Massachusetts Institute of Technology

Principal Inertia Components and Applications

Flavio P. Calmon, Ali Makhdoumi, Muriel Médard,
Mayank Varia, Mark Christiansen, Ken R. Duffy *

Abstract

We explore properties and applications of the Principal Inertia Components (PICs) between two discrete random variables X and Y. The PICs lie in the intersection of information and estimation theory, and provide a fine-grained decomposition of the dependence between X and Y. Moreover, the PICs describe which functions of X can or cannot be reliably inferred (in terms of MMSE) given an observation of Y. We demonstrate that the PICs play an important role in information theory, and they can be used to characterize information-theoretic limits of certain estimation problems. In privacy settings, we prove that the PICs are related to fundamental limits of perfect privacy.
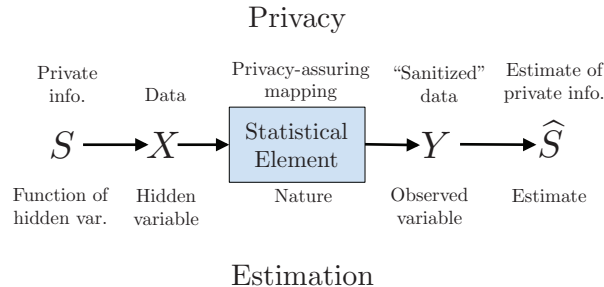
Contents

Figure 1: Problem central to both estimation and privacy.

1 Introduction

There is a fundamental limit to how much we can learn from data. The problem of determining which functions of a hidden variable can or cannot be estimated from a noisy observation is at the heart of estimation, statistical learning theory [1], and numerous other applications of interest. For example, one of the main goals of prediction is to determine a function of a hidden variable that can be reliably inferred from the output of a system.

Privacy and security applications are concerned with the inverse problem: guaranteeing that a certain set of functions of a hidden variable *cannot* be reliably estimated given the output of a system. Examples of such functions are the identity of an individual whose information is contained in a supposedly anonymous dataset [2], sensitive information of a user who joined a database [3,4], the political preference of a set of users who disclosed their movie ratings [5–7], among others. On the one hand, estimation methods attempt to extract as much information as possible from data. On the other hand, privacy-assuring systems seek to minimize the information about a secret variable that can be reliably estimated from disclosed data. The relationship between privacy and estimation is similar to the one noted by Shannon between cryptography and communication [8]: they are connected fields, but with different goals. As illustrated in Fig. 1, estimation and privacy are concerned with the same fundamental problem, and can be simultaneously studied through an information-theoretic lens.

In this paper, we discuss information-theoretic tools to address challenges in privacy, security and estimation. By studying fundamental models that are common to these fields, we derive information-theoretic metrics and associated results that simultaneously (i) delineate the fundamental limits of estimation and (ii) characterize the security properties of privacy-assuring systems.

We focus on the question that is central to privacy and estimation (illustrated in Fig. 1): How well can a random variable S, that is correlated with a hidden variable X, be estimated given an observation of Y? The information-theoretic metrics presented here seek to quantify properties of the random mapping from X to Y that can be translated into bounds on the error of estimating S given an observation of Y. These bounds, which are often at the heart of information-theoretic converse proofs [9], provide universal, algorithm-independent guarantees on what can (or cannot) be learned from Y. With a characterization of these bounds in hand, we study properties of random mappings that seek to achieve privacy in terms of how well an adversary can estimate a secret S given the output of the mapping Y.

The results in this paper are situated at the intersection of estimation, privacy and security. We derive a set of general sharp bounds on how well certain classes of functions of a hidden variable can(not) be estimated from a noisy observation. The bounds are expressed in terms of different information metrics of the joint distribution of the hidden and observed variables, and provide converse (negative) results: If an information metric is small, then not only the hidden variable cannot be reliably estimated, but also any non-trivial function of the hidden variable cannot be inferred with probability of error or mean-squared error smaller than a certain threshold.

These results are applicable to both estimation and privacy. For estimation and statistical learning theory, they shed light on the fundamental limits of learning from noisy data, and can help guide the design of practical learning algorithms. In particular, the converse bounds can be used to derive minimax lower bounds (the same way Fano-style inequalities are used [10]). Furthermore, as illustrated in this paper, the proposed bounds are useful for creating security and privacy metrics, for characterizing the inherent trade-off between privacy and utility in statistical data disclosure problems and for studying the fundamental limits of perfect privacy. The tools used to derive the converse bounds are based on a set of statistics known as the Principal Inertia Components (PICs).

1.1 Principal Inertia Components

The PICs provide a fine-grained decomposition of the dependence between two random variables. Well-studied statistical methods for estimating the PICs [11,12] can lead to results on the (im)possibility of estimating a large classes of functions by using bounds based on the PICs and standard statistical tests. We show how PICs can be used to characterize the information-theoretic limits of certain estimation problems. The PICs generalize other measures that are used in information theory, such as maximal correlation [13] and χ^2-dependence [14]. The largest and smallest PIC play an important role in estimation and privacy (discussed in Sections 4 and 5). We also study properties of the sum of the k largest principal inertia components. Below we list a few key properties of the PICs studied in this paper.

1. **Overview of the PICs:** We present an overview of the PICs and their different interpretations, summarized in Theorem 1. For two discrete random variables X and Y, we denote the k largest PICs by $\lambda_1(X;Y), \lambda_2(X;Y), \ldots, \lambda_k(X;Y)$.

2. **Sum of the PICs:** We propose a measure of dependence termed k-correlation which is defined as the sum of the k largest PICs, i.e., $\mathcal{J}_k(X;Y) \triangleq \sum_{i=1}^{k} \lambda_i(X;Y)$. This metric satisfies two key properties: (i) convexity in $p_{Y|X}$ (Theorem 2); (ii) Data Processing Inequality (Theorem 3). The latter is also satisfied by $\lambda_1(X;Y), \ldots, \lambda_d(X;Y)$ individually, where $d = \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1$. Both maximal correlation and the χ^2-dependence between X and Y are special cases of k-correlation, with $\mathcal{J}_1(X;Y) = \rho_m(X;Y)^2$ and $\mathcal{J}_d(X;Y) = \chi^2(X;Y)$ (cf. notation in Section 1.3).

3. **Largest PIC** The largest PIC satisfies $\lambda_1(X;Y) = \rho_m(X;Y)^2$, where $\rho_m(X;Y)$ is the *maximal correlation* between X and Y, defined as [15]

$$\rho_m(X;Y) \triangleq \max_{\substack{\mathbb{E}[f(X)]=\mathbb{E}[g(Y)]=0 \\ \mathbb{E}[f(X)^2]=\mathbb{E}[g(Y)^2]=1}} \mathbb{E}\left[f(X)g(Y)\right]. \tag{1}$$

   We show that both the probability of error and the minimum mean-squared error (MMSE) of estimating any function of a hidden variable X given an observation Y are closely related to the largest PIC.

   By making use of the fact that the PICs satisfy the Data Processing Inequality (DPI), we are able to derive a family of bounds for the smallest average error of estimating X having observed Y $P_e(X|Y)$ (cf. (5) and notation in Section 1.3) in terms of the marginal distribution of X, p_X, and $\lambda_1(X;Y), \ldots, \lambda_d(X;Y)$, described in Theorem 6. This result sheds light on the relationship of $P_e(X|Y)$ with the PICs.

   One immediate consequence of Theorem 6 is a useful scaling law for $P_e(X|Y)$ in terms of the largest PIC, the maximal correlation. Let $X = 1$ be the most likely outcome for X. Corollary 4 proves that the advantage an adversary (who has access to Y) has of guessing X over guessing the most likely outcome $X = 1$ satisfies

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

4. **Smallest PIC** We show that the smallest PIC determines when perfect privacy, defined in Section 5, can be achieved with non-trivial utility in the model depicted in Fig. 1. More specifically, perfect privacy can be achieved with non-trivial utility if and only if the smallest PIC is 0 (Theorem 10).

1.2 Organization of the Paper

This paper is organized as follows. The rest of this section introduces notation and discusses related work. In Section 2, we present the PICs and their multiple characterizations (Theorem 1). We also introduce the definition of k-correlation, and demonstrate several properties of both k-correlation and, more broadly, the PICs, including convexity and the DPI. In Section 3, we apply the PICs to problems in information theory. In Section 4, we derive bounds on error probability and other estimation-theoretic results based on the PICs. Finally, in Section 5, we demonstrate how the PICs play an important role in privacy and can be used for determining privacy-assuring mappings. We first summarize the main results obtained by applying the PICs to information theory, estimation theory and privacy.

Applications to Information Theory

We present several distinct applications of the PICs to information theory. In Section 3.2, we demonstrate that the PICs correspond to the singular values of certain channel transformation matrices, and there

effect on input distributions to the channel bear an interpretation similar to that of filter coefficients in a linear filter [16]. This is illustrated through an example in binary additive noise channels, where we argue that the binary symmetric channel is akin to a low-pass filter. We show how the PICs, and particularly the largest PIC, can be used to derive bounds on information metrics between one-bit functions of a hidden variable X and a correlated observation Y. We apply these results to the "one-bit function conjecture" [17] We do not solve this conjecture here. Nevertheless, we present further evidence for its validity, and introduce another conjecture based on our results.

The new conjecture (cf. Conjecture 1) generalizes the "one-bit function conjecture". It states that, given a symmetric distribution $p_{X,Y}$, if we generate a new distribution $q_{X,Y}$ by making all the PICs of $p_{X,Y}$ equal to the largest one, then the new distribution is more informative about bits of X. By more informative, we mean that, for any 1-bit function b, $I(b(X); Y)$ is larger under $q_{X,Y}$ than under $p_{X,Y}$. Indeed, from an estimation-theoretic perspective, increasing the PICs imply that any function of X can be estimated with smaller MMSE when considering $q_{X,Y}$ than $p_{X,Y}$. Furthermore, in this case, we show that $q_{X,Y}$ is a q-ary symmetric channel. This conjecture, if proven, would imply as a corollary the original one-bit function conjecture.

We do show that our results on the PICs can be used to resolve the one-bit function conjecture in a specific setting in Section 3.6. Instead of considering the mutual information between $b(X)$ and Y, we study the mutual information between $b(X)$ and a one-bit estimator $\hat{b}(Y)$. We show in Theorem 5 that, when $\hat{b}(Y)$ is an unbiased estimator, the information that $\hat{b}(Y)$ carries about $b(X)$ can be upper-bounded for a range of dependence metrics (e.g. mutual information). This result also leads to bounds on estimation error probability.

Applications to Estimation Theory

In Section 4, we derive converse bounds on estimation error based on the PICs. In particular, we provide lower bounds on (i) the probability of correctly guessing a hidden variable X given an observation Y and (ii) on the MMSE of estimating X given Y. These results are stated in terms of the PICs between X and Y, and provide algorithm-independent bounds on estimation. We also extend these bounds to the functional setting, and show that the advantage over a random guess of correctly estimating a function of X given an observation of Y is upper-bounded by the largest PIC between X and Y. More specifically, we propose a family of lower bounds for the error probability of estimating X given Y based on the PICs of $p_{X,Y}$ and the marginal distribution of X in Theorems 6 and 9. We also extend these bounds for the probability of correctly estimating a function of the hidden variable X given an observation of Y.

These results are based on a more general framework for deriving bounds on error probability, discussed in Section 4.1. At the heart of this framework are rate-distortion (test-channel) formulations that allow bounds on information metrics to be translated into bounds on estimation. These formulations, in turn, are based on convex programs that minimize the average estimation error over all possible distributions that satisfy a bound on a given information metric. The solution of such convex programs are called the error-rate functions. We study extremal properties of error-rate function and, by revisiting a result by Ahlswede [18], we show how to extend the error-rate function to quantify not only the smallest average error of estimating a hidden variable, but also of estimating any function of a hidden variable.

Applications to Privacy

When referring to privacy in this paper, we consider the setting studied by Calmon and Fawaz in [19]. Using Fig. 1 as reference, we study the problem of disclosing data X to a third-party in order to derive some utility based on X. At the same time, some information correlated with X, denoted by S, is supposed to remain private. The engineering goal is to create a random mapping, called the privacy-assuring mapping, that transforms X into a new data Y that achieves a certain target utility, while minimizing the information revealed about S. For example, X can represent movie ratings that a user intends to disclose to a third-party in order to receive movie recommendations [5–7,20]. At the same time, the user may want to keep her political preference S secret. We allow the user to distort movie ratings in her data X in order to generate a new data Y. The goal would then be to find privacy-assuring mappings that minimize the number of distorted entries in Y given a privacy constraint (e.g. the third-party cannot guess S with significant advantage over a random guess). In general, X is not restricted to be the data of an individual user, and can also represent multidimensional data derived from different sources.

