_{Y|X}(y|x)p_{Z|Y}(z|y)$ (i.e. X, Y, Z form a Markov chain).

Given an observation of Y, the estimation problem considered here is to find a function $h(Y) = \hat{X}$ that minimizes the average error probability P_e, defined as

$$P_e \triangleq \Pr\left\{\hat{X} \neq X\right\}. \tag{1}$$

Note that $X \to Y \to \hat{X}$. P_e is minimized when $\hat{X}$ is the maximum-likelihood estimate of X.

The column vector with all entries equal to 1 is represented by $\mathbf{1}$. The length of the vector will be clear from the context. For any given matrix A, we denote by $\sigma_k(A)$ the k-th largest singular value of A. If A is hermitian, we denote the k-th largest eigenvalue of A by $\Lambda_k(A)$. We denote by $\mathcal{S}^m_{++}$ the set of positive definite matrices in $\mathbb{R}^{m\times m}$. Furthermore,

$$\mathcal{T}_{m,n} \triangleq \left\{A \in \mathbb{R}^{m\times n} : A \text{ is row-stochastic, } [A]_{i,j} \geq 0\right\}. \tag{2}$$

For a given measure of information (correlation) $\mathcal{I}(X;Y)$ between X and Y (such as mutual information or maximal correlation), we denote $\mathcal{I}(X;Y) = \mathcal{I}(p_X, P_{Y|X})$ when we wish to highlight $\mathcal{I}(X;Y)$ as a function p_X and the transition matrix $P_{Y|X}$.

B. Overview of main results

Assume that the joint distribution $p_{X,Y}$ is unknown, but that the marginal distribution p_X is given. Furthermore, assume that a certain measure of information (correlation) $\mathcal{I}(X;Y)$ between X and Y is bounded above by θ, i.e. $\mathcal{I}(X;Y) \leq \theta$. In practice, the value of θ and p_X could be determined, for example, from multiple i.i.d. samples drawn according to $p_{X,Y}$. The number of samples available might be insufficient to characterize $p_{X,Y}$, but enough to estimate θ and p_X reliably. Under these assumptions, what can be said about the smallest P_e possible? Our goal in this paper is to derive lower bounds of the form $P_e \geq L_{\mathcal{I}}(p_X, \theta)$, creating a limit on how well X can be inferred from Y.

The characterization of $L_{\mathcal{I}}(p_X, \theta)$ for different measures of information $\mathcal{I}$ is particularly relevant for applications in privacy and security, where X is a variable that should remain hidden (e.g. plaintext). A lower bound for P_e can then be viewed as a security metric: regardless of an adversary's computational resources, he will not be able to guess X with an average estimation error smaller than $L_{\mathcal{I}}(p_X, \theta)$ given an observation of Y. Therefore, by simply estimating θ and calculating $L_{\mathcal{I}}(p_X, \theta)$ we are able to evaluate the privacy threat incurred by an adversary that has access to Y.

If $\mathcal{I}(X;Y) = I(X;Y)$, where $I(X;Y)$ is the mutual information between X and Y, then Fano's inequality [2] provides a lower bound for P_e. However, in practice, several other statistics are used in addition to mutual information in order to capture the information (correlation) between X and Y. In this work, we focus on one particular metric, namely the *principal inertia components* of $p_{X,Y}$, denoted by the vector $(\lambda_1, \ldots, \lambda_d)$, where $d = \min\{m-1, n-1\}$, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. The exact definition of the principal inertias is presented in Section III.

1) Bounds based on principal inertia components: The principal inertias generalize other measures that are used in information theory. In particular, $\lambda_1 = \rho_m^2(X;Y)$, where $\rho_m(X;Y)$ is the *maximal correlation* between X and Y. Given

$$\mathcal{S} \triangleq \{(f(X), g(Y)) : \mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0, \\ \mathbb{E}\left[f^2(X)\right] = \mathbb{E}\left[g^2(Y)\right] = 1\},$$

the maximal correlation $\rho_m(X;Y)$ is defined as [3]

$$\rho_m(X;Y) = \max_{(f(X),g(Y))\in\mathcal{S}} \mathbb{E}[f(X)g(Y)].$$

In section III and appendix B, we discuss how to compute the principal inertias and provide two alternative characterizations. Compared to mutual information, the principal inertias provide a finer-grained decomposition of the correlation between X and Y.

We propose a metric of information called k-correlation, defined as $\mathcal{J}_k(X;Y) \triangleq \sum_{i=1}^k \lambda_i$. This metric satisfies two key properties:

- Convexity in $p_{Y|X}$ (Theorem 1);
- Data Processing Inequality (Theorem 2). This is also satisfied by $\lambda_1, \ldots, \lambda_d$ individually.

By making use of the fact that the principal inertia components satisfy the Data Processing Inequality, we are able to derive a family of bounds for P_e in terms of p_X and $\lambda_1, \ldots, \lambda_d$, described in Theorem 3. This result sheds light on the relationship of P_e with the principal inertia components.

One immediate consequence of Theorem 3 is a useful scaling law for P_e in terms of the largest principal inertia (i.e. maximal correlation). Let $X = 1$ be the most likely outcome for X. Corollary 3 proves that the advantage an adversary has of guessing X, over the trivial solution of simply guessing the most likely outcome of X (i.e. $X = 1$), satisfies

$$\mathsf{Adv}(X;Y) \triangleq |1 - p_X(1) - P_e| \leq O\left(\sqrt{\lambda_1}\right). \tag{3}$$

2) Bounding the estimation error of functions: For most security applications, minimizing the probability that an adversary guesses the hidden variable X from an observation of Y is insufficient. Cryptographic definitions of security, and in particular semantic security [1], require that an adversary has negligible advantage in guessing any function of the input given an observation of the output. In light of this, we present bounds for the best possible average error achievable for estimating functions of X given an observation of Y.

Still assuming that $p_{X,Y}$ is unknown, p_X is given and $\mathcal{I}(X;Y) \leq \theta$, we present in Theorem 6 a method for adapting bounds of the form $P_e \geq L_{\mathcal{I}}(p_X, \theta)$ into bounds for the average estimation error of functions of X given Y. This method depends on $\mathcal{I}$ satisfying a few technical assumptions (stated in section V), foremost of which is the existence of a lower bound $L_{\mathcal{I}}(p_X, \theta)$ that is Schur-concave[1] in p_X for a fixed θ. Theorem 6 then states that, under these assumptions, for any deterministic, surjective function $f : \mathcal{X} \to \{1, \ldots, M\}$, we can obtain a lower bound for the average estimation error of f by computing $L_{\mathcal{I}}(p_U, \theta)$, where U is a random variable that is a function X.

Note that Schur-concavity of $L_{\mathcal{I}}(p_X, \theta)$ is crucial for this result. In Theorem 4, we show that this condition is always satisfied when $\mathcal{I}(X;Y)$ is concave in p_X for a fixed $p_{Y|X}$, convex in $p_{Y|X}$ for a fixed p_X, and satisfies the Data Processing Inequality. This generalizes a result by Ahlswede [4] on the extremal properties of rate-distortion functions. Consequently, Fano's inequality can be adapted in order to bound the average estimation error of functions, as shown in Corollary 4. By observing that a particular form of the bound stated in Theorem 3 is Schur-concave, we also present a bound for the error probability of estimating functions in terms of the maximal correlation, as shown in Corollary 5.

C. Background

The joint distribution matrix $P_{X,Y}$ can be viewed as a contingency table and decomposed using standard techniques from correspondence analysis [5], [6]. We note that this decomposition was originally investigated by Hirschfield [7], Gebelein [8] and later by Rényi [3]. For a quick overview of correspondence analysis, we refer the reader to [9].