We present necessary and sufficient conditions for achieving perfect privacy while disclosing a non-trivial amount of useful information when both S and X have finite support $\mathcal{S}$ and $\mathcal{X}$, respectively. We

4

prove that the smallest PIC of $p_{S,X}$ plays a central role for achieving perfect privacy (i.e. $I(S;Y) = 0$): If $|\mathcal{X}| \leq |\mathcal{S}|$, then perfect privacy is achievable with $I(X;Y) > 0$ if and only if the smallest PIC of $p_{S,X}$ is 0. Since $I(S;Y) = 0$ if and only if $S \perp\!\!\!\perp Y$, this fundamental result holds for any privacy metric where statistical independence implies perfect privacy. We also provide an explicit lower bound for the amount of useful information that can be released while guaranteeing perfect privacy, and demonstrate how to construct $p_{Y|X}$ in order to achieve this bound.

In addition, we derive general bounds for the minimum amount of disclosed private information $I(S;Y)$ given that, on average, at least t bits of useful information are revealed, i.e. $I(X;Y) \geq t$. These bounds are sharp, and delimit the achievable privacy-utility region for the considered setting. Adopting an analysis related to the information bottleneck [21] and for characterizing linear contraction coefficients in strong DPIs in [22,23], we determine the smallest achievable ratio between disclosed private and useful information, i.e. $\inf_{p_{Y|X}} I(S;Y)/I(X;Y)$. We prove that this value is upper-bounded by the smallest PIC, and is zero if and only if the smallest PIC is zero. In this case, we present an explicit construction of a privacy-assuring mapping that discloses a non-trivial amount of useful information while guaranteeing perfect privacy. We also show that when the data is composed by multiple i.i.d. samples (S^n, X^n), the smallest PIC decreases exponentially in n. Consequently, as the number of samples n increases, we can achieve a more favorable trade-off between disclosing useful and private information. Finally, we motivate potential future applications of the PICs as a design driver for privacy assuring mappings in our final remarks in Section 6.

1.3 Notation

Capital letters (e.g. X and Y) are used to denote random variables, and calligraphic letters (e.g. $\mathcal{X}$ and $\mathcal{Y}$) denote sets. The exceptions are (i) $\mathcal{I}$, which will be used in Section 4 to denote a non-specified measure of dependence, and (ii) T, which will denote the conditional expectation operator (defined below). The support set of random variables X and Y are denoted by $\mathcal{X}$ and $\mathcal{Y}$, respectively. If X and Y have finite support sets $|\mathcal{X}| < \infty$ and $|\mathcal{Y}| < \infty$, then we denote the joint probability mass function (pmf) of X and Y as $p_{X,Y}$, the conditional pmf of Y given X as $p_{Y|X}$, and the marginal distributions of X and Y as p_X and p_Y, respectively. We denote the fact that X is distributed according to p_X by $X \sim p_X$. When $p_{X,Y,Z}(x,y,z) = p_X(x)p_{Y|X}(y|x)p_{Z|Y}(z|y)$ (i.e. X, Y, Z form a Markov chain), we write $X \to Y \to Z$. We denote independence of two random variables X and Y by $X \perp\!\!\!\perp Y$.

For positive integers j, k, n, $j \leq k$, we define $[n] \triangleq \{1, \ldots, n\}$ and $[j, k] \triangleq \{j, j+1, \ldots, k\}$. For any $x \in \mathbb{R}$, $[x]^+$ is defined as x if $x \geq 0$ and 0 otherwise. Matrices are denoted in bold capital letters (e.g. $\mathbf{X}$) and vectors in bold lower-case letters (e.g. $\mathbf{x}$). The (i,j)-th entry of a matrix $\mathbf{X}$ is given by $[\mathbf{X}]_{i,j}$. Furthermore, for $\mathbf{x} \in \mathbb{R}^n$, we let $\mathbf{x} = (x_1, \ldots, x_n)$. We denote by $\mathbf{1}$ the vector with all entries equal to 1, and the dimension of $\mathbf{1}$ will be clear from the context. The singular values of a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ are denoted by $\sigma_1(\mathbf{X}), \ldots, \sigma_m(\mathbf{X})$. For a matrix $\mathbf{X}$, we denote its k-th Ky Fan norm [24, Eq. (7.4.8.1)] by $\|\mathbf{X}\|_k \triangleq \sum_{i=1}^{k} \sigma_i(\mathbf{X})$.

For a random variable X with discrete support and $X \sim p_X$, the entropy of X is given by

$$H(X) \triangleq -\mathbb{E}\left[\log\left(p_X(X)\right)\right].$$

If Y has a discrete support set and $X, Y \sim p_{X,Y}$, the mutual information between X and Y is

$$I(X;Y) \triangleq \mathbb{E}\left[\log\left(\frac{p_{X,Y}(X,Y)}{p_X(X)p_Y(Y)}\right)\right].$$

The basis of the logarithm will be clear from the context. The χ^2-information between X and Y is

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

We denote the binary entropy function $h_b : [0,1] \to \mathbb{R}$ as

$$h_b(x) \triangleq -x\log x - (1-x)\log(1-x), \tag{2}$$

where, as usual, $0\log 0 \triangleq 0$.

Let X and Y be discrete random variables with finite support sets $\mathcal{X} = [m]$ and $\mathcal{Y} = [n]$, respectively. Then we define the joint distribution matrix $\mathbf{P}$ as an $m \times n$ matrix with $[\mathbf{P}]_{i,j} \triangleq p_{X,Y}(i,j)$. We denote by $\mathbf{p}_X$ (respectively, $\mathbf{p}_Y$) the vector with i-th entry equal to $p_X(i)$ (resp. $p_Y(i)$). $\mathbf{D}_X = \text{diag}(\mathbf{p}_X)$ and $\mathbf{D}_Y = \text{diag}(\mathbf{p}_Y)$ are matrices with diagonal entries equal to $\mathbf{p}_X$ and $\mathbf{p}_Y$, respectively, and all other entries equal to 0. The matrix $\mathbf{P}_{Y|X} \in \mathbb{R}^{m \times n}$ is defined as $[\mathbf{P}_{Y|X}]_{i,j} \triangleq p_{Y|X}(j|i)$. Note that $\mathbf{P} = \mathbf{D}_X \mathbf{P}_{Y|X}$.

5

For any real-valued random variable X, we denote the L_p-norm of X as

$$\|X\|_p \triangleq \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p}.$$

The set of all functions that when composed with a random variable X with distribution p_X result in an L_2-norm smaller than 1 is given by

$$\mathcal{L}_2(p_X) \triangleq \{f : \mathcal{X} \to \mathbb{R} \mid \|f(X)\|_2 \le 1\}. \tag{3}$$

The operators $T_X : \mathcal{L}_2(p_Y) \to \mathcal{L}_2(p_X)$ and $T_Y : \mathcal{L}_2(p_X) \to \mathcal{L}_2(p_Y)$ denote conditional expectation, where

$$(T_X g)(x) = \mathbb{E}\left[g(Y)|X = x\right] \text{ and } (T_Y f)(y) = \mathbb{E}\left[f(X)|Y = y\right], \tag{4}$$

respectively. Observe that T_X and T_Y are adjoint operators.

For X and Y with discrete support sets, we denote by $P_e(X|Y)$ the smallest average probability of error of estimating X given an observation of Y, defined as

$$P_e(X|Y) = \min_{X \to Y \to \hat{X}} \Pr\{X \ne \hat{X}\}, \tag{5}$$

where the minimum is taken over all distributions $p_{\hat{X}|Y}$ such that $X \to Y \to \hat{X}$. The advantage of correctly estimating X given an observation of Y over a random guess is defined as:

$$\mathsf{Adv}(X|Y) = 1 - P_e(X|Y) - \max_{x \in \mathcal{X}} p_X(x). \tag{6}$$

The MMSE of estimating X from an observation of Y is given by

$$\mathsf{mmse}(X|Y) \triangleq \min_{X \to Y \to \hat{X}} \mathbb{E}\left[(X - \hat{X})^2\right],$$

where the minimum is taken over all distributions $p_{\hat{X}|Y}$ such that $X \to Y \to \hat{X}$. Note that, from Jensen's inequality, it is sufficient to consider $\hat{X}$ a deterministic mapping of Y. For any $X \to Y \to g(Y)$ with $\|g(Y)\|_2 = \alpha$ and $\|X\|_2 = \sigma$

$$\mathbb{E}\left[(X - g(Y))^2\right] \ge \sigma^2 + \alpha^2 - 2\alpha\|\mathbb{E}\left[X|Y\right]\|_2,$$

with equality if and only if $g(Y)$ is proportional to $\mathbb{E}\left[X|Y\right]$. Minimizing the right-hand side over all α, we find that the MMSE estimator of X from Y is $g(y) = \mathbb{E}\left[X|Y = y\right]$, and

$$\mathsf{mmse}(X|Y) = \|X\|_2^2 - \|\mathbb{E}\left[X|Y\right]\|_2^2. \tag{7}$$

For a given joint distribution $p_{X,Y}$ and corresponding joint distribution matrix $\mathbf{P}$, the set of all vectors contained in the unit cube in $\mathbb{R}^n$ that satisfy $\|\mathbf{P}\mathbf{x}\|_1 = a$ is given by

$$\mathcal{C}^n(a, \mathbf{P}) \triangleq \{\mathbf{x} \in \mathbb{R}^n | 0 \le x_i \le 1, \|\mathbf{P}\mathbf{x}\|_1 = a\}. \tag{8}$$

We represent the set of all $m \times n$ probability distribution matrices by $\mathcal{P}_{m,n}$.

For $x^n \in \{-1, 1\}^n$ and $\mathcal{S} \subseteq [n]$,

$$\chi_{\mathcal{S}}(x^n) \triangleq \prod_{i \in \mathcal{S}} x_i \tag{9}$$

(we consider $\chi_{\emptyset}(x) = 1$). For $y^n \in \{-1, 1\}^n$, $a^n = x^n \oplus y^n$ is the vector resulting from the entrywise product of x^n and y^n, i.e. $a_i = x_i y_i$, $i \in [n]$.

Given two probability distributions p_X and q_X and $f(t)$ a smooth convex function defined for $t > 0$ with $f(1) = 0$, the f-divergence is defined as [25]

$$D_f(p_X\|q_X) \triangleq \sum_x q_X(x) f\left(\frac{p_X(x)}{q_X(x)}\right). \tag{10}$$

The f-information is given by

$$I_f(X; Y) \triangleq D_f(p_{X,Y}\|p_X p_Y). \tag{11}$$

When $f(x) = x \log(x)$, then $I_f(X; Y) = I(X; Y)$. A study of information metrics related to f-information was given in [26] in the context of channel coding converses.

1.4 Related Work

The joint distribution matrix $\mathbf{P}$ can be viewed as a contingency table and decomposed using standard techniques from correspondence analysis [11,27]. For an overview of correspondence analysis, we refer the reader to [28]. The term "principal inertia components", used here, is borrowed from the correspondence analysis literature [11]. However, the study of the PICs of the joint distribution of two random variables or, equivalently, the spectrum of the conditional expectation operator, predates correspondence analysis, and goes back to the work of Hirschfeld [29], Gebelein [30], Sarmanov [31] and Rényi [15], having also appeared in the work of Witsenhausen [32] and Ahlswede and Gács [22]. The PICs are also related to strong DPIs and contraction coefficients, being recently investigated by Anantharam *et al.* [23], Polyanskiy [33], Raginsky [34], Calmon *et al.* [35], Makur and Zheng [36], among others. Recently, Liu *et al.* [37] provided a unified perspective on several functional inequalities used in the study of strong DPIs and hypercontractivity. The PICs also play a role in Euclidean Information Theory [38], since they related to χ^2-divergence and, consequently, to local approximations of mutual information and related measures.

The largest principal inertia component is equal to $\rho_m(X;Y)^2$, 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 [15,31]). Ahslwede and Gács studied maximal correlation in the context of contraction coefficients in strong data processing inequalities [22], and more recently Anantharam *et al.* presented in [23] an overview of different characterizations of maximal correlation, as well as its application in information theory. Estimating the maximal correlation is also the goal of the Alternating Conditional Expectation (ACE) algorithm introduced by Breiman and Friedman [12], further analyzed by Buja [39], and recently investigated in [40].

The DPI for the PICs was shown by Kang and Ulukus in [41, Theorem 2] in a different setting than the one considered here. Kang and Ulukus made 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.

Lower bounds on the average estimation error can be found using Fano-style inequalities. Recently, Guntuboyina *et al.* ([42,43]) 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 $P_e(X|Y)$.

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 [44,45]. 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 [17] 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 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 seems to be non-trivial. Further investigations of this problem was done in [46–49]. In particular, [46] studies a related problem in a continuous setting by considering that X and Y are Gaussian random vectors. Recently, Samorodnitsky [50] presented a proof of the conjecture in the high noise regime.

Information-theoretic formulations for privacy have appeared in [51–55]. For an overview, we refer the reader to [19,53] and the references therein. The privacy against statistical inference framework considered here was further studied in [5,6,56]. The results presented in this paper are closely connected to the study of hypercontractivity coefficients and strong data processing results, such as in [22,23,33,34,57]. PIC-based analysis were used in the context of security in [58,59]. Extremal properties of privacy were also investigated in [60,61], and in particular [62] builds upon some of the results introduced here. For more details on designing privacy-assuring mappings and applications with real-world data, we refer the reader to [5–7,19,20,63].

We note that the privacy against statistical inference setting is related to differential privacy [3,4]. In the classic differential privacy setting, the output of a statistical query over a database is masked against small perturbations of the data contained in the database. Assuming this centralized statistical database setting, the private variable S can represent an individual user's entry to the database, and the variable X the output of a query over the database. Unlike in differential privacy, here we consider an additional distortion constraint, which can be chosen according to the application at hand. In the privacy funnel setting [63], the distortion constraint is given in terms of the mutual information between X and the perturbed query output Y. Connections between differential privacy and the privacy setting depicted in Fig. 1 as well as connections between differential privacy and PICs are studied in [19,64].
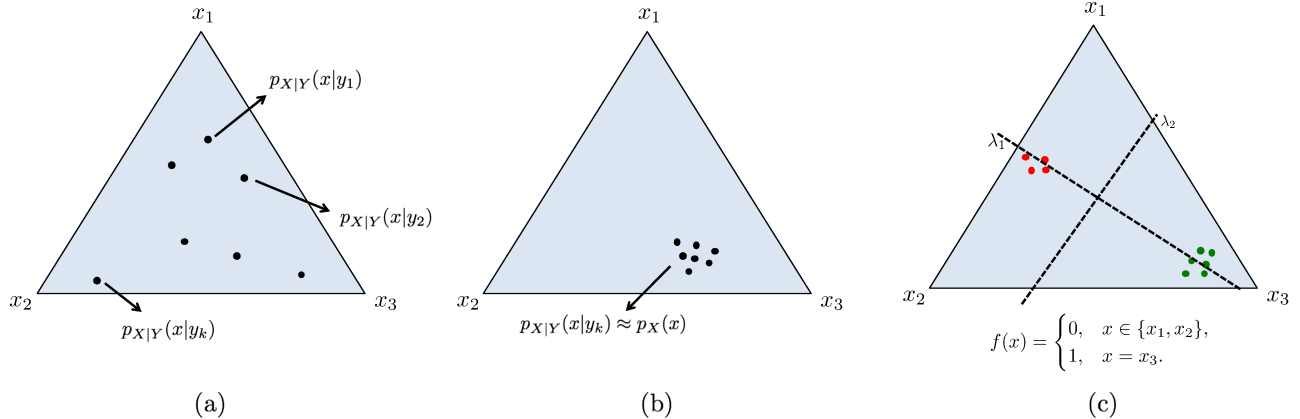
Figure 2: Geometric interpretation of the PICs for $\mathcal{X} = \{x_1, x_2, x_3\}$ and X uniformly distributed. In (a), each point on the simplex corresponds to a posterior distribution $p_{X|Y}(\cdot|y)$ induced on $\mathcal{X}$ by an observation of $Y = y_k$. If all the posterior distribution points are close together (b), then X and Y are approximately independent. If these points are far apart (c), then there may exist a function of X that can be approximately reliably estimated given an observation of Y (in this case, a binary function). The PICs can be intuitively understood as a measure of inertia of the posterior distribution vectors on the simplex.

2   Principal Inertia Components

We introduce in this section the Principal Inertia Components (PICs) of the joint distribution of two random variables X and Y. The PICs provide a fine-grained decomposition of the statistical dependence between X and Y, and are dependence measures that lie in the intersection of information and estimation theory. The PICs possess several desirable information-theoretic properties (e.g. satisfy the DPI, convexity, tensorization, etc.), and describe which functions of X can or cannot be reliably inferred (in terms of MMSE) given an observation of Y. The latter interpretation is discussed in more detail in Section 4.

2.1   A Geometric Interpretation of the PICs

We give an intuitive geometric interpretation of the PICs before presenting their formal definition in the next section. Let X and Y be related through a conditional distribution (channel), denoted by $p_{Y|X}$. For each $y \in \mathcal{Y}$, $p_{X|Y}(\cdot|y)$ will be a vector on the $|\mathcal{X}|$-dimensional simplex, and the position of these vectors on the simplex will determine the nature of the relationship between X and Y (Fig. 2). If $p_{X|Y}$ is fixed, what can be learned about X given an observation of Y, or the degree of accuracy of what can be inferred about X *a posteriori*, will then depend on the marginal distribution p_Y. The value $p_Y(y)$, in turn, ponderates the corresponding vector $p_{X|Y}(\cdot|y)$ akin to a mass. As a simple example, if $|\mathcal{X}| = |\mathcal{Y}|$ and the vectors $p_{X|Y}(\cdot|y)$ are located on distinct corners of the simplex, then X can be perfectly learned from Y. As another example, assume that the vectors $p_{X|Y}(\cdot|y)$ can be grouped into two clusters located near opposite corners of the simplex. If the sum of the masses induced by p_Y for each cluster is approximately $1/2$, then one may expect to reliably infer on the order of 1 unbiased bit of X from an observation of Y.

The above discussion naturally leads to considering the use of techniques borrowed from classical mechanics. For a given inertial frame of reference, the mechanical properties of a collection of distributed point masses can be characterized by the moments of inertia of the system. The moments of inertia measure how the weight of the point masses is distributed around the center of mass. An analogous metric exists for the distribution of the vectors $p_{X|Y}$ and masses p_Y in the simplex, and it is the subject of study of a branch of applied statistics called *correspondence analysis* ([11,28]). In correspondence analysis, the joint distribution $p_{X,Y}$ is decomposed in terms of the PICs, which, in some sense, are analogous to the moments of inertia of a collection of point masses. For more related literature, we refer the reader back to Section 1.4.

8

2.2 Definition and Characterizations of the PICs

We start with the definition of principal inertia components. In this paper we focus on the discrete case, since two of our main goals are (i) derive lower bounds on average estimation error probability and (ii) apply these results to privacy, where private data is often categorical. In addition, tools from correspondence analysis [27] can be used for estimating the PICs in the discrete setting. Nevertheless, the definition below is not limited to discrete random variables, and can be directly extended to general probability measures under compactness of the operator $T_X T_Y$ (cf. [32, Section 3]).

Definition 1. Let X and Y be random variables with support sets $\mathcal{X}$ and $\mathcal{Y}$, respectively, and joint distribution $p_{X,Y}$. In addition, let $f_0 : \mathcal{X} \to \mathbb{R}$ and $g_0 : \mathcal{Y} \to \mathbb{R}$ be the constant functions $f_0(x) = 1$ and $g_0(y) = 1$. For $k \in \mathbb{Z}_+$, we (recursively) define

$$\lambda_k(X;Y) = \max \left\{ \mathbb{E}\left[f(X)g(Y)\right]^2 \,\middle|\, f \in \mathcal{L}_2(p_X), g \in \mathcal{L}_2(p_Y), \mathbb{E}\left[f(X)f_j(X)\right] = 0, \right.$$
$$\left. \mathbb{E}\left[g(Y)g_j(Y)\right] = 0, j \in \{0, \ldots, k-1\} \right\}, \tag{12}$$

where

$$(f_k, g_k) \triangleq \operatorname{argmax} \left\{ \mathbb{E}\left[f(X)g(Y)\right]^2 \,\middle|\, f \in \mathcal{L}_2(p_X), g \in \mathcal{L}_2(p_Y), \mathbb{E}\left[f(X)f_j(X)\right] = 0, \right.$$
$$\left. \mathbb{E}\left[g(Y)g_j(Y)\right] = 0, j \in \{0, \ldots, k-1\} \right\}. \tag{13}$$

The values $\lambda_k(X;Y)$ are called the *principal inertia components* (PICs) of $p_{X,Y}$. The functions f_k and g_k are called the *principal functions* of X and Y.

Observe that the PICs satisfy $\lambda_k(X;Y) \leq 1$, since $f_k \in \mathcal{L}_2(p_X)$ $g_k \in \mathcal{L}_2(p_Y)$ and

$$\mathbb{E}\left[f(X)g(Y)\right] \leq \|f(X)\|_2 \|g(Y)\|_2 \leq 1.$$

Thus, from Definition 1, $\lambda_{k+1}(X;Y) \leq \lambda_k(X;Y) \leq 1$. When both random variables X and Y have a finite support set, we have the following definition.

Definition 2. For $\mathcal{X} = [m]$ and $\mathcal{Y} = [n]$, let $\mathbf{P} \in \mathbb{R}^{m \times n}$ be a matrix with entries $[\mathbf{P}]_{i,j} = p_{X,Y}(i,j)$, and $\mathbf{D}_X \in \mathbb{R}^{m \times m}$ and $\mathbf{D}_Y \in \mathbb{R}^{n \times n}$ be diagonal matrices with diagonal entries $[\mathbf{D}_X]_{i,i} = p_X(i)$ and $[\mathbf{D}_Y]_{j,j} = p_Y(j)$, respectively, where $i \in [m]$ and $j \in [n]$. We define

$$\mathbf{Q} \triangleq \mathbf{D}_X^{-1/2} \mathbf{P} \mathbf{D}_Y^{-1/2}. \tag{14}$$

We denote the singular value decomposition of $\mathbf{Q}$ by $\mathbf{Q} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.

The next theorem provides four equivalent characterizations of the PICs.

Theorem 1. *The following characterizations of the PICs are equivalent:*

(1) The characterization given in Definition 1 where, for f_k and g_k given in (13), $g_k(Y) = \frac{\mathbb{E}[f_k(X)|Y]}{\|\mathbb{E}[f_k(X)|Y]\|_2}$ and $f_k(X) = \frac{\mathbb{E}[g_k(Y)|X]}{\|\mathbb{E}[g_k(Y)|X]\|_2}$.

(2) [32, Section 3] Consider the conditional expectation operator $T_Y : \mathcal{L}_2(p_X) \to \mathcal{L}_2(p_Y)$, defined in (4). Then

$$\left(1, \sqrt{\lambda_1(X;Y)}, \sqrt{\lambda_2(X;Y)}, \ldots \right)$$

are the singular values of T_Y.

(3) For any $k \in \mathbb{Z}_+$,

$$1 - \lambda_k(X;Y) = \min \left\{ \mathsf{mmse}(f(X)|Y) \,\middle|\, f \in \mathcal{L}_2(p_X), \|f(X)\|_2 = 1, \mathbb{E}\left[f(X)h_j(X)\right] = 0, j \in \{0, \ldots, k-1\} \right\}, \tag{15}$$

where

$$h_k \triangleq \operatorname{argmin} \left\{ \mathsf{mmse}(f(X)|Y) \,\middle|\, f \in \mathcal{L}_2(p_X), \|f(X)\|_2 = 1, \mathbb{E}\left[f(X)h_j(X)\right] = 0, j \in \{0, \ldots, k-1\} \right\}. \tag{16}$$

If $\lambda_k(X;Y)$ is unique, then $h_k = f_k$ given in (13).

Finally, if both $\mathcal{X}$ and $\mathcal{Y}$ are defined over finite supports, the following characterization is also equivalent.

(4) $\sqrt{\lambda_k(X;Y)}$ is the $(k+1)$-st largest singular value of $\mathbf{Q}$. The principal functions f_k and g_k in (13) correspond to the columns of the matrices $\mathbf{D}_X^{-1/2}\mathbf{U}$ and $\mathbf{D}_Y^{-1/2}\mathbf{V}$, respectively, where $\mathbf{Q} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}$.