The largest principal inertia of $P_{X,Y}$ is equal to $\rho_m^2(X;Y)$, where $\rho_m(X;Y)$ is the *maximal correlation* between X and Y. Maximal correlation has been widely studied in the information theory and statistics literature (e.g [3]). Anantharam *et al.* present in [10] an overview of different characterizations of maximal correlation, as well as its application in information theory.

The Data Processing Inequality for the principal inertias was shown by Kang and Ulukus in [11, Theorem 2] in a different setting than the one considered here. Kang and Ulukus make use of the decomposition of the joint distribution matrix to derive outer bounds for the rate-distortion region achievable in certain distributed source and channel coding problems.

[1]A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *Schur-concave* if for all $x, y \in \mathbb{R}^n$ where x is majorized by y, then $f(x) \geq f(y)$.

Lower bounds on the average estimation error can be found using Fano-type inequalities. Recently, Guntuboyina *et al.* ([12], [13]) presented a family of sharp bounds for the minmax risk in estimation problems involving general f-divergences. These bounds generalize Fano's inequality and, under certain assumptions, can be extended in order to lower bound P_e.

Most information-theoretic approaches for estimating or communicating functions of a random variable are concerned with properties of specific functions given i.i.d. samples of the hidden variable X, such as in the functional compression literature [14], [15]. These results are rate-based and asymptotic, and do not immediately extend to the case where the function $f(X)$ can be an arbitrary member of a class of functions, and only a single observation is available.

More recently, Kumar and Courtade [16] investigated boolean functions in an information-theoretic context. In particular, they analyzed which is the most informative (in terms of mutual information) 1-bit function (i.e. $M = 2$) for the case where X is composed by n i.i.d. Bernoulli(1/2) random variables, and Y is the result of passing X through a discrete memoryless binary symmetric channel. Even in this simple case, determining the most informative function is non-trivial.

Bellare *et al.* [17] considered the standard wiretap setting [18], and proved the equivalence between semantic security and minimizing the maximum mutual information over all possible input message distributions. Since semantic security [1] is achieved only when an adversary's advantage of correctly computing a function of the hidden variable given an observation of the output is negligibly small, the results in [17] are closely related to the ones presented here.

III. A measure of information based on principal inertias

In this section we discuss how the joint probability matrix $P_{X,Y}$ can be decomposed into principal inertia components[2], and introduce the k-correlation measure. We also prove that the k-correlation measure is convex in $p_{Y|X}$ and satisfies the Data Processing Inequality. Several equivalent characterizations of the principal inertias have appeared in the literature (e.g. [8] and [9]). We discuss two of these characterizations in appendix B.

Consider the singular value decomposition [19] of the matrix $D_X^{-1/2} P_{X,Y} D_Y^{-1/2}$, given by

$$D_X^{-1/2} P_{X,Y} D_Y^{-1/2} = U \Sigma V^T, \tag{4}$$

and define $\widetilde{A} \triangleq D_X^{1/2} U$ and $\widetilde{B} \triangleq D_Y^{1/2} V$. Then

$$P_{X,Y} = \widetilde{A} \Sigma \widetilde{B}^T, \tag{5}$$

where $\widetilde{A}^T D_X^{-1} \widetilde{A} = \widetilde{B}^T D_Y^{-1} \widetilde{B} = I$.

Definition 1. The square of the diagonal entries of $\widetilde{\Sigma}$ are called the *principal inertias*, and are denoted by $\lambda_1, \ldots, \lambda_d$, where $d = \min(m-1, n-1)$. Throughout this paper, we assume that principal inertias are ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.

[2]The term *principal inertias* is borrowed from the correspondence analysis literature [6].

It can be shown that $\widetilde{A}$, $\widetilde{B}$ and Σ have the form

$$\widetilde{A} = [\mathbf{p}_X \quad A], \ \widetilde{B} = [\mathbf{p}_Y \quad B], \tag{6}$$
$$\Sigma = \operatorname{diag}\left(1, \sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d}\right),$$

and, consequently, the joint distribution can be written as

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) + \sum_{k=1}^{d} \sqrt{\lambda_k} b_{y,k} a_{x,k}, \tag{7}$$

where $a_{x,k}$ and $b_{y,k}$ are the entries of A and B in (6), respectively,

Based on the decomposition of the joint distribution matrix, we define below a measure of information between X and Y based on the principal inertias.

Definition 2. Let $\|A\|_k$ denote the k-th Ky Fan norm[3] [19, Example 7.4.8] of a matrix A. For $1 \leq k \leq d$, we define the *k-correlation* between X and Y as

$$\mathcal{J}_k(X;Y) \triangleq \|D_X^{-1/2} P_{X,Y} D_Y^{-1} P_{X,Y}^T D_X^{-1/2}\|_k - 1 \tag{8}$$
$$= \sum_{i=1}^{k} \lambda_i. \tag{9}$$

Note that

$$\mathcal{J}_1(X;Y) = \rho_m^2(X;Y),$$

where $\rho_m(X;Y)$ is the *maximal correlation* of (X,Y) [10], and

$$\mathcal{J}_d(X;Y) = \mathbb{E}_{X,Y}\left[\frac{p_{X,Y}(X,Y)}{p_X(X)p_Y(Y)}\right] - 1 = \chi^2.$$

We now show that k-correlation and, consequently, maximal correlation, is convex in $p_{Y|X}$ for a fixed p_X and satisfies the Data Processing Inequality.

1) Convexity in $p_{Y|X}$: We use the next lemma to prove convexity of $\mathcal{J}_k(X;Y)$ in the transition probability $p_{X,Y}$.

Lemma 1. *For $W \in \mathcal{S}_{++}^m$ and $1 \leq k \leq m$, the function $h_k : \mathbb{R}^{m\times n} \times \mathcal{S}_{++}^n \to \mathbb{R}$ defined as*

$$h_k(C,W) \triangleq \|CW^{-1}C^T\|_k \tag{10}$$

is convex.

Proof: Let $Q \triangleq CW^{-1}C^T$. Since Q is positive semidefinite, $\|Q\|_k$ is the sum of the k largest eigenvalues of Q, and can be written as [20], [21]:

$$h_k(C,W) = \|Q\|_k = \max_{Z^TZ = I_k} \operatorname{tr}\left(Z^TQZ\right). \tag{11}$$

Let Z be fixed and $Z^TZ = I_k$, and denote the i-th column of Z by $\mathbf{z}_i$. Note that $g(\mathbf{a},W) \triangleq \mathbf{a}^TW^{-1}\mathbf{a}$ is convex [22, Example 3.4] and, consequently, $g(C^T\mathbf{z}_i, W)$ is also convex in C and W. Since the sum of convex functions is itself convex, then $\operatorname{tr}\left(Z^TQZ\right) = \sum_{i=1}^m g(C^T\mathbf{z}_i, W)$ is also convex in X and Y. The result follows by noting that the pointwise supremum over an infinite set of convex functions is also a convex function [22, Sec. 3.2.3]. ∎

[3]For $A \in \mathbb{R}^{m\times n}$, $\|A\|_k = \sum_{i=1}^k \sigma_i$, where $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$ are the singular values of A.

Theorem 1. *For a fixed p_X, $\mathcal{J}_k(X;Y)$ is convex in $p_{Y|X}$.*

Proof: Note that $\mathcal{J}_k(X;Y) = h_k(D_XP_{Y|X}, D_Y) - 1$, where h_k is defined in equation (10). For a fixed p_X, D_Y is a linear combination of $p_{Y|X}$. Therefore, since h_k is convex (Lemma 1), and composition with an affine mapping preserves convexity, the result follows. ∎

2) A data processing result: In the next theorem, we prove that the principal inertias satisfy the Data Processing Inequality.