Proof. We will prove that (1) $\iff$ (2), (1) $\iff$ (3), finally and (1) $\iff$ (4).

- (1) $\iff$ (2). First observe that for $f \in \mathcal{L}_2(p_X)$ and $g \in \mathcal{L}_2(p_Y)$

$$\mathbb{E}\left[f(X)g(Y)\right] = \mathbb{E}\left[g(Y)\mathbb{E}\left[f(X)|Y\right]\right] \leq \|g(Y)\|_2\|\mathbb{E}\left[f(X)|Y\right]\|_2 \leq \|\mathbb{E}\left[f(X)|Y\right]\|_2,$$

where the first inequality follows from the Cauchy-Schwarz inequality, with equality if and only if $g(Y) = \frac{\mathbb{E}[f(X)|Y]}{\|\mathbb{E}[f(X)|Y]\|_2}$. The equivalence then follows by noting that

$$\sqrt{\lambda_1(X;Y)} = \max_{\substack{\mathbb{E}[f(X)]=\mathbb{E}[g(X)]=0 \\ \|f(X)\|_2=\|g(Y)\|_2=1}} \mathbb{E}\left[f(X)g(Y)\right]$$

$$= \max_{\substack{\mathbb{E}[f(X)]=\mathbb{E}[g(X)]=0 \\ \|f(X)\|_2=\|g(Y)\|_2=1}} \mathbb{E}\left[\mathbb{E}\left[g(Y)f(X)|Y\right]\right]$$

$$= \max_{\substack{\mathbb{E}[f(X)]=0 \\ \|f(X)\|_2=1}} \|\mathbb{E}\left[f(X)|Y\right]\|_2, \tag{17}$$

where the last equality follows by setting $g(Y) = \frac{\mathbb{E}[f(X)|Y]}{\|\mathbb{E}[(X)|Y]\|_2}$. Inverting the roles of f and g, we find $f(X) = \frac{\mathbb{E}[g(Y)|X]}{\|\mathbb{E}[(Y)|X]\|_2}$. Since this last expression is the second largest singular value of the conditional expectation operator T_Y (the largest being 1), the result follows for $\lambda_1(X;Y)$. The equivalent result for the other PICs follows by adding orthogonality constraints and the min-max properties of singular values (cf. Rayleigh-Ritz Theorem [24, Theorem 4.2.2]).

- (1) $\iff$ (3). The result follows from $\lambda_k(X;Y) = \|\mathbb{E}\left[f_k(X)|Y\right]\|_2^2$ in (17) and by noting that the MMSE can be written as (7). Consequently, maximizing $\|\mathbb{E}\left[f(X)|Y\right]\|$ is equivalent to minimizing the MMSE in (15).

- (1) $\iff$ (4). Let $f \in \mathcal{L}_2(p_X)$ and $g \in \mathcal{L}_2(p_Y)$. Define the column-vectors $\mathbf{f} \triangleq (f(1),\ldots,f(m))^T$ and $\mathbf{g} \triangleq (g(1),\ldots,g(n))^T$. Then

$$\mathbb{E}\left[f(X)g(Y)\right] = \mathbf{f}^T\mathbf{P}\mathbf{g}$$

and

$$\mathbf{f}^T\mathbf{D}_X\mathbf{f} = \mathbf{g}^T\mathbf{D}_Y\mathbf{g} = 1.$$

For $\mathbf{Q} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ given in Definition 2, put $\mathbf{u} \triangleq \mathbf{U}^T\mathbf{D}_X^{1/2}\mathbf{f}$ and $\mathbf{v} \triangleq \mathbf{V}\mathbf{D}_Y^{1/2}\mathbf{g}$. Then $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$, and

$$\mathbb{E}\left[f(X)g(Y)\right] = \mathbf{u}^T\mathbf{\Sigma}\mathbf{v}.$$

The result then follows directly from the variational characterization of singular values [24, Theorem 7.3.8].

Assuming unique PICs, note that the column-vectors $(\mathbf{f}_0, \mathbf{f}_1, \ldots, \mathbf{f}_d)$ corresponding to the functions $(f_0, f_1, \ldots, f_d)$ are the first $d+1$ columns of $\mathbf{D}_X^{-1/2}\mathbf{U}$, and the column-vectors $(\mathbf{g}_0, \mathbf{g}_1, \ldots, \mathbf{g}_d)$ corresponding to the functions $(g_0, g_1, \ldots, g_d)$ are the first $d+1$ of $\mathbf{D}_Y^{-1/2}\mathbf{V}$. In addition, let $\mathbf{z}_k \in \mathbb{R}^n$ be the column vector with entries $\mathbb{E}\left[f_k(X)|Y=j\right]$. Then

$$\mathbf{z}_k = \mathbf{f}^T\mathbf{P}\mathbf{D}_Y^{-1} = \mathbf{f}_k^T\mathbf{D}_X^{1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{D}_Y^{-1/2} = \sqrt{\lambda_k(X;Y)}\mathbf{g}_k,$$

so $\lambda_k(X;Y) = \|\mathbb{E}\left[f_k(X)|Y\right]\|_2^2$ and once again we find $g_k(Y) = \frac{\mathbb{E}[f_k(X)|Y]}{\|\mathbb{E}[f_k(X)|Y]\|_2}$. $\qquad\square$

The previous theorem provides different operational characterization of the PICs. Characterization (1), presented in Definition 1 implies that the principal functions of X and Y are the solution to the following problem: Consider two parties, namely Alice and Bob, where Alice has access to an observation of X and Bob has access to an observation Y. Alice and Bob's goal is to produce zero-mean, unit variance functions $f(X)$ and $g(Y)$, respectively, that maximizes the correlation $\mathbb{E}\left[f(X)g(Y)\right]$ without any additional information beyond their respective observations of X and Y. The optimal choice of functions is f_1 and g_1, given in the theorem. Moreover,

$$\lambda_1(X;Y) = \rho_m(X;Y)^2.$$

10

Characterization (3) above proves that the PICs are the solution to another related question: Given a noisy observation Y of a hidden variable X, what is the unit-variance, zero-mean function of X that can be estimated with the smallest mean-squared error? It follows directly from (15) that the function is $f_1(X)$, and the minimum MMSE is $1 - \lambda_1(X;Y)$. Indeed, since they are orthonormal, the principal functions form a basis for the zero-mean functions in $\mathcal{L}_2(p_X)$ (we revisit this point in the Section 6). Characterization (4) lends itself to the geometric interpretation discussed in Section 2.1.

The next result states the well-known tensorization property the PICs between sequences of independent random variables (e.g. [23,32,65]). We present a proof of the discrete case here for the sake of completeness.

Lemma 1. *Let* $(X_1, Y_1) \perp\!\!\!\perp (X_2, Y_2)$, $d_1 = \min\{|\mathcal{X}_1|, |\mathcal{Y}_1|\} - 1 < \infty$ *and* $d_2 = \min\{|\mathcal{X}_2|, |\mathcal{Y}_2|\} - 1 < \infty$. *Then the PICs of* $p_{(X_1, X_2),(Y_1, Y_2)}$ *are* $\lambda_i(X_1, Y_1)\lambda_j(X_2, Y_2)$ *for* $(i, j) \in [0, d_1] \times [0, d_2]$, *where* $\lambda_0(X_1, Y_1) = \lambda_0(X_2, Y_2) = 1$. *Furthermore, denoting the principal functions* (X_1, Y_1) *by* f_i *and of* (X_2, Y_2) *by* $\widetilde{f}_j$, *then the principal functions of* $p_{(X_1, X_2),(Y_1, Y_2)}$ *are of the form* $(x_1, x_2) \mapsto f_i(x_1)\widetilde{f}_j(x_2)$. *In particular*

$$\lambda_1((X_1, X_2); (Y_1, Y_2)) = \max\{\lambda_1(X_1; Y_1), \lambda_1(X_2; Y_2)\}.$$

Proof. Let $[\mathbf{Q}_1]_{i,j} = \frac{p_{X_1, Y_1}(i,j)}{\sqrt{p_{X_1}(i)p_{Y_1}(j)}}$ and $[\mathbf{Q}_2]_{i,j} = \frac{p_{X_2, Y_2}(i,j)}{\sqrt{p_{X_2}(i)p_{Y_2}(j)}}$. Denoting by $\mathbf{Q}$ the decomposition in Definition 1 of $p_{(X_1, X_2),(Y_1, Y_2)}$ then, from the independence assumption, $\mathbf{Q} = \mathbf{Q}_1 \otimes \mathbf{Q}_2$, where $\otimes$ is the Kronecker product. The result follows directly from the fact that the singular values of the Kronecker product of two matrices are the Kronecker product of the singular values (and equivalently for the singular vectors) [66, Theorem 4.2.15]. □

2.3 k-correlation

In this section we introduce the k-correlation $\mathcal{J}_k(X;Y)$ between two random variables, which is equivalent to the sum of the k largest PICs. We prove that k-correlation is convex in $p_{Y|X}$ and satisfies the DPI.

Definition 3. We define the *k-correlation* between X and Y as

$$\mathcal{J}_k(X;Y) \triangleq \sum_{i=1}^{k} \lambda_i(X;Y). \tag{18}$$

For finite $\mathcal{X}$ and $\mathcal{Y}$, the k-correlation is given by

$$\mathcal{J}_k(X;Y) \triangleq \|\mathbf{Q}\mathbf{Q}^T\|_k - 1. \tag{19}$$

Note that

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

and for finite $\mathcal{X}$ and $\mathcal{Y}$, $d = \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1$,

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

We demonstrate next that k-correlation and, consequently, maximal correlation, is convex in $p_{Y|X}$ for a fixed p_X and satisfies a form of the DPI, i.e. if $X \to Y \to Z$, then $\mathcal{J}_k(X;Y) \le \mathcal{J}_k(X;Z)$. These results hold for both discrete and continuous random variables (under appropriate compactness assumptions),

Theorem 2. *For a fixed* p_X, $\mathcal{J}_k(X;Y)$ *is convex in* $p_{Y|X}$.

Proof. First note that $\|\mathbb{E}[f(X)|Y]\|_2^2$ is convex $p_{X,Y}$, since for any $U \to (X, Y)$

$$\mathbb{E}_Y\left[\left(\mathbb{E}_{X|Y}[f(X)|Y]\right)^2\right] = \mathbb{E}_Y\left[\left(\mathbb{E}_{U|Y}\left[\mathbb{E}_{X|Y,U}[f(X)|Y, U]\right]\right)^2\right]$$

$$\le \mathbb{E}_Y\left[\mathbb{E}_{U|Y}\left[\left(\mathbb{E}_{X|Y,U}[f(X)|Y, U]\right)^2\right]\right]$$

$$= \mathbb{E}_U\left[\mathbb{E}_{Y|U}\left[\left(\mathbb{E}_{X|Y,U}[f(X)|Y, U]\right)^2\right]\right],$$

where the inequality follows from Jensen's inequality. Consequently, for any $\{f_1, \ldots, f_k\} \subseteq \mathcal{L}_2(p_X)$, $\sum_{i=1}^{k} \|\mathbb{E}[f_i(X)|Y]\|_2^2$ is convex in $p_{X,Y}$ and thus, for a fixed p_X, convex in $p_{Y|X}$. From Theorem 1 and the Poincaré separation theorem [24, Corollary 4.3.16]

$$\sum_{i=1}^{k} \lambda_i(X;Y) = \max_{\substack{\{f_i\}_{i=1}^{k} \subseteq \mathcal{L}_2(p_X) \\ f_i \perp f_j, i \neq j \\ \mathbb{E}[f_i] = 0}} \sum_{i=1}^{k} \|\mathbb{E}[f_i(X)|Y]\|_2^2.$$

11

Since the pointwise supremum of convex functions is convex [67, Sec 3.2.3], it follows that for fixed p_X $\mathcal{J}_k(X;Y)$ is convex in $p_{Y|X}$. $\square$

The following lemma will be used to prove that the PICs satisfy the DPI.

Lemma 2 (DPI for MMSE). *For $X \to Y \to Z$ and any $f \in \mathcal{L}_2(p_X)$, $\mathbb{E}[f(X)] = 0$,*

$$\| \mathbb{E}[f(X)|Z] \|_2^2 \leq \lambda_1(Y;Z) \| \mathbb{E}[f(X)|Y] \|_2^2. \tag{20}$$

Consequently, $\mathsf{mmse}(f(X)|Y) \leq \mathsf{mmse}(f(X)|Z)$.