Theorem 2. *Assume that $X' \to X \to Y$, where X' is a discrete random variable with finite support. Let $\lambda_1, \lambda_2, \ldots, \lambda_d$ and $\lambda_1', \lambda_2', \ldots, \lambda_d'$ denote the principal inertias of $P_{X,Y}$ and $P_{X',Y}$, respectively. Then $\lambda_1 \geq \lambda_1', \lambda_2 \geq \lambda_2', \ldots, \lambda_d \geq \lambda_d'$.*

Remark 1. This data processing result was also proved by Kang and Ulukus in [11, Theorem 2], even though they do not make the explicit connection with maximal correlation and principal inertias. A weaker form of Theorem 2 can be derived using a clustering result presented in [6, Sec. 7.5.4] and originally due to Deniau *et al.* [23]. We use a different proof technique from the one in [6, Sec. 7.5.4] and [11, Theorem 2] to show result stated in the theorem, and present the proof here for completeness. Finally, a related data processing result was stated in [24, Eq. (31)].

Proof: Assume without loss of generality that $\mathcal{X}' = \{1, \ldots, m'\}$ is the support set of X'. Then $P_{X',Y} = FP_{X,Y}$, where F is a $m' \times m$ column stochastic matrix. Note that F represents the conditional distribution of the mapping $X' \to X$, where the (i,j)-th entry of F is $p_{X'|X}(i|j)$.

Consider the decomposition of $P_{X',Y} = FP_{X,Y}$:

$$\begin{aligned} S' &= D_{X'}^{-1/2}\left(P_{X'Y} - \mathbf{p}_{X'}\mathbf{p}_Y^T\right) D_Y^{-1/2} \\ &= D_{X'}^{-1/2} F \left(P_{XY} - \mathbf{p}_X\mathbf{p}_Y^T\right) D_Y^{-1/2} \\ &= D_{X'}^{-1/2} F D_X^{1/2} S, \end{aligned}$$

where S is given by

$$S \triangleq D_X^{-1/2}\left(P_{X,Y} - \mathbf{p}_X\mathbf{p}_Y^T\right) D_Y^{-1/2}. \tag{12}$$

Note that the singular values of S' are the principal inertias $\lambda_1', \ldots, \lambda_d'$.

Let $E = D_{X'}^{-1/2} F D_X^{1/2}$, where that the size of E is $m' \times m$. Since $[FD_X]_{i,j} = p_{X',X}(i,j)$, then the (i,j)-th entry of E is

$$[E]_{i,j} = \frac{p_{X',X}(i,j)}{\sqrt{p_{X'}(i)p_X(j)}}.$$

Observe that E has the same form as (4), and, therefore, $\|E\|_1 = 1$. Let $H = S^TS - S'^TS'$. Then for $\mathbf{y} \in \mathbb{R}^n$ and $S\mathbf{y} = \mathbf{z}$:

$$\begin{aligned} \mathbf{y}^TH\mathbf{y} &= \mathbf{y}^TS^TS\mathbf{y} - \mathbf{y}^TS'^TS'\mathbf{y} \\ &= \|\mathbf{z}\|_2 - \|E\mathbf{z}\|_2 \\ &\geq \|\mathbf{z}\|_2 - \|E\|_1\|\mathbf{z}\|_2 \\ &= 0. \end{aligned}$$

Consequently, H is positive semidefinite. Since H is symmetric, it follows from Weyl's theorem [19, Theorem 4.3.1] that

for $k = 1, \ldots, n$,

$$\begin{aligned}\Lambda_k(S'^T S') &\le \Lambda_k(S'^T S' + H) \\ &= \Lambda_k(S^T S) \\ &= \lambda_k.\end{aligned}$$

Since $\Lambda_k(S'^T S') = \lambda'_k$, the result follows. ∎

The next corollary is a direct consequence of the previous theorem.

Corollary 1. *For $X' \to X \to Y$ forming a Markov chain, $J_k(X';Y) \le J_k(X;Y)$.*

IV. A lower bound for the estimation error probability in terms of the principal inertias

Throughout the rest of the paper, we assume without loss of generality that p_X is sorted in decreasing order, i.e. $p_X(1) \ge p_X(2) \ge \cdots \ge p_X(m)$.

Definition 3. Let $\mathbf{\Lambda}(P_{X,Y})$ denote the vector of principal inertias of a joint distribution matrix $P_{X,Y}$ sorted in decreasing order, i.e. $\mathbf{\Lambda}(P_{X,Y}) = (\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_d)$. We denote $\mathbf{\Lambda}(P_{X,Y}) \le \boldsymbol{\lambda}$ if $\widetilde{\lambda}_1 \le \lambda_1, \ldots, \widetilde{\lambda}_d \le \lambda_d$ and

$$\mathcal{R}(\mathbf{q}_X, \boldsymbol{\lambda}) \triangleq \left\{P_{X,Y} \middle| \mathbf{p}_X = \mathbf{q}_X \text{ and } \mathbf{\Lambda}(P_{X,Y}) \le \boldsymbol{\lambda}\right\}. \quad (13)$$

In the next theorem we present a Fano-like bound for the estimation error probability of X that depends on the marginal distribution p_X and on the principal inertias.

Theorem 3. *For $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$, define*

$$k^* \triangleq \max\left\{k \in \{1, \ldots, m\} \;\middle|\; p_X(k) - \mathbf{p}_X^T \mathbf{p}_X \ge 0\right\}, \quad (14)$$

$$\begin{aligned}f_0^*(\mathbf{p}_X, \boldsymbol{\lambda}) \triangleq &\sum_{i=1}^{k^*} \lambda_i p_X(i) + \sum_{i=k^*+1}^{m} \lambda_{i-1} p_X(i) \\ &- \lambda_{k^*} \mathbf{p}_X^T \mathbf{p}_X,\end{aligned}$$

$$\begin{aligned}g_0(\beta, \mathbf{p}_X, \boldsymbol{\lambda}) &\triangleq f_0^*(\mathbf{p}_X, \boldsymbol{\lambda}) + \sum_{i=1}^{m} \left([p_X(i) - \beta]^+\right)^2, \\ U_0(\beta, \mathbf{p}_X, \boldsymbol{\lambda}) &\triangleq \beta + \sqrt{g_0(\beta, \mathbf{p}_X, \boldsymbol{\lambda})}, \\ U_1(\mathbf{p}_X, \boldsymbol{\lambda}) &\triangleq \min_{0 \le \beta \le p_X(2)} U_0(\beta, \mathbf{p}_X, \boldsymbol{\lambda}).\end{aligned}$$

Then for any $P_{X,Y} \in \mathcal{R}(\mathbf{p}_X, \boldsymbol{\lambda})$,

$$P_e \ge 1 - U_1(\mathbf{p}_X, \boldsymbol{\lambda}). \quad (15)$$

Proof: The proof of the theorem is presented in the appendix. ∎

Remark 2. If $\lambda_i = 1$ for all $1 \le i \le d$, (15) reduces to $P_e \ge 0$. Furthermore, if $\lambda_i = 0$ for all $1 \le i \le d$, (15) simplifies to $P_e \ge 1 - p_X(1)$.

We now present a few direct but powerful corollaries of the result in Theorem 3. We note that a bound similar to (16) below has appeared in the context of bounding the minmax decision risk in [25, (3.4)]. However, the proof technique used in [25] does not lead to the general bound presented in Theorem 3.

Corollary 2. *If X is uniformly distributed in $\{1, \ldots, m\}$, then*

$$P_e \ge 1 - \frac{1}{m} - \frac{\sqrt{(m-1)\chi^2}}{m}. \quad (16)$$

Furthermore, if only the maximal correlation $\rho_m(X;Y) = \sqrt{\lambda_1}$ is given, then

$$\begin{aligned}P_e &\ge 1 - \frac{1}{m} - \sqrt{\lambda_1}\left(1 - \frac{1}{m}\right) \\ &= 1 - \frac{1}{m} - \rho_m(X;Y)\left(1 - \frac{1}{m}\right).\end{aligned}$$

Corollary 3. *For any pair of variables (X, Y) with marginal distribution in X equal to p_X and maximal correlation (largest principal inertia) $\rho_m^2(X;Y) = \lambda_1$, we have for all $\beta \ge 0$*

$$P_e \ge 1 - \beta - \sqrt{\lambda_1\left(1 - \sum_{i=1}^{m} p_X^2(i)\right) + \sum_{i=1}^{m}\left([p_X(i) - \beta]^+\right)^2}. \quad (17)$$

In particular, setting $\beta = p_X(2)$,

$$\begin{aligned}P_e &\ge 1 - p_X(2) - \sqrt{\lambda_1\left(1 - \sum_{i=1}^{m} p_X^2(i)\right) + \left(p_X(1) - p_X(2)\right)^2} \\ &\ge 1 - p_X(1) - \rho_m(X;Y)\sqrt{\left(1 - \sum_{i=1}^{m} p_X^2(i)\right)}. \quad (18)\end{aligned}$$

Remark 3. The bounds (17) and (18) are particularly insightful in showing how the error probability scales with the input distribution and the maximal correlation. For a given $p_{X,Y}$, recall that $\mathsf{Adv}(X;Y)$, defined in (3), is the advantage of correctly estimating X from an observation of Y over a random guess of X when Y is unknown. Then, from equation (18)

$$\begin{aligned}\mathsf{Adv}(X;Y) &\le \rho_m(X;Y)\sqrt{\left(1 - \sum_{i=1}^{m} p_X^2(i)\right)} \\ &= O(\rho_m(X;Y)).\end{aligned}$$

Therefore, the advantage of estimating X from Y decreases at least linearly with the maximal correlation between X and Y.

V. Lower bounds on estimating functions

For any function $f : \mathcal{X} \to \mathcal{U}$, we denote by $\hat{f}$ the maximum-likelihood estimator of $f(X)$ given an observation of Y. For a given integer $1 \le M \le |\mathcal{X}|$, we define

$$\mathcal{F}_M \triangleq \left\{f : \mathcal{X} \to \mathcal{U} \;\middle|\; f \text{ is surjective and } |\mathcal{U}| = M\right\}$$

and

$$P_{e,M} \triangleq \min_{f \in \mathcal{F}_M} \Pr\{f(X) \ne \hat{f}\}. \quad (19)$$

$P_{e,|\mathcal{X}|}$ is simply the error probability of estimating X from Y, i.e. $P_{e,|\mathcal{X}|} = P_e$.

Throughout this section we adopt the following additional assumption.

Assumption 1. *An upper bound θ for a given measure of information $\mathcal{I}(X;Y)$ between X and Y is given, i.e. $\mathcal{I}(X;Y) \leq \theta$. Furthermore, $\mathcal{I}(X;Y)$ satisfies the Data Processing Inequality, is convex in $p_{Y|X}$ for a given p_X, and is invariant to row and column permutations of the joint distribution matrix $p_{X,Y}$. Finally, we also assume that the marginal distribution of X, given by p_X, is known.*

Under this assumption, what can be said about $P_{e,M}$? In the next sections we present a general procedure to derive non-trivial lower bounds for $P_{e,M}$.

A. Extremal properties of the error-rate function

Before investigating how to bound $P_{e,M}$, we first study how to bound P_e in a more general setting than the one in section IV. Note that

$$P_e \geq \min_{P_{Y|X},E} \; 1 - \text{tr}\left(D_X P_{Y|X} E\right) \\ \text{s.t. } \mathcal{I}(p_X, P_{Y|X}) \leq \theta,\; P_{Y|X} \in \mathcal{T}_{m,n},\; E \in \mathcal{T}_{n,m}.$$

Here E denotes the mapping from Y to $\hat{X}$. By fixing $P_{Y|X}$ and taking the dual in E of the previous convex program, we can verify that E will always be a row-stochastic matrix with entries equal to 0 or 1. Since $\mathcal{I}(X;Y)$ satisfies the Data Processing Inequality, $P_e \geq e_{\mathcal{I}}(p_X, \theta)$, where $e_{\mathcal{I}}(p_X, \theta)$ is defined below.

Definition 4. The *error-rate function* $e_{\mathcal{I}}(p_X, \theta)$ is the solution of the following convex program:

$$e_{\mathcal{I}}(p_X, \theta) \triangleq \min_{P_{\hat{X}|X}} \; 1 - \text{tr}\left(D_X P_{\hat{X}|X}\right) \\ \text{s.t. } \mathcal{I}(p_X, P_{\hat{X}|X}) \leq \theta,\; P_{\hat{X}|X} \in \mathcal{T}_{m,m}\;. \tag{20}$$

Due to convexity of $\mathcal{I}(p_X, P_{\hat{X}|X})$ in $P_{\hat{X}|X}$, it follows directly that $e_{\mathcal{I}}(p_X, \theta)$ is convex in θ for a fixed p_X. Furthermore, the cost function (20) is equal to the average Hamming distortion $\mathbb{E}_{X,\hat{X}}\left[d_H(X, \hat{X})\right]$ between X and $\hat{X}$. Therefore, $e_{\mathcal{I}}(p_X, \theta)$ has a dual relationship[4] with the rate-distortion problem

$$R_{\mathcal{I}}(p_X, \Delta) \triangleq \min_{P_{\hat{X}|X}} \; \mathcal{I}(p_X, P_{\hat{X}|X}) \\ \text{s.t. } \mathbb{E}_{X,\hat{X}}\left[d_H(X, \hat{X})\right] \leq \Delta,\; P_{\hat{X}|X} \in \mathcal{T}_{m,m}.$$

We will now prove that, for a fixed θ (respectively, fixed Δ), $e_{\mathcal{I}}(p_X, \theta)$ (resp. $R_{\mathcal{I}}(p_X, \Delta)$) is *Schur-concave* in p_X if $\mathcal{I}(p_X, P_{Y|X})$ is concave in p_X for a fixed $P_{Y|X}$. Ahlswede [4, Theorem 2] proved this result for the particular case where $\mathcal{I}(X;Y) = I(X;Y)$ by investigating the properties of the explicit characterization of the rate-distortion function under Hamming distortion. The proof presented here is considerably simpler and more general, and is based on a proof technique used by Ahlswede in [4, Theorem 1].

[4]The authors thank Prof. Yury Polyansky (MIT) for pointing out the dual relationship.

Theorem 4. *If $\mathcal{I}(p_X, P_{Y|X})$ is concave in p_X for a fixed $P_{Y|X}$, then $e_{\mathcal{I}}(p_X, \theta)$ and $R_{\mathcal{I}}(p_X, \Delta)$ are Schur-concave in p_X for a fixed θ and Δ, respectively.*