Proof. The proof is in Appendix A. $\square$

Lemma 2 leads to the following theorem.

Theorem 3 (DPI for the PICs). *Assume that $X \to Y \to Z$. Then $\lambda_k(X;Z) \leq \lambda_1(Y;Z)\lambda_k(X;Y)$ for all k.*

Proof. A direct consequence of Theorem 1 is that for any two random variables X, Y

$$\lambda_k(X;Y) = \min_{\{f_i\}_{i=1}^k \subseteq \mathcal{L}_2(p_X)} \max_{\substack{f \in \mathcal{L}_2(p_X) \\ \mathbb{E}[f(X)f_i(X)]=0}} \| \mathbb{E}[f(X)|Y] \|_2^2,$$

and equivalently for $\lambda_k(X;Z)$. The result then follows directly from (20). $\square$

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

Corollary 1. *For $X \to Y \to Z$ forming a Markov chain, $\mathcal{J}_k(X;Z) \leq \lambda_1(Y;Z)\mathcal{J}_k(X;Y)$.*

Remark 1. The data processing result in Theorem 3 and the previous corollary was proved by Kang and Ulukus in [41, Theorem 2] and applied to problems in distributed source and channel coding, even though they do not make the explicit connection with maximal correlation and PICs. A weaker form of Theorem 3 can be derived using a clustering result presented in [11, Sec. 7.5.4] and originally due to Deniau *et al.* [68]. We use a different proof technique from the one in [11, Sec. 7.5.4] and [41, 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 [33].

In the next three sections of the paper, we demonstrate the fundamental role of PICs in problems in information theory, estimation theory, and privacy.

3 Applications of the Principal Inertia Components to Information Theory

In this section, we present results that connect the PICs with other information-theoretic metrics. As seen in Section 2, the distribution of the vectors $p_{Y|X}$ in the simplex or, equivalently, the PICs of the joint distribution of X and Y, are inherently connected to how an observation of Y is statistically related to X. In this section, we explore this connection within an information theoretic framework. We show that, under certain assumptions, the PICs play an important part in estimating a one-bit function of X, namely $b(X)$ where $b : \mathcal{X} \to \{0, 1\}$, given an observation of Y: they can be understood as the singular values (or filter coefficients) in the linear transformation of $p_{b(X)|X}$ into $p_{b(X)|Y}$ determined by the channel transition matrix. Alternatively, the PICs can bear an interpretation as the transform of the distribution of the noise in certain additive-noise channels, in particular when X and Y are binary strings. We also show that maximizing the PICs is equivalent to maximizing the first-order term of the Taylor series expansion of certain convex dependence measures between $b(X)$ and Y. We conjecture that, for symmetric distributions of X and Y and a given upper bound on the value of the largest PIC, $I(b(X);Y)$ is maximized when all the principal inertia components have the same value as the largest principal inertia component. For uniformly distributed X and Y, this is equivalent to Y being the result of passing X through a q-ary symmetric channel. This conjecture, if proven, would imply the conjecture made by Kumar and Courtade in [17].

Finally, we study the Markov chain $B \to X \to Y \to \widehat{B}$, where B and $\widehat{B}$ are binary random variables, and the role of the principal inertia components in characterizing the relation between B and $\widehat{B}$. We show that this relation is linked to solving a non-linear maximization problem, which, in turn, can be solved when $\widehat{B}$ is an unbiased estimate of B (i.e. $\mathbb{E}[B] = \mathbb{E}[\widehat{B}]$), the joint distribution of X and Y is

symmetric and $\Pr\{B = \widehat{B} = 0\} \geq \mathbb{E}[B]^2$. We illustrate this result for the setting where X is a binary string and Y is the result of sending X through a memoryless binary symmetric channel. We note that this is a similar setting to the one considered by Anantharam *et al.* in [47].

The rest of the section is organized as follows. Section 3.1 introduces the notion of conforming distributions and ancillary results. Section 3.2 presents results concerning the role of the PICs in inferring one-bit functions of X from an observation of Y and in the transformation of p_X into p_Y in certain symmetric settings. We argue that, in such settings, the PICs can be viewed as singular values (filter coefficients) in a linear transformation. In particular, results for binary channels with additive noise are derived using techniques inspired by Fourier analysis of Boolean functions. Furthermore, Section 3.2 also introduces a conjecture that encompasses the one made by Kumar and Courtade in [17]. Finally, Section 3.6 provides further evidence for this conjecture by investigating the Markov chain $B \to X \to Y \to \widehat{B}$ where B and $\widehat{B}$ are binary random variables. Throughout this section we assume X and Y are discrete random variables defined over a finite support set.

3.1   Conforming distributions

In this section we shall focus on probability distributions that meet the following definition.

Definition 4. A joint distribution $p_{X,Y}$ is said to be *conforming* if the corresponding matrix $\mathbf{P}$ satisfies $\mathbf{P} = \mathbf{P}^T$ and $\mathbf{P}$ is positive-semidefinite.

Conforming distributions are particularly interesting since they are closely related to symmetric channels[1]. In addition, if a joint distribution is conforming, then its eigenvalues are equal to (the square root of) its PICs when its marginal distributions are identical. We shall illustrate this relation in the following two lemmas and in Section 3.2.

Remark 2. If X and Y have a conforming joint distribution, then they have the same marginal distribution. Consequently, $\mathbf{D} \triangleq \mathbf{D}_X = \mathbf{D}_Y$, and $\mathbf{P} = \mathbf{D}^{1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^T\mathbf{D}^{1/2}$ (cf. Definition 2 for notation).

Lemma 3. *If $\mathbf{P}$ is conforming, then the corresponding conditional distribution matrix $\mathbf{P}_{Y|X}$ is positive semi-definite. Furthermore, for any symmetric channel $\mathbf{P}_{Y|X} = \mathbf{P}_{Y|X}^T$, there is an input distribution p_X (namely, the uniform distribution) such that the PICs of $\mathbf{P} = \mathbf{D}_X\mathbf{P}_{Y|X}$ correspond to the square of the eigenvalues of $\mathbf{P}_{Y|X}$. In this case, if $\mathbf{P}_{Y|X}$ is also positive-semidefinite, then the resulting $\mathbf{P}$ is conforming.*

Proof. Let $\mathbf{P}$ be conforming and $\mathcal{X} = \mathcal{Y} = [m]$. Then $\mathbf{P}_{Y|X} = \mathbf{D}^{-1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^T\mathbf{D}^{1/2} = \left(\mathbf{D}^{-1/2}\mathbf{U}\right)\mathbf{\Sigma}\left(\mathbf{D}^{-1/2}\mathbf{U}\right)^{-1}$. It follows that diag $(\mathbf{\Sigma})$ are the eigenvalues of $\mathbf{P}_{Y|X}$, and, consequently, $\mathbf{P}_{Y|X}$ is positive semi-definite.

Now let $\mathbf{P}_{Y|X} = \mathbf{P}_{Y|X}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$. The entries of $\mathbf{\Lambda}$ here are the eigenvalues of $\mathbf{P}_{Y|X}$ and not necessarily positive. Since $\mathbf{P}_{Y|X}$ is symmetric, it is also doubly stochastic, and for X uniformly distributed Y is also uniformly distributed. Thus, the resulting joint distribution matrix $\mathbf{P}$ is symmetric, and $\mathbf{P} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T/m$. It follows directly that the principal inertia components of $\mathbf{P}$ are the diagonal entries of $\mathbf{\Lambda}^2$, and if $\mathbf{P}_{Y|X}$ is positive-semidefinite then $\mathbf{P}$ is conforming. $\qquad\square$

The q-ary symmetric channel, defined below, is of particular interest to some of the results derived in the following subsections.

Definition 5. The q-ary symmetric channel with crossover probability $\epsilon \leq 1 - q^{-1}$, also denoted as (ϵ, q)-SC, is defined as the channel with input X and output Y where $\mathcal{X} = \mathcal{Y} = [q]$ and

$$p_{Y|X}(y|x) = \begin{cases} 1 - \epsilon & \text{if } x = y \\ \dfrac{\epsilon}{q-1} & \text{if } x \neq y. \end{cases}$$

In the rest of this section, we assume that X and Y have a conforming joint distribution matrix with $\mathcal{X} = \mathcal{Y} = [q]$ and PICs $\lambda_k(X;Y) = \sigma_k^2$ for $k \in [d-1]$. The following lemma shows that a conforming $\mathbf{P}$ with uniform marginals can be transformed into the joint distribution of a q-ary symmetric channel with input distribution p_X by setting $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_{q-1}^2$, i.e. making all principal inertia components equal to the largest one.

---

[1]We say that a channel is symmetric if $\mathbf{P}_{Y|X} = \mathbf{P}_{Y|X}^T$.

Lemma 4. *Let* $\mathbf{P}$ *be a conforming joint distribution matrix of* X *and* Y, *with* $\mathcal{X} = \mathcal{Y} = [q]$, $\mathbf{P} = \mathbf{D}^{1/2}\mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T\mathbf{D}^{1/2}$, *where* $\mathbf{D} = \mathbf{D}_X$ *and* $\boldsymbol{\Sigma} = \mathrm{diag}\,(1, \sigma_1, \ldots, \sigma_d)$. *For* $\widetilde{\boldsymbol{\Sigma}} = \mathrm{diag}\,(1, \sigma_1, \ldots, \sigma_1)$, *let* X *and* $\widetilde{Y}$ *have joint distribution* $\widetilde{\mathbf{P}} = \mathbf{D}^{1/2}\mathbf{U}\widetilde{\boldsymbol{\Sigma}}\mathbf{U}^T\mathbf{D}^{1/2}$. *Then,* $\widetilde{Y}$ *is output of a channel with input* X *and probability transition matrix*

$$\mathbf{P}_{\widetilde{Y}|X} = \sigma_1\mathbf{I} + (1 - \sigma_1)\mathbf{1}\mathbf{p}_X^T. \tag{21}$$

In particular, if X *is uniform,* $\widetilde{Y}$ *is the output of an* (ϵ, q)-*SC with input* X, *where*

$$\epsilon = \frac{(q - 1)(1 - \rho_m(X; Y))}{q}. \tag{22}$$

Proof. The first column of $\mathbf{U}$ is $\mathbf{p}_X^{1/2}$. Therefore

$$\begin{aligned} \widetilde{\mathbf{P}} &= \mathbf{D}^{1/2}\mathbf{U}\widetilde{\boldsymbol{\Sigma}}\mathbf{U}^T\mathbf{D}^{1/2} \\ &= \sigma_1\mathbf{D} + (1 - \sigma_1)\mathbf{p}_X\mathbf{p}_X^T. \end{aligned} \tag{23}$$

By left multiplying $\widetilde{\mathbf{P}}$ by $\mathbf{D}^{-1}$, we obtain the channel transition matrix given in (21). $\qquad\square$

Remark 3. For X, Y and $\widetilde{Y}$ given in the previous lemma, a natural question that arises is whether Y is a degraded version of $\widetilde{Y}$, i.e. $X \to \widetilde{Y} \to Y$. Unfortunately, this is not true in general, since the matrix $\mathbf{U}\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\mathbf{U}^T$ does not necessarily contain only positive entries, although it is doubly-stochastic. However, since the PICs of X and $\widetilde{Y}$ upper bound the PICs of X and Y, it is natural to expect that, at least in some sense, $\widetilde{Y}$ is more informative about X than Y. This intuition is indeed correct for certain estimation problems where a one-bit function of X is to be inferred from a single observation Y or $\widetilde{Y}$, and will be investigated in the next subsection. In addition, using the characterization of the PICs in Theorem 1, it follows that *any* function of X can be inferred with smaller MMSE from $\widetilde{Y}$ than from Y. Consequently, even if, for example $I(X; \widetilde{Y}) \leq I(X; Y)$, any function of X can be estimated with smaller MMSE for $\widetilde{Y}$ than from Y.