Proof: Consider two probability distributions p_X and q_X defined over $\mathcal{X} = \{1, \ldots, m\}$. As usual, let $p_X(1) \geq p_X(2) \geq \cdots \geq p_X(m)$ and $q_X(1) \geq q_X(2) \geq \cdots \geq q_X(m)$. Furthermore, assume that p_X majorizes q_X, i.e. $\sum_{i=1}^{k} q_X(i) \leq \sum_{i=1}^{k} p_X(i)$ for $1 \leq k \leq m$. Therefore q_X is a convex combination of permutations of p_X [26], and can be written as $q_Z = \sum_{i=1}^{l} a_i \pi_i p_X$ for some $l \geq 1$, where $a_i \leq 0$, $\sum a_i = 1$ and π_i are permutation operators, i.e. $\pi_i p_X = p_{\pi_i X}$. Hence, for a fixed $A \in \mathcal{T}_{m,n}$:

$$\begin{aligned} \mathcal{I}(q_X, A) &= \mathcal{I}\left(\sum_{i=1}^{l} a_i \pi_i p_X, A\right) \\ &\leq \sum_{i=1}^{l} a_i \mathcal{I}(\pi_i p_X, A), \\ &= \sum_{i=1}^{l} a_i \mathcal{I}(p_X, \pi_i A \pi_i), \end{aligned}$$

where the inequality follows from the concavity assumption and from $\mathcal{I}(X;Y)$ being invariant to row and column permutations of the joint distribution matrix $p_{X,Y}$. Consequently, from equation (20),

$$\begin{aligned} e_{\mathcal{I}}(q_X, \theta) &= \inf_{A \in \mathcal{T}_{m,m}} \left\{ 1 - \sum_{i=1}^{l} a_i \text{tr}\left(D_X \pi_i A \pi_i\right) : \right. \\ &\qquad \left. \sum_{i=1}^{l} a_i \mathcal{I}(p_X, \pi_i A \pi_i) \leq \theta \right\} \\ &\geq \inf_{A_1, \ldots, A_l \in \mathcal{T}_{m,m}} \left\{ \sum_{i=1}^{l} a_i (1 - \text{tr}\left(D_X A_i\right)) : \right. \\ &\qquad \left. \sum_{i=1}^{l} a_i \mathcal{I}(p_X, A_i) \leq \theta \right\} \\ &= \inf_{\theta_1, \ldots, \theta_l \geq 0} \left\{ \sum_{i=1}^{l} a_i e_{\mathcal{I}}(p_X, \theta_i) : \sum_{i=1}^{l} a_i \theta_i = \theta \right\} \\ &\geq \inf_{\theta_1, \ldots, \theta_l \geq 0} \left\{ e_{\mathcal{I}}\left(p_X, \sum a_i \theta_i\right) : \sum_{i=1}^{l} a_i \theta_i = \theta \right\} \\ &= e_{\mathcal{I}}\left(p_X, \theta\right), \end{aligned}$$

where the last inequality follows from the convexity of the error-rate function. Since this holds for any q_X that is majorized by p_X, $e_{\mathcal{I}}(p_X, \theta)$ is Schur-concave. Schur-concavity of $R_{\mathcal{I}}(p_X, \Delta)$ follows directly from its dual relationship with $e_{\mathcal{I}}(p_X, \theta)$. ∎

For $\mathcal{I} = \mathcal{J}_k$, the convex program (20) might be difficult to compute due to the constraint on the sum of the singular values. The next theorem presents a convex program that evaluates a lower bound for $e_{\mathcal{J}_k}(p_X, \theta)$ and can be solved using standard methods.

Theorem 5.

$$e_{\mathcal{J}_k}(p_X,\theta) \geq \min_{P_{\hat{X}|X}} \; 1-\operatorname{tr}\left(D_X P_{\hat{X}|X}\right)$$
$$\text{s.t.} \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{p_X(i)p^2_{\hat{X}|X}(j|i)}{y_j} \leq \theta+1, \quad (21)$$
$$P_{\hat{X}|X} \in \mathcal{T}_{m,m},$$
$$\sum_{j=1}^{m} p_X(i)p_{\hat{X}|X}(j|i) = y_j,\ 1\leq j\leq m.$$

Proof: Let $F \triangleq D_X^{-1/2}P_{XY}D_Y^{-1/2}$. Then

$$\mathcal{J}_k(X;Y) = \|FF^T\|_k - 1.$$

Let

$$c_i \triangleq \sum_{j=1}^{m}\frac{p_X(i)p^2_{\hat{X}|X}(j|i)}{y_j}$$

be the i-th diagonal entry of FF^T. By using the fact that the eigenvalues majorize the diagonal entries of a Hermitian matrix ([19, Theorem 4.3.45]), we find

$$\sum_{i=1}^{k} c_i \leq \|FF^T\|_k,$$

and the result follows. Note that convexity of the constraint (21) follows from the fact that the perspective of a convex function is convex [22, Sec. 2.3.3]. ∎

B. Bounds for $P_{e,M}$

Still adopting assumption 1, a lower bound for $P_{e,M}$ can be derived as long as $e_{\mathcal{I}}(p_X,\theta)$ or a lower bound for $e_{\mathcal{I}}(p_X,\theta)$ is Schur-concave in p_X.

Theorem 6. *For a given M, $1\leq M\leq m$, and p_X, let $U=g(X)$, where $g_M:\{1,\dots,m\}\to\{1,\dots,M\}$ is defined as*

$$g_M(x) \triangleq \begin{cases} 1 & 1\leq x\leq m-M+1 \\ x-m+M & m-M+2\leq x\leq m\ . \end{cases}$$

Let p_U be the marginal distribution[5] of U. Assume that, for a given measure of information $\mathcal{I}(X;Y)$, there exists a function $L_{\mathcal{I}}(\cdot,\cdot)$ such that for all distributions q_X and any θ, $e_{\mathcal{I}}(q_X,\theta)\geq L_{\mathcal{I}}(q_X,\theta)$. Under assumption 1, if $L_{\mathcal{I}}(p_X,\theta)$ is Schur-concave in p_X, then

$$P_{e,M} \geq L_{\mathcal{I}}(p_U,\theta)\ . \quad (22)$$

Proof: The result follows from the following chain of inequalities:

$$\begin{aligned} P_{e,M} &\overset{(a)}{\geq} \min_{f\in\mathcal{F}_M,\widetilde{\theta}}\left\{e_{\mathcal{I}}\left(p_{f(X)},\widetilde{\theta}\right) : \widetilde{\theta}\leq\theta\right\} \\ &\geq \min_{f\in\mathcal{F}_M}\left\{e_{\mathcal{I}}\left(p_{f(X)},\theta\right)\right\} \\ &\overset{(b)}{\geq} \min_{f\in\mathcal{F}_M}\left\{L_{\mathcal{I}}\left(p_{f(X)},\theta\right)\right\} \\ &\overset{(c)}{\geq} L_{\mathcal{I}}(p_U,\theta), \end{aligned}$$

where (a) follows from the Data Processing Inequality, (b) follows from $e_{\mathcal{I}}(q_X,\theta)\geq L_{\mathcal{I}}(q_X,\theta)$ for all q_X, and θ and (c) follows from the Schur-concavity of the lower bound and by observing that p_U majorizes $p_{f(X)}$ for every $f\in\mathcal{F}_M$. ∎

The following two corollaries illustrate how Theorem 6 can be used for different measures of information, namely mutual information and maximal correlation.

Corollary 4. *Let $I(X;Y)\leq\theta$. Then*

$$P_{e,M} \geq d^*$$

where d^* is the solution of

$$h_b(d^*) + d^*\log(m-1) = \min\{H(U)-\theta, 0\},$$

and $h_b(\cdot)$ is the binary entropy function.

Proof: $R_I(p_X,\delta)$ is the well known rate-distortion function under Hamming distortion, which satisfies ([27, (9.5.8)]) $R_I(p_X,\delta)\geq H(X)-h_b(d^*)-d^*\log(m-1)$. The result follows from Theorem 4, since mutual information is concave in p_X. ∎

Corollary 5. *Let $\mathcal{J}_1(X;Y)=\rho_m(X;Y)\leq\theta$. Then*

$$P_{e,M} \geq 1-p_U(1)-\theta\sqrt{\left(1-\sum_{i=1}^{M}p_U^2(i)\right)}\ .$$

Proof: The proof follows directly from Theorems 1, 2 and Corollary 3, by noting that (18) is Schur-concave in p_X. ∎

VI. Concluding remarks

We illustrated in this paper how the principal inertia-decomposition of the joint distribution matrix can be applied to derive useful bounds for the average estimation error. The principal inertias are a more refined metric of the correlation between X and Y than, say, mutual information. Furthermore, the principal inertia components can be used in metrics, such as k-correlation, that share several properties with mutual information (e.g. convexity).

Furthermore, we also introduced a general method for bounding the average estimation error of functions of a hidden random variable. This method depends on the Schur-concavity of a lower bound for the error-rate function. We proved that the $e_{\mathcal{I}}(p_X,\theta)$ itself is Schur-concave whenever the measure of information is concave in p_X. It remains to be shown if $e_{\mathcal{I}}(p_X,\theta)$ is Schur-concave for more general measures of information (such as k-correlation), and finding the necessary and sufficient conditions for Schur-concavity would be of both theoretical and practical interest.

Finally, the creation of bounds for P_e and $P_{e,M}$ given constraints on different metrics of information is a promising avenue of research. Most information-theoretic lower bounds for the average estimation error are based on mutual information. However, in statistics, a wide range of metrics are used to estimate the information between an observed and a hidden variable. Relating such metrics with the fundamental limits of inference is relevant for practical applications in both security and machine learning.

[5]The pmf of U is $p_U(1)=\sum_{i=1}^{m-M+1}p_X(i)$ and $p_U(k)=p_X(m-M+k)$ for $k=2,\dots,M$.

APPENDIX A
PROOF OF THEOREM 3

Theorem 3 follows directly from the next two lemmas.

Lemma 2. *Let the marginal distribution $\mathbf{p}_X$ and the principal inertias $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ be given, where $d = m - 1$. Then for any $P_{X,Y} \in \mathcal{R}(\mathbf{p}_X, \boldsymbol{\lambda})$, $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq p_X(2)$*

$$P_e \geq 1 - \beta - \sqrt{f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) + \sum_{i=1}^{m} \left([p_X(i) - \beta]^+\right)^2},$$

where

$$\begin{aligned} f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) = &\sum_{i=2}^{d+1} p_X(i)(\lambda_{i-1} + c_i - c_{i-1}) \\ &+ p_X(1)(c_1 + \alpha) - \alpha \mathbf{p}_X^T \mathbf{p}_X \ , \end{aligned} \tag{23}$$

and $c_i = [\lambda_i - \alpha]^+$ for $i = 1, \ldots, d$ and $c_{d+1} = 0$.

Proof: Let X and Y have a joint distribution matrix $P_{X,Y}$ with marginal p_X and principal inertias individually bounded by $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$. We assume without loss of generality that $d = m - 1$, where $|\mathcal{X}| = m$. This can always be achieved by adding inertia components equal to 0.

Consider $X \to Y \to \hat{X}$, where $\hat{X}$ is the estimate of X from Y. The mapping from Y to $\hat{X}$ can be described without loss of generality by a $|\mathcal{Y}| \times |\mathcal{X}|$ row stochastic matrix, denoted by F, where the (i, j)-th entry is the probability $p_{\hat{X}|Y}(j|i)$. The probability of correct estimation P_c is then

$$P_c = \Pr\left\{\hat{X} = X\right\} = \operatorname{tr}\left(P_{X,X'}\right),$$

where $P_{X,X'} \triangleq P_{X,Y} F$.

The matrix $P_{X,X'}$ can be decomposed according to (5), resulting in

$$P_c = \operatorname{tr}\left(D_X^{1/2} U \widetilde{\Sigma} V^T D_{X'}^{1/2}\right) = \operatorname{tr}\left(\widetilde{\Sigma} V^T D_{X'}^{1/2} D_X^{1/2} U\right), \tag{24}$$

where

$$\begin{aligned} U &= \left[\mathbf{p}_X^{1/2} \ \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_m\right], \\ V &= \left[\mathbf{p}_{X'}^{1/2} \ \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_m\right], \\ \widetilde{\Sigma} &= \operatorname{diag}\left(1, \widetilde{\lambda}_1^{1/2}, \ldots, \widetilde{\lambda}_d^{1/2}\right), \\ D_{X'} &= \operatorname{diag}\left(\mathbf{p}_{X'}\right), \end{aligned}$$

and $\widetilde{U}$ and $\widetilde{V}$ are orthogonal matrices. The probability of correct detection can be written as

$$\begin{aligned} P_c &= \mathbf{p}_X^T \mathbf{p}_{X'} + \sum_{k=2}^{m} \sum_{i=1}^{m} \left(\widetilde{\lambda}_{k-1} p_X(i) p_{X'}(i)\right)^{1/2} u_{k,i} v_{k,i} \\ &= \mathbf{p}_X^T \mathbf{p}_{X'} + \sum_{k=2}^{m} \sum_{i=1}^{m} \widetilde{\lambda}_{k-1}^{1/2} \widetilde{u}_{k,i} \widetilde{v}_{k,i} \end{aligned}$$

where $u_{k,i} = [\mathbf{u}_k]_i$, $v_{k,i} = [\mathbf{v}_k]_i$, $\widetilde{u}_{k,i} = p_X(i) u_{k,i}$ and $\widetilde{v}_{k,i} = p_{X'}(i) v_{k,i}$. Applying the Cauchy-Schwarz inequality twice, we obtain

$$\begin{aligned} P_c &\leq \mathbf{p}_X^T \mathbf{p}_{X'} + \sum_{i=1}^{m} \left(\sum_{k=2}^{m} \widetilde{v}_{k,i}^2\right)^{1/2} \left(\sum_{k=2}^{m} \widetilde{\lambda}_{k-1} \widetilde{u}_{k,i}^2\right)^{1/2} \\ &= \mathbf{p}_X^T \mathbf{p}_{X'} + \sum_{i=1}^{m} \left(p_{X'}(i)(1 - p_{X'}(i)) \sum_{k=2}^{m} \widetilde{\lambda}_{k-1} \widetilde{u}_{k,i}^2\right)^{1/2} \\ &\leq \mathbf{p}_X^T \mathbf{p}_{X'} + \left(1 - \sum_{i=1}^{m} p_{X'}^2(i)\right)^{1/2} \left(\sum_{i=1}^{m} \sum_{k=2}^{m} \widetilde{\lambda}_{k-1} \widetilde{u}_{k,i}^2\right)^{1/2} \end{aligned} \tag{25}$$

Let $\overline{U} = [\mathbf{u}_2 \cdots \mathbf{u}_m]$ and $\Sigma = \operatorname{diag}\left(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_d\right)$. Then

$$\begin{aligned} \sum_{i=1}^{m} \sum_{k=2}^{m} \widetilde{\lambda}_{k-1} \widetilde{u}_{k,i}^2 &= \operatorname{tr}\left(\Sigma \overline{U}^T D_X \overline{U}\right) \\ &\leq \sum_{k=1}^{d} \sigma_k \widetilde{\lambda}_k, \\ &\leq \sum_{k=1}^{d} \sigma_k \lambda_k. \end{aligned} \tag{26}$$

where $\sigma_k = \Lambda_k(\overline{U}^T D_X \overline{U})$. The first inequality follows from the application of Von-Neumman's trace inequality and the fact that $\overline{U}^T D_X \overline{U}$ is symmetric and positive semi-definite. The second inequality follows by observing that the principal inertias satisfy the data processing inequality and, therefore, $\widetilde{\lambda}_k \leq \lambda_k$.