3.2 One-bit Functions and Channel Transformations

Let $B \to X \to Y$, where B is a binary random variable. When X and Y have a conforming probability distribution, the PICs of X and Y have a particularly interesting interpretation: they can be understood as the filter coefficients in a linear transformation from $p_{B|X}$ into $p_{B|Y}$, as we explain next. Consider the joint distribution of B and Y, denoted here by $\mathbf{B}$, given by

$$\mathbf{B} \triangleq [\mathbf{x} \quad \mathbf{1} - \mathbf{x}]^T\mathbf{P} = [\mathbf{x} \quad \mathbf{1} - \mathbf{x}]^T\mathbf{P}_{X|Y}\mathbf{D}_Y = [\mathbf{y} \quad \mathbf{1} - \mathbf{y}]^T\mathbf{D}_Y, \tag{24}$$

where $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$ are column-vectors with entries $x_i = p_{B|X}(0|i)$ and $y_j = p_{B|Y}(0|j)$. In particular, if B is a deterministic function of X, $\mathbf{x} \in \{0, 1\}^m$.

If $\mathbf{P}$ is conforming and $\mathcal{X} = \mathcal{Y} = [m]$, then $\mathbf{P} = \mathbf{D}^{1/2}\mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T\mathbf{D}^{1/2}$, where $\mathbf{D} = \mathbf{D}_X = \mathbf{D}_Y$. Assuming $\mathbf{D}$ fixed, the joint distribution $\mathbf{B}$ is entirely specified by the linear transformation of $\mathbf{x}$ into $\mathbf{y}$. Denoting $\mathbf{T} \triangleq \mathbf{U}^T\mathbf{D}^{1/2}$, this transformation is done in three steps:

1. (Linear transform) $\widehat{\mathbf{x}} \triangleq \mathbf{T}\mathbf{x}$,
2. (Filter) $\widehat{\mathbf{y}} \triangleq \boldsymbol{\Sigma}\widehat{\mathbf{x}}$, where the diagonal of $\boldsymbol{\Sigma}^2$ are the PICs of X and Y,
3. (Inverse transform) $\mathbf{y} = \mathbf{T}^{-1}\widehat{\mathbf{y}}$.

Note that $\widehat{x}_1 = \widehat{y}_1 = 1 - \mathbb{E}[B]$ and $\widehat{\mathbf{y}} = \mathbf{T}\mathbf{y}$. Consequently, the PICs of X and Y correspond to the singular values (or filter coefficients) of the linear transformation of $p_{B|X}(0|\cdot)$ into $p_{B|Y}(0|\cdot)$.

A similar interpretation can be made for symmetric channels, where $\mathbf{P}_{Y|X} = \mathbf{P}_{Y|X}^T = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ and $\mathbf{P}_{Y|X}$ acts as the matrix of the linear transformation of $\mathbf{p}_X$ into $\mathbf{p}_Y$. Note that $\mathbf{p}_Y = \mathbf{P}_{Y|X}\mathbf{p}_X$, and, consequently, $\mathbf{p}_X$ is transformed into $\mathbf{p}_Y$ in the same three steps as before:

1. (Linear transform) $\widehat{\mathbf{p}_X} = \mathbf{U}^T\mathbf{p}_X$,
2. (Filter) $\widehat{\mathbf{p}_Y} \triangleq \boldsymbol{\Lambda}\widehat{\mathbf{p}_X}$, where the diagonal of $\boldsymbol{\Lambda}^2$ is the PICs of X and Y in the particular case when X is uniformly distributed (Lemma 3),
3. (Inverse transform) $\mathbf{p}_Y = \mathbf{U}\widehat{\mathbf{p}_Y}$.

From this perspective, the vector $\mathbf{z} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{1}m^{-1/2}$ can be understood as a proxy for the noise effect of the channel. Note that $\sum_i z_i = 1$. However, the entries of $\mathbf{z}$ are not necessarily positive, and $\mathbf{z}$ might not be a probability distribution.

We now illustrate these ideas by investigating binary channels with additive noise in the next section, where $\mathbf{T}$ will correspond to the well-known Walsh-Hadamard transform matrix.

3.3 Example: Binary Additive Noise Channels

In this example, let $\mathcal{X}^n, \mathcal{Y}^n \subseteq \{-1, 1\}^n$ be the support sets of X^n and Y^n, respectively. We define two sets of channels that maps X^n to Y^n. In each set definition, we assume the conditions for $p_{Y^n|X^n}$ to be a valid probability distribution (i.e. non-negativity and unit sum).

Definition 6. The set of *parity-changing channels* of block-length n, denoted by $\mathcal{A}_n$, is defined as:

$$\mathcal{A}_n \triangleq \left\{ p_{Y^n|X^n} \mid \forall \mathcal{S} \subseteq [n], \ \exists c_\mathcal{S} \in [-1, 1] \text{ s.t. } \mathbb{E}\left[\chi_\mathcal{S}(Y^n) | X^n \right] = c_\mathcal{S} \chi_\mathcal{S}(X^n) \right\}, \tag{25}$$

where $\chi_\mathcal{S}(\cdot)$ is defined in (9). The set of all *binary additive noise channels* is given by

$$\mathcal{B}_n \triangleq \left\{ p_{Y^n|X^n} \mid \exists Z^n \text{ s.t. } Y^n = X^n \oplus Z^n, \ \text{supp}(Z^n) \subseteq \{-1, 1\}^n, Z^n \perp\!\!\!\perp X^n \right\}. \tag{26}$$

The definition of parity-changing channels is inspired by results from the literature on Fourier analysis of Boolean functions. For an overview of the topic we refer the reader to the survey [69]. The set of binary additive noise channels, in turn, is widely used in the information theory literature. The following lemma shows that both characterizations are equivalent.

Lemma 5. *For $\mathcal{A}_n$ and $\mathcal{B}_n$ given in (25) and (26), respectively, $\mathcal{A}_n = \mathcal{B}_n$.*

Proof. The proof is in Appendix B.

$\square$

The previous theorem suggests that there is a correspondence between the coefficients $c_\mathcal{S}$ in (25) and the distribution of the additive noise Z^n in the definition of $\mathcal{B}_n$. The next result shows that this is indeed the case and, when X^n is uniformly distributed, the coefficients $c_\mathcal{S}^2$ correspond to the PICs of X^n and Y^n.

Theorem 4. *Let $p_{Y^n|X^n} \in \mathcal{B}_n$, and $X^n \sim p_{X^n}$. Then $\mathbf{P}_{X^n, Y^n} = \mathbf{D}_{X^n} \mathbf{H}_{2^n} \mathbf{\Lambda} \mathbf{H}_{2^n}$, where $\mathbf{H}_l$ is the $l \times l$ normalized Hadamard matrix[2] (hence $\mathbf{H}_l^2 = \mathbf{I}$). Furthermore, for $Z^n \sim p_{Z^n}$, $\text{diag}(\mathbf{\Lambda}) = 2^{n/2} \mathbf{H}_{2^n} \mathbf{p}_{Z^n}$, and the diagonal entries of $\mathbf{\Lambda}$ are equal to $c_\mathcal{S}$ in (25). Finally, if X is uniformly distributed, then $c_\mathcal{S}^2$ are the principal inertia components of X^n and Y^n.*

Proof. Let $p_{Y^n|X^n} \in \mathcal{A}_n$ be given. From Lemma 5 and the definition of $\mathcal{A}_n$, it follows that $\chi_\mathcal{S}(Y^n)$ is a right eigenvector of $p_{Y^n|X^n}$ with corresponding eigenvalue $c_\mathcal{S}$. Since $\chi_\mathcal{S}(Y^n) 2^{-n/2}$ corresponds to a row of $\mathbf{H}_{2^n}$ for each $\mathcal{S}$ (due to the Kronecker product construction of the Hadamard matrix) and $\mathbf{H}_{2^n}^2 = \mathbf{I}$, then $\mathbf{P}_{X^n, Y^n} = \mathbf{D}_{X^n} \mathbf{H}_{2^n} \mathbf{\Lambda} \mathbf{H}_{2^n}$. Finally, note that $\mathbf{p}_Z^T = 2^{-n/2} \mathbf{1}^T \mathbf{\Lambda} \mathbf{H}_{2^n}$. From Lemma 3, it follows that $c_\mathcal{S}^2$ are the PICs of X^n and Y^n if X^n is uniformly distributed. $\square$

Remark 4. Theorem 4 suggests that one possible method for estimating the distribution of the additive binary noise Z^n is to estimate its effect on the parity bits of X^n and Y^n. In this case, we are estimating the coefficients $a_\mathcal{S}$ of the Walsh-Hadamard transform of p_{Z^n}. This approach was studied by Raginsky *et al.* in [70] and in other learning literature (see [71] and the references therein).

Theorem 4 illustrates the filtering role of the principal inertia components (discussed in Section 3.2) in binary additive noise channels. If X^n is uniform, then the vector of conditional probabilities $\mathbf{p}_X$ is transformed into the vector of *a posteriori* probabilities $\mathbf{p}_Y$ by: (i) taking the Hadamard transform of $\mathbf{p}_X$, (ii) filtering the transformed vector according to the coefficients $c_\mathcal{S}$ (these coefficients have a one-to-one mapping to the entries of the vector resulting from the Hadamard transform of $\mathbf{p}_Z$), and (iii) taking the inverse Hadamard transform to recover $\mathbf{p}_Y$.

---

[2] We define the normalized Hadamard matrix $\mathbf{H}_{2^k}$ as $\mathbf{H}_1 \triangleq [1]$,

$$\mathbf{H}_2 \triangleq \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

and $\mathbf{H}_{2^k} \triangleq \mathbf{H}_2 \otimes \mathbf{H}_{2^{k-1}}$.

3.4 Quantifying the Information of a Boolean Function of the Input of a Noisy Channel

We now investigate the connection between the PICs and f-information (cf. Eq. (11)) in the context of one-bit functions of X. Recall from the discussion in the beginning of this section and, in particular, equation (24), that for a binary B and $B \to X \to Y$, the distribution of B and Y is entirely specified by the transformation of $\mathbf{x}$ into $\mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are vectors with entries equal to $p_{B|X}(0|\cdot)$ and $p_{B|Y}(0|\cdot)$, respectively.

For $\mathbb{E}[B] = 1 - a$, the f-information between B and Y is given by (cf. (11))

$$I_f(B;Y) = \mathbb{E}\left[af\left(\frac{p_B(0|Y)}{a}\right) + (1-a)f\left(\frac{1-p_B(0|Y)}{1-a}\right)\right].$$

For $0 \leq r,s \leq 1$, and since f is smooth with $f(1) = 0$, we can expand $f\left(\frac{r}{s}\right)$ around 1 as

$$f\left(\frac{r}{s}\right) = \sum_{k=1}^{\infty} \frac{f^{(k)}(1)}{k!}\left(\frac{r-s}{r}\right)^k.$$

Denoting

$$c_k(\alpha) \triangleq \frac{1}{a^{k-1}} + \frac{(-1)^k}{(1-a)^{k-1}},$$

the f-information can then be expressed as

$$I_f(B;Y) = \sum_{k=2}^{\infty} \frac{f^{(k)}(1)c_k(a)}{k!}\mathbb{E}\left[(p_B(0|Y) - a)^k\right]. \tag{27}$$

Similarly to [25, Chapter 4], for a fixed $\mathbb{E}[B] = 1 - a$, maximizing the PICs of X and Y will always maximize the first term in the expansion (27). To see why this is the case, observe that

$$\begin{aligned}
\mathbb{E}\left[(p_{B|Y}(0|Y) - a)^2\right] &= (\mathbf{y} - a)^T \mathbf{D}_Y (\mathbf{y} - a) \\
&= \mathbf{y}^T \mathbf{D}_Y \mathbf{y} - a^2 \\
&= \mathbf{x}^T \mathbf{D}_X^{1/2} \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^T \mathbf{D}_X^{1/2}\mathbf{x} - a^2.
\end{aligned} \tag{28}$$

For a fixed a and any $\mathbf{x}$ such that $\mathbf{x}^T\mathbf{1} = a$, (28) is non-decreasing in the diagonal entries of $\mathbf{\Sigma}^2$ which, in turn, are exactly the PICs of X and Y. Equivalently, (28) is non-decreasing in the χ^2-divergence between $p_{X,Y}$ and $p_X p_Y$.

However, we do note that increasing the PICs does not increase the f-information between B and Y in general. Indeed, for a fixed $\mathbf{U}$, $\mathbf{V}$ and marginal distributions of X and Y, increasing the PICs might not even lead to a valid probability distribution matrix $\mathbf{P}$.

Nevertheless, if $\mathbf{P}$ is conforming and X and Y are uniformly distributed over $[q]$, as shown in Lemma 4, by increasing the PICs we can define a new random variable $\widetilde{Y}$ that results from sending X through a (ϵ, q)-SC, where ϵ is given in (22). In this case, the f-information between B and Y has a simple expression when B is a function of X.

Lemma 6. *Let* $B \to X \to \widetilde{Y}$, *where* $B = b(X)$ *for some* $b : [q] \to \{0,1\}$, $\mathbb{E}[B] = 1 - a$ *where* aq *is an integer,* X *is uniformly distributed in* $[q]$ *and* $\widetilde{Y}$ *is the result of passing* X *through a* (ϵ, q)-SC *with* $\epsilon \leq (q-1)/q$. *Then*

$$I_f(B;\widetilde{Y}) = a^2 f(1 + \sigma_1 c) + 2a(1-a)f(1-\sigma_1) + (1-a)^2 f(1 + \sigma_1 c^{-1}) \tag{29}$$

where $\sigma_1 = \rho_m(X;\widetilde{Y}) = 1 - \epsilon q(q-1)^{-1}$ *and* $c \triangleq (1-a)a^{-1}$. *In particular, for* $f(x) = x\log x$, *then* $I_f(X;\widetilde{Y}) = I(X;\widetilde{Y})$, *and for* $\sigma_1 = 1 - 2\delta$

$$I(B;\widetilde{Y}) = h_b(a) - \alpha h_b(2\delta(1-a)) - (1-a)h_b(2\delta a) \tag{30}$$

$$\leq 1 - h_b(\delta). \tag{31}$$

where $h_b(\cdot)$ *is the binary entropy function, defined in* (2).

Proof. Since B is a deterministic function of X and aq is an integer, $\mathbf{x}$ is a vector with aq entries equal to 1 and $(1-a)q$ entries equal to 0. It follows from (23) that

$$
\begin{aligned}
I_f(B; \widetilde{Y}) =& \frac{1}{q} \sum_{i=1}^{q} af\left(\frac{(1-\sigma_1)a + x_i\sigma_1}{a}\right) + (1-a)f\left(\frac{1-(1-\sigma_1)a - x_i\sigma_i}{1-a}\right) \\
=& a^2 f\left(1 + \sigma_1 \frac{1-a}{a}\right) + 2a(1-a)f(1-\sigma_1) + (1-a)^2 f\left(1 + \sigma_1 \frac{a}{1-a}\right).
\end{aligned}
$$

Letting $f(x) = x \log x$, (30) follows immediately. Since (30) is concave in a and symmetric around $a = 1/2$, it is maximized at $a = 1/2$, resulting in (31). $\qquad\square$

3.5 On the "Most Informative Bit"

We now return to channels with additive binary noise, analyzed in Section 3.3. Let X^n be a uniformly distributed binary string of length n ($\mathcal{X} = \{-1, 1\}$) and Y^n be the result of passing X^n through a memoryless binary symmetric channel with crossover probability $\delta \leq 1/2$. Kumar and Courtade conjectured [17] that for all binary B and $B \to X^n \to Y^n$ we have

$$
I(B; Y^n) \leq 1 - h_b(\delta). \quad \text{(conjecture)} \tag{32}
$$

It is sufficient to consider B a function of X^n, denoted by $B = b(X^n)$, $b : \{-1, 1\}^n \to \{0, 1\}$, and we make this assumption henceforth.

From the discussion in Section 3.3, for the memoryless binary symmetric channel $Y^n = X^n \oplus Z^n$, where Z^n is an i.i.d. string with $\Pr\{Z_i = 1\} = 1 - \delta$, and any $\mathcal{S} \in [n]$,

$$
\begin{aligned}
\mathbb{E}\left[\chi_{\mathcal{S}}(Y^n)|X^n\right] &= \chi_{\mathcal{S}}(X^n)\left(\Pr\{\chi_{\mathcal{S}}(Z^n) = 1\} - \Pr\{\chi_{\mathcal{S}}(Z^n) = -1\}\right) \\
&= \chi_{\mathcal{S}}(X^n)\left(2\Pr\{\chi_{\mathcal{S}}(Z^n) = 1\} - 1\right) \\
&= \chi_{\mathcal{S}}(X^n)(1 - 2\delta)^{|\mathcal{S}|}.
\end{aligned}
$$

It follows directly that $c_{\mathcal{S}} = (1-2\delta)^{|\mathcal{S}|}$ for all $\mathcal{S} \subseteq [n]$. Consequently, from Theorem 4, the principal inertia components of X^n and Y^n are of the form $(1-2\delta)^{2|\mathcal{S}|}$ for some $\mathcal{S} \subseteq [n]$. Observe that the principal inertia components act, broadly speaking, as a low pass filter on the vector of conditional probabilities $\mathbf{x}$ given in (24), since it attenuates the high order interaction terms in the Walsh-Hadamard transform of $\mathbf{x}$.

Can the noise distribution be modified so that the principal inertia components act as an all-pass filter? More specifically, what happens when $\widetilde{Y}^n = X^n \oplus W^n$, where W^n is such that the principal inertia components between X^n and $\widetilde{Y}^n$ satisfy $\sigma_i = 1 - 2\delta$? Then, from Lemma 4, $\widetilde{Y}^n$ is the result of sending X^n through a $(\epsilon, 2^n)$-SC with $\epsilon = 2\delta(1 - 2^{-n})$. Therefore, from (31),

$$
I(B; \widetilde{Y}^n) \leq 1 - h_b(\delta).
$$

For any function $b : \{-1, 1\}^n \to \{0, 1\}$ such that $B = b(X^n)$, from standard results in Fourier analysis of Boolean functions [69, Prop. 1.1], $b(X^n)$ can be expanded as

$$
b(X^n) = \sum_{\mathcal{S} \subseteq [n]} \beta_{\mathcal{S}} \chi_{\mathcal{S}}(X^n).
$$

The value of B is uniquely determined by the action of b on $\chi_{\mathcal{S}}(X^n)$. Consequently, for a fixed function b, one could expect that $\widetilde{Y}^n$ should be more informative about B than Y^n, since the parity bits $\chi_{\mathcal{S}}(X^n)$ are more reliably estimated from $\widetilde{Y}^n$ than from Y^n. Indeed, the memoryless binary symmetric channel attenuates $\chi_{\mathcal{S}}(X^n)$ exponentially in $|\mathcal{S}|$, acting (as argued previously) as a low-pass filter. In addition, if one could prove that for any fixed b the inequality $I(B; Y^n) \leq I(B; \widetilde{Y}^n)$ holds, then (32) would be proven true. This motivates the following conjecture.

Conjecture 1. *For all $b : \{-1, 1\}^n \to \{0, 1\}$ and $B = b(X^n)$*

$$
I(B; Y^n) \leq I(B; \widetilde{Y}^n).
$$

We note that Conjecture 1 is false if B is not a deterministic function of X^n. In the next section, we provide further evidence for this conjecture by investigating information metrics between B and an estimate $\widehat{B}$ derived from Y^n.

3.6 One-bit Estimators

Let $B \to X \to Y \to \widehat{B}$, where B and $\widehat{B}$ are binary random variables with $\mathbb{E}[B] = 1 - a$ and $\mathbb{E}[\widehat{B}] = 1 - b$. Again, we let $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$ be the column vectors with entries $x_i = p_{B|X}(0|i)$ and $y_j = p_{\widehat{B}|Y}(0|j)$. The joint distribution matrix of B and $\widehat{B}$ is given by

$$\mathbf{P}_{B,\widehat{B}} = \begin{pmatrix} z & a - z \\ b - z & 1 - a - b + z \end{pmatrix}, \tag{33}$$

where $z = \mathbf{x}^T \mathbf{P} \mathbf{y} = \Pr\{B = \widehat{B} = 0\}$. For fixed values of a and b, the joint distribution of B and $\widehat{B}$ only depends on z.

Let $f : \mathcal{P}_{2\times 2} \to \mathbb{R}$, and, with a slight abuse of notation, we also denote f as a function of the entries of the 2×2 matrix as $f(a, b, z)$. If f is convex in z for a fixed a and b, then f is maximized at one of the extreme values of z. Examples of such functions f include mutual information and expected error probability. Therefore, characterizing the maximum and minimum values of z is equivalent to characterizing the maximum value of f over all possible mappings $X \to B$ and $Y \to \widehat{B}$. This leads to the following definition.

Definition 7. For a fixed $\mathbf{P}$ and given $\mathbb{E}[B] = 1 - a$ and $\mathbb{E}[\widehat{B}] = 1 - b$, the minimum and maximum values of z over all possible mappings $X \to B$ and $Y \to \widehat{B}$ are defined as

$$z_l^*(a, b, \mathbf{P}) \triangleq \min_{\substack{\mathbf{x} \in \mathcal{C}^m(a, \mathbf{P}^T) \\ \mathbf{y} \in \mathcal{C}^n(b, \mathbf{P})}} \mathbf{x}^T \mathbf{P} \mathbf{y} \quad \text{and} \quad z_u^*(a, b, \mathbf{P}) \triangleq \max_{\substack{\mathbf{x} \in \mathcal{C}^m(a, \mathbf{P}^T) \\ \mathbf{y} \in \mathcal{C}^n(b, \mathbf{P})}} \mathbf{x}^T \mathbf{P} \mathbf{y},$$

respectively, and $\mathcal{C}^n(a, \mathbf{P})$ is defined in (8).

The next lemma provides a simple upper-bound for $z_u^*(a, b, \mathbf{P})$ in terms of the largest principal inertia components or, equivalently, the maximal correlation between X and Y.

Lemma 7. $z_u^*(a, b, \mathbf{P}) \leq ab + \rho_m(X; Y)\sqrt{a(1-a)b(1-b)}$.

Proof. The proof is in Appendix B. $\qquad\square$

Remark 5. An analogous result was derived by Witsenhausen [32, Thm. 2] for bounding the probability of agreement of a common bit derived from two correlated sources.

We will focus in the rest of this section on functions and corresponding estimators that are (i) unbiased ($a = b$) and (ii) satisfy $z = \Pr\{\hat{B} = B = 0\} \geq a^2$. The set of all such mappings is given by

$$\mathcal{H}(a, \mathbf{P}) \triangleq \left\{ (\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in \mathcal{C}^m(a, \mathbf{P}^T), \mathbf{y} \in \mathcal{C}^n(a, \mathbf{P}), \mathbf{x}^T \mathbf{P} \mathbf{y} \geq a^2 \right\}.$$

The next results provide upper and lower bounds on z for the mappings in $\mathcal{H}(a, \mathbf{P})$.

Lemma 8. *Let $0 \leq a \leq 1/2$ and $\mathbf{P}$ be fixed. For any $(\mathbf{x}, \mathbf{y}) \in \mathcal{H}(a, \mathbf{P})$*

$$a^2 \leq z \leq a^2 + \rho_m(X; Y)a(1 - a), \tag{34}$$

where $z = \mathbf{x}^T \mathbf{P} \mathbf{y}$.

Proof. The lower bound for z follows directly from the definition of $\mathcal{H}(a, \mathbf{P})$, and the upper bound follows from Lemma 7. $\qquad\square$

The previous lemma allows us to provide an upper bound over the mappings in $\mathcal{H}(a, \mathbf{P})$ for the f-information between B and $\widehat{B}$ when I_f is non-negative.

Theorem 5. *For any non-negative I_f and fixed a and $\mathbf{P}$,*

$$\sup_{(\mathbf{x}, \mathbf{y}) \in \mathcal{H}(a, \mathbf{P})} I_f(B; \hat{B}) \leq a^2 f(1 + \sigma_1 c) + 2a(1 - a)f(1 - \sigma_1) + (1 - a)^2 f(1 + \sigma_1 c^{-1}) \tag{35}$$

where here $\sigma_1 = \rho_m(X; \widetilde{Y})$ and $c \triangleq (1 - a)a^{-1}$. In particular, for $a = 1/2$,

$$\sup_{(\mathbf{x}, \mathbf{y}) \in \mathcal{H}(1/2, \mathbf{P})} I_f(B; \hat{B}) \leq \frac{1}{2}(f(1 - \sigma_1) + f(1 + \sigma_1)). \tag{36}$$

18

Proof. Using the matrix form of the joint distribution between B and $\widehat{B}$ given in (33), for $\mathbb{E}[B] = \mathbb{E}\left[\widehat{B}\right] = 1 - a$, the f information is given by

$$I_f(B; \hat{B}) = a^2 f\left(\frac{z}{a^2}\right) + 2a(1-a)f\left(\frac{a-z}{a(1-a)}\right) + (1-a)^2 f\left(\frac{1-2a+z}{(1-a)^2}\right). \tag{37}$$

Consequently, (37) is convex in z. For $(\mathbf{x}, \mathbf{y}) \in \mathcal{H}(a, \mathbf{P})$, it follows from Lemma 8 that z is restricted to the interval in (34). Since $I_f(B; \hat{B})$ is non-negative by assumption, $I_f(B; \hat{B}) = 0$ for $z = a^2$ and (37) is convex in z, then $I_f(B; \hat{B})$ is non-decreasing in z for z in (34). Substituting $z = a^2 + \rho_m(X; Y)a(1-a)$ in (37), inequality (35) follows. $\qquad\square$

Remark 6. Note that the right-hand side of (35) matches the right-hand side of (29), and provides further evidence for Conjecture 1 by demonstrating that the conjecture holds for the specific case when $B \to X \to Y \to \hat{B}$ and $\mathbb{E}[B] = \mathbb{E}[\hat{B}]$. Moreover, this result indicates that, for conforming probability distributions, the information between a binary function and its corresponding unbiased estimate is maximized when all the PICs have the same value.