We will now find an upper bound for (26) by bounding the eigenvalues σ_k. First, note that $\overline{U}\ \overline{U}^T = I - \mathbf{p}_X^{1/2} \left(\mathbf{p}_X^{1/2}\right)^T$ and consequently

$$\begin{aligned} \sum_{k=1}^{d} \sigma_k &= \operatorname{tr}\left(\overline{U}^T D_X \overline{U}\right) \\ &= \operatorname{tr}\left(D_X \left(I - \mathbf{p}_X^{1/2} \left(\mathbf{p}_X^{1/2}\right)^T\right)\right) \\ &= 1 - \sum_{i=1}^{m} p_X^2(i) \ . \end{aligned} \tag{27}$$

Second, note that $\overline{U}^T D_X \overline{U}$ is a principal submatrix of $U^T D_X U$, formed by removing the first row and columns of $U^T D_X U$. It then follows from Cauchy's interlacing theorem that

$$p_X(m) \leq \sigma_{m-1} \leq p_X(m-1) \leq \cdots \leq p_X(2) \leq \sigma_1 \leq p_X(1). \tag{28}$$

Combining the restriction (27) and (28), an upper bound for (26) can be found by solving the following linear program

$$\begin{aligned} \max_{\sigma_i} \quad & \sum_{i=1}^{d} \lambda_i \sigma_i \\ \text{subject to} \quad & \sum_{i=1}^{d} \sigma_i = 1 - \mathbf{p}_X^T \mathbf{p}_X, \\ & p_X(i+1) \leq \sigma_i \leq p_X(i), \ i = 1, \ldots, d \ . \end{aligned} \tag{29}$$

Let $\delta_i \triangleq p_X(i)-p_X(i+1)$ and $\gamma_i \triangleq \lambda_i p_X(i+1)$. The dual of (29) is

$$\begin{aligned} \min_{y_i,\alpha} \quad & \alpha\left(p_X(1)-\mathbf{p}_X^T\mathbf{p}_X\right)+\sum_{i=1}^{m-1}\delta_i y_i+\gamma_i \qquad (30)\\ \text{subject to} \quad & y_i \geq [\lambda_i-\alpha]^+ ,\ i=1,\ldots,d\ . \end{aligned}$$

For any given value of α, the optimal values of the dual variables y_i in (30) are

$$y_i=[\lambda_i-\alpha]^+=c_i,\ i=1,\ldots,d\ .$$

Therefore the linear program (30) is equivalent to

$$\min_{\alpha} f_0(\alpha,\mathbf{p}_X,\boldsymbol{\lambda}), \qquad (31)$$

where $f_0(\alpha,\mathbf{p}_X,\boldsymbol{\lambda})$ is defined in the statement of the theorem.

Denote the solution of (29) by $f_P^*(\mathbf{p}_X,\boldsymbol{\lambda})$ and of (30) by $f_D^*(\mathbf{p}_X,\boldsymbol{\lambda})$. It follows that (26) can be bounded

$$\begin{aligned} \sum_{k=1}^{d}\sigma_k\lambda_k &\leq f_P^*(\mathbf{p}_X,\boldsymbol{\lambda})\\ &= f_D^*(\mathbf{p}_X,\boldsymbol{\lambda})\\ &\leq f_0(\alpha,\mathbf{p}_X,\boldsymbol{\lambda})\ \forall\ \alpha\in\mathbb{R}. \qquad (32)\end{aligned}$$

We may consider $0\leq\alpha\leq 1$ in (32) without loss of generality.

Using (32) to bound (25), we find

$$P_c \leq \mathbf{p}_X^T\mathbf{p}_{X'}+\left[f_0(\alpha,\mathbf{p}_X,\boldsymbol{\lambda})\left(1-\sum_{i=1}^{m}p_{X'}^2(i)\right)\right]^{1/2} \qquad (33)$$

The previous bound can be maximized over all possible output distributions $p_{X'}$ by solving:

$$\begin{aligned} \max_{x_i} \quad & \left[f_0(\alpha,\mathbf{p}_X,\boldsymbol{\lambda})\left(1-\sum_{i=1}^{m}x_i^2\right)\right]^{1/2}+\sum_{i=1}^{m}p_X(i)x_i \qquad (34)\\ \text{subject to} \quad & \sum_{i=1}^{m}x_i=1,\\ & x_i\geq 0, i=1,\ldots,m\ . \end{aligned}$$

The dual function of (34) over the additive constraint is

$$\begin{aligned} L(\beta)=\max_{x_i\geq 0} \quad & \beta+\left[f_0(\alpha,\mathbf{p}_X,\boldsymbol{\lambda})\left(1-\sum_{i=1}^{m}x_i^2\right)\right]^{1/2}\\ & +\sum_{i=1}^{m}(p_X(i)-\beta)x_i\\ =\beta+&\sqrt{f_0(\alpha,\mathbf{p}_X,\boldsymbol{\lambda})+\sum_{i=1}^{m}\left([p_X(i)-\beta]^+\right)^2}. \qquad (35)\end{aligned}$$

Since $L(\beta)$ is an upper bound of (34) for any β and, therefore, is also an upper bound of (33), then

$$P_c \leq \beta+\sqrt{f_0(\alpha,\mathbf{p}_X,\boldsymbol{\lambda})+\sum_{i=1}^{m}\left([p_X(i)-\beta]^+\right)^2}. \qquad (36)$$

Note that we can consider $0\leq\beta\leq p_X(2)$ in (36), since $L(\beta)$ is increasing for $\beta>p_X(2)$. Taking $P_e=1-P_c$, the result follows. ∎

The next result tightens the bound introduced in lemma 2 by optimizing over all values of α.

Lemma 3. *Let $f_0^*(\mathbf{p}_X,\boldsymbol{\lambda})\triangleq\min_\alpha f_0(\alpha,\mathbf{p}_X,\boldsymbol{\lambda})$ and k^* be defined as in* (14)*. Then*

$$\begin{aligned} f_0^*(\mathbf{p}_X,\boldsymbol{\lambda})=&\sum_{i=1}^{k^*}\lambda_i p_X(i)+\sum_{i=k^*+1}^{m}\lambda_{i-1}p_X(i)\\ &-\lambda_{k^*}\mathbf{p}_X^T\mathbf{p}_X\ , \qquad (37)\end{aligned}$$

where $\lambda_m=0$.

Proof: Let $\mathbf{p}_X$ and $\boldsymbol{\lambda}$ be fixed, and $\lambda_k\leq\alpha\leq\lambda_{k-1}$, where we define $\lambda_0=1$ and $\lambda_m=0$. Then $c_i=\lambda_i-\alpha$ for $1\leq i\leq k-1$ and $c_i=0$ for $k\leq i\leq d$ in (23). Therefore

$$\begin{aligned} f_0(\alpha,\mathbf{p}_X,\boldsymbol{\lambda})=&\sum_{i=1}^{k-1}\lambda_i p_X(i)+\alpha p_X(k)\\ &+\sum_{i=k+1}^{m}\lambda_{i-1}p_X(i)-\alpha\mathbf{p}_X^T\mathbf{p}_X \qquad (38)\end{aligned}$$

Note that (38) is convex in α, and is decreasing when $p_X(k)-\mathbf{p}_X^T\mathbf{p}_X\leq 0$ and increasing when $p_X(k)-\mathbf{p}_X^T\mathbf{p}_X\geq 0$. Therefore, $f_0(\alpha,\mathbf{p}_X,\boldsymbol{\lambda})$ is minimized when $\alpha=\lambda_k$ such that $p_X(k)\geq\mathbf{p}_X^T\mathbf{p}_X$ and $p_X(k-1)\leq\mathbf{p}_X^T\mathbf{p}_X$. If $p_X(k)-\mathbf{p}_X^T\mathbf{p}_X\geq 0$ for all k (i.e. p_X is uniform), then we can take $\alpha=0$. The result follows. ∎

APPENDIX B
EQUIVALENT CHARACTERIZATIONS OF THE PRINCIPAL INERTIAS AND k-CORRELATION

In this appendix we discuss two distinct characterizations of the principal inertia components. The first characterization is based on the work of Gebelein [8] and the overview presented in [10]. The second characterization is based on the overview presented in [9], and is analogous to the definition of moments of inertia from classical mechanics.