Following the same approach from Lemma 6, we find the next bound for the mutual information between B and $\widehat{B}$.

Corollary 2. *For $0 \le a \le 1$ and $\rho_m(X; Y) = 1 - 2\delta$*

$$\sup_{(p_{B|X}, p_{\widehat{B}|Y}) \in \mathcal{H}(a, \mathbf{P})} I(B; \hat{B}) \le 1 - h_b(\delta).$$

We provide next a few application examples for the results derived in this section.

Example 1 (Memoryless Binary Symmetric Channels with Uniform Inputs). We turn our attention back to the setting considered in Section 3.3. Let Y^n be the result of passing X^n through a memoryless binary symmetric channel with crossover probability δ, X^n uniformly distributed, and $B \to X^n \to Y^n \to \widehat{B}$. Then $\rho_m(X^n; Y^n) = 1 - 2\delta$ and, from (40), when $\mathbb{E}[B] = 1/2$,

$$\Pr\{B \ne \widehat{B}\} \ge \delta.$$

Consequently, inferring any unbiased one-bit function of the input of a binary symmetric channel is at least as hard (in terms of error probability) as inferring a single output from a single input.

Using the result from Corollary 2, it follows that when $\mathbb{E}[B] = \mathbb{E}\left[\widehat{B}\right] = a$ and $\Pr\{B = \widehat{B} = 0\} \ge a^2$, then

$$I(B; \widehat{B}) \le 1 - h_b(\delta). \tag{38}$$

Remark 7. Anantharam *et al.* presented in [47] a computer aided proof that the upper bound (38) holds for any $B \to X^n \to Y^n \to \widehat{B}$. Nevertheless, we highlight that the methods introduced here allowed an analytical derivation of (38) for unbiased estimators.

Example 2 (Lower Bounding the Estimation Error Probability). For z given in (33), the average estimation error probability is given by $\Pr\{B \ne \widehat{B}\} = a + b - 2z$, which is a convex (linear) function of z. If a and b are fixed, then the error probability is minimized when z is maximized. Therefore

$$\Pr\{B \ne \widehat{B}\} \ge a + b - 2z_u^*(a, b).$$

Using the bound from Lemma 7, it follows that

$$\Pr\{B \ne \widehat{B}\} \ge a + b - 2ab - 2\rho_m(X; Y)\sqrt{a(1-a)b(1-b)}. \tag{39}$$

The bound (39) is exactly the bound derived by Witsenhausen in [32, Thm 2.]. Furthermore, minimizing the right-hand side of (39) over $0 \le b \le 1/2$, we arrive at

$$\Pr\{B \ne \widehat{B}\} \ge \frac{1}{2}\left(1 - \sqrt{1 - 4a(1-a)(1-\rho_m(X; Y)^2)}\right). \tag{40}$$

This result suggests that the PICs are particularly useful for deriving bounds on error probability. We explore this fact in the next section, and show that (40) is a particular form of a more general bound derived in Theorem 6.

19

4   Application to Estimation: Bounds on Error Probability

In this section we derive lower bounds on error-probability based on the PICs (cf. section 2). Before presenting these bounds, we discuss the general approach used for deriving lower bounds, which can be extended to other measures of dependence. This approach is particularly useful for proving information-theoretic security and privacy guarantees.

Recall the central estimation-theoretic problem: Given an observation of a random variable Y, what can we learn about a correlated, hidden variable X? Such questions are relevant for different application areas. For example, in a symmetric-key encryption setup, X can be the plaintext message, and Y the ciphertext and any additional side information available to an adversary. If there is an encryption mechanism in place that guarantees that the mutual information between an individual symbol of the plaintext X and a ciphertext Y is at most 0.01 bits [72], how well can an adversary guess individual symbols of X? How does this result depend on the distribution of the plaintext source? Are there other information measures besides mutual information for deriving such bounds on estimation?

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 may be possible to estimate certain correlation (dependence) measure of X and Y reliably, such as maximal correlation, χ^2 or mutual information. In general, we will denote this measure as $\mathcal{I}(X;Y)$.

Given an upper bound θ on a certain dependence measure $\mathcal{I}$, i.e. $\mathcal{I}(X;Y) \leq \theta$, is it possible to determine a lower bound for the average error of estimating X from Y over all possible estimators? We answer this question in the affirmative. In particular, the problem of computing such a bound for a given distribution p_X and θ is equivalent to computing a distortion-rate function, presented in Definition 9. When the estimation metric is error probability, we call the corresponding distortion-rate function the *error-rate function*, denoted by $e_{\mathcal{I}}(p_X, \theta)$ and given in Definition 10. In the context of security and privacy, this bound characterizes the best estimation of the plaintext that a (computationally unbounded) adversary can make given an observation of the output of the system in terms of the statistic of the distribution of the input and output. This allows, for example, guarantees on correlation measures frequently used in security and privacy settings to be translated into bounds on the estimation error.

Recall that X and Y are discrete random variables with support $\mathcal{X} = [m]$ and $\mathcal{Y} = [n]$, and, consequently, the joint pmf $p_{X,Y}$ can be displayed as the entries of a matrix $\mathbf{P} \in \mathbb{R}^{m \times n}$, where $[\mathbf{P}]_{i,j} = p_{X,Y}(i,j)$. The problem of determining the estimator $\hat{X}$ of X given an observation of Y then reduces to finding a row-stochastic matrix $\mathbf{P}_{\hat{X}|Y} \in \mathbb{R}^{n \times m}$ that is the solution of

$$P_e(X|Y) = \min_{\mathbf{P}_{\hat{X}|Y}} 1 - \mathrm{tr}\left(\mathbf{P} \times \mathbf{P}_{\hat{X}|Y}\right). \tag{41}$$

Note that the previous minimization is a linear program, and by taking its dual the reader can verify that the optimal $\mathbf{P}_{\hat{X}|Y}$ is the maximum a posteriori (MAP) estimator, as expected.

We highlight again that in applications the joint distribution matrix $\mathbf{P}$ may not be known exactly – only a given dependence measure $\mathcal{I}(p_{X,Y})$ may be known. Equation (41) hints that dependence measures that depend on the spectrum of $\mathbf{P}$ may lead to sharp lower bounds for error probability. Indeed, the trace of the product of two matrices is closely related to their spectra (cf. Von Neumman's trace inequality [24, Thm. 7.4.1.1]). This motivates the following question: Are there information measures that capture the spectrum of a joint distribution matrix $\mathbf{P}$? This naturally leads to the consideration of measures of dependence and lower bounds on estimation error based on the PICs. These bounds are derived in Section 4.2, but we first provide an overview of our approach in Section 4.1.

Owing to the nature of the joint distribution, it may be infeasible to estimate X from Y with small estimation error. It is, however, possible that a non-trivial function $f(X)$ exists that is of interest to a learner and can be estimated reliably from Y. 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 learning, privacy, security and information theory. In particular, this setting is related to the information bottleneck method [73] and functional compression [44], where the goal is to compress X into Y such that Y still preserves information about $f(X)$.

For most security applications, minimizing the average error of estimating a hidden variable X from an observation of Y is insufficient. As argued in [59], cryptographic definitions of security, and in particular semantic security [74], 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.

Assuming that $p_{X,Y}$ is unknown, p_X is given and a bound $\mathcal{I}(X;Y) \leq \theta$ is known (where $\mathcal{I}$ is not restricted to being mutual information), we present in Theorem 8 a method for adapting bounds for error probability into bounds for the average estimation error of functions of X given Y. This method depends on a few technical assumptions on the dependence measure (stated in Definition 8 and in Theorem 8), foremost of which is the existence of a lower bound for the error-rate function that is Schur-concave[3] in p_X for a fixed θ. Theorem 8 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 $e_{\mathcal{I}}(p_U, \theta)$, where U is a random variable that is a function X.

Note that Schur-concavity is crucial for this result. In Theorem 7, 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 DPI. This generalizes a result by Ahlswede [18] 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 5. By observing that a particular form of the bound stated in Theorem 6 is Schur-concave, we present in the next section a bound for the error probability of estimating functions in terms of the maximal correlation, stated in Corollary 6.

4.1 A Convex Program for Mapping Information Guarantees to Bounds on Estimation

Throughout the rest of the paper, we let X and Y be two random variables drawn from finite sets $\mathcal{X}$ and $\mathcal{Y}$. We have the following definition.

Definition 8. We say that a function $\mathcal{I}$ that maps any joint probability mass function (pmf) to a non-negative real number is a *dependence measure* (equivalently *measure of dependence*) if for any discrete random variables W, X, Y and Z (i) $\mathcal{I}(p_{X,Y})$ is convex in $p_{Y|X}$ for a fixed p_X, (ii) $\mathcal{I}$ satisfies the DPI, i.e. if $X \to Y \to Z$ then $\mathcal{I}(p_{X,Z}) \leq \mathcal{I}(p_{X,Y})$, and (iii) if W is a one-to-one mapping of Y and Z is a one-to-one mapping of X, then $\mathcal{I}(p_{W,Z}) = \mathcal{I}(p_{X,Y})$ (invariance property). We overload the notation of $\mathcal{I}$ and let $\mathcal{I}(p_{X,Y}) = \mathcal{I}(p_X, p_{Y|X})$ in order to make the dependence on the marginal distribution and the channel (transition probability) clear. Furthermore, we also denote $\mathcal{I}(p_{X,Y}) = \mathcal{I}(X;Y)$ when the distribution is clear from the context. Examples of dependence measures includes maximal correlation, defined in (1), and mutual information.

Now consider the standard estimation setup where a hidden variable X should be estimated from an observed random variable Y. We assume that the joint distribution between $p_{X,Y}$ is not known, but the marginal distribution p_X is known, and that $\mathcal{I}(p_{X,Y}) \leq \theta$ (e.g. security constraint) for a given dependence measure $\mathcal{I}$. Since $\mathcal{I}$ satisfies the DPI, for any estimate $\hat{X}$ of X such that $X \to Y \to \hat{X}$ we have $\mathcal{I}(X;\hat{X}) \leq \mathcal{I}(X;Y) \leq \theta$. The problem of translating a bound on $\mathcal{I}$ into a constraint on how well a hidden variable X can (on average) be estimated from Y given an error function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be approximated by solving the optimization problem

$$\inf_{p_{\hat{X}|X}} \quad \mathbb{E}\left[d(X, \hat{X})\right] \tag{42}$$

$$\text{s.t.} \quad \mathcal{I}(X; \hat{X}) \leq \theta. \tag{43}$$

This motivates the following definition.

Definition 9. We denote the smallest (average) estimation error $D_{\mathcal{I},d}$ for a given dependence measure $\mathcal{I}$ and estimation cost function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as

$$D_{\mathcal{I},d}(p_X, \theta) \triangleq \inf_{p_{\hat{X}|X}} \left\{ \mathbb{E}\left[d(X, \hat{X})\right] \Big| \mathcal{I}(p_X, p_{\hat{X}|X}) \leq \theta \right\}, \tag{44}$$

where the infimum is over all conditional distributions $p_{\hat{X}|X}$.

Observe that for any $p_{Y|X}$ that satisfies $\mathcal{I}(p_X, p_{Y|X}) \leq \theta$

$$D_{\mathcal{I},d}(p_X, \theta) \leq \inf_{p_{\hat{X}|Y}} \left\{ \mathbb{E}\left[d(X, \hat