A. Correlation characterization

Let $\mathcal{S}$ be a collection of random variables defined as

$$\begin{aligned}\mathcal{S}\triangleq\{(f(X),g(Y)):&\mathbb{E}\left[f(X)\right]=\mathbb{E}\left[g(Y)\right]=0,\\ &\mathbb{E}\left[f^2(X)\right]=\mathbb{E}\left[g^2(Y)\right]=1\}\ .\end{aligned}$$

Then, for $1\leq k\leq d$, we can compute the principal inertias recursively as

$$\begin{aligned}\lambda_k^{1/2}&=\max_{(f(X),g(Y))\in\mathcal{S}_k}\mathbb{E}\left[f(X)g(Y)\right],\\ (f_k(X),g_k(Y))&=\operatorname*{argmax}_{(f(X),g(Y))\in\mathcal{S}_k}\mathbb{E}\left[f(X)g(Y)\right],\end{aligned}$$

where $\mathcal{S}_1=\mathcal{S}$ and

$$\begin{aligned}\mathcal{S}_k=\{(f(X),g(Y))\in\mathcal{S}:&\mathbb{E}\left[f(X)f_i(X)\right]=0,\\ &\mathbb{E}\left[g(Y)g_i(Y)\right]=0, i=1,\ldots,k-1\}.\end{aligned}$$

for $2\leq k\leq d$. We can verify that $f_k(x)=a_{x,k}/p_X(x)$ and $g_k(x)=b_{y,k}/p_Y(y)$.

B. Spatial characterization

Let S be defined in equation (12). Then the square of the singular values of S are the principal inertias of $P_{X,Y}$ [6]. The decomposition of S can be interpreted as the moment of inertia of a set of masses located in discrete points in space, as described below. We will change the notation slightly in this appendix in order to make this analogy clear.

Consider an n-dimensional Euclidean space V with a symmetric positive definite form $Q = D_Y$. For $x, y \in V$ we let $\langle x, y\rangle = x^T Q y$, $\|x\|_Q = \sqrt{\langle x, x\rangle}$ and $d(x, y) = \|x - y\|_Q$.

Let $x_1, \ldots, x_m \in V$, where each point is $x_i = \mathbf{p}_{Y|X=i}$. We associate to each point x_i a mass $w_i = p_X(i)$, $1 \leq i \leq m$. The *barycenter* (center of mass) $\bar{x}$ of the points $x_1, \ldots, x_m$ is simply $\bar{x} = \mathbf{p}_Y$.

Let $G = P_{Y|X}$. If we translate the space so the barycenter $\bar{x}$ is the origin, the new coordinates of $x_1, \ldots, x_m$ are then the rows of $C = G - \mathbf{1}\bar{x}^T$, We denote $C^T = [c_1, \ldots, c_m]$. We define the *moment of inertia* $\mathcal{I}$ of the collection of m points as the weighted sum of the squared distances of each point to the barycenter:

$$\mathcal{I} \triangleq \sum_{i=1}^{m} w_i d^2(x_i - \bar{x}) \tag{39}$$

$$= \operatorname{tr}\left(D_X C Q C^T\right). \tag{40}$$

We now ask: What is the subspace $W^t \in V$ of dimension $t \leq m$ where the projection of $x_1, \ldots, x_m$ has the largest moment of inertia? To answer this question we need to determine a basis $a_1, \ldots, a_t$ of W^t. This is equivalent to solving the following optimization:

$$\mathcal{I}_t \triangleq \max_{a_1,\ldots,a_t} \sum_{j=1}^{d} \|D_X^{1/2} C Q a_j\|_2^2 \tag{41}$$

$$\text{s.t.} \|a_j\|_Q = 1, a_j \in V \;\; j = 1, \ldots, t$$

$$\langle a_i, a_j\rangle = 0, \;\; 1 \leq i < j \leq t$$

Note that $S = D_X^{1/2} C Q^{1/2}$, and the decomposition in (12) can be interpreted accordingly. The solution of (41) is exactly the sum of the square of the t largest singular values of S, which, in turn, is equal to $J_t(X;Y)$.

ACKNOWLEDGEMENT

The authors would like to thank Prof. Shafi Goldwasser and Prof. Yury Polyanskiy for the insightful discussions and suggestions throughout the course of this work.

REFERENCES

[1] S. Goldwasser and S. Micali, "Probabilistic encryption," *J. Comput. Syst. Sci.*, vol. 28, no. 2, pp. 270–299, Apr. 1984.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition*, 2nd ed. Wiley-Interscience, Jul. 2006.

[3] A. Rényi, "On measures of dependence," *Acta Math. Hung.*, vol. 10, no. 3-4, pp. 441–451, Sep. 1959.

[4] R. Ahlswede, "Extremal properties of rate distortion functions," *IEEE Trans. on Info. Theory*, vol. 36, no. 1, pp. 166–171, 1990.

[5] M. Greenacre, *Correspondence Analysis in Practice, Second Edition*, 2nd ed. Chapman and Hall/CRC, May 2007.

[6] M. J. Greenacre, *Theory and Applications of Correspondence Analysis*. Academic Pr, Mar. 1984.

[7] H. O. Hirschfeld, "A connection between correlation and contingency," in *Math Proc. Cambridge*, vol. 31, 1935, pp. 520–524.

[8] H. Gebelein, "Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung," *ZAMM-Z. Angew. Math. Me.*, vol. 21, no. 6, pp. 364–379, 1941.

[9] M. Greenacre and T. Hastie, "The geometric interpretation of correspondence analysis," *J. Am. Stat. Assoc.*, vol. 82, no. 398, pp. 437–447, Jun. 1987.

[10] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover," arXiv e-print 1304.6133, Apr. 2013.

[11] W. Kang and S. Ulukus, "A new data processing inequality and its applications in distributed source and channel coding," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 56–69, 2011.

[12] A. Guntuboyina, "Lower bounds for the minimax risk using -divergences, and applications," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2386–2399, 2011.

[13] A. Guntuboyina, S. Saha, and G. Schiebinger, "Sharp inequalities for f-divergences," *arXiv:1302.0336*, Feb. 2013.

[14] V. Doshi, D. Shah, M. Médard, and M. Effros, "Functional compression through graph coloring," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3901 –3917, Aug. 2010.

[15] A. Orlitsky and J. Roche, "Coding for computing," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 903 –917, Mar. 2001.

[16] G. R. Kumar and T. A. Courtade, "Which boolean functions are most informative?" *arXiv:1302.2512*, Feb. 2013.

[17] M. Bellare, S. Tessaro, and A. Vardy, "Semantic security for the wiretap channel," in *Advances in Cryptology – CRYPTO 2012*, ser. Lecture Notes in Comput. Sci. Springer, Jan. 2012, no. 7417, pp. 294–311.

[18] Y. Liang, H. V. Poor, and S. Shamai (Shitz), "Information theoretic security," *Found. Trends Commun. Inf. Theory*, vol. 5, no. 4, pp. 355–580, Apr. 2009.

[19] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge University Press, Oct. 2012.

[20] K. Fan, "On a theorem of Weyl concerning eigenvalues of linear transformations I," *P. Natl. Acad. Sci. USA*, vol. 35, no. 11, pp. 652–655, Nov. 1949.

[21] M. L. Overton and R. S. Womersley, "On the sum of the largest eigenvalues of a symmetric matrix," *SIAM J. Matrix Anal. A.*, vol. 13, no. 1, pp. 41–45, Jan. 1992.

[22] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, UK; New York: Cambridge University Press, 2004.

[23] C. Deniau, G. Oppenheim, and J. P. Benzécri, "Effet de l'affinement d'une partition sur les valeurs propres issues d'un tableau de correspondance," *Cahiers de l'analyse des données*, vol. 4, no. 3, pp. 289–297.

[24] Y. Polyanskiy, "Hypothesis testing via a comparator." [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/htstruct.pdf

[25] A. Guntuboyina, "Minimax lower bounds," Ph.D., Yale University, United States – Connecticut, 2011.

[26] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: theory of majorization and its applications*. New York: Springer Series in Statistics, 2011.

[27] R. G. Gallager, *Information theory and reliable communication*. New York: Wiley, 1968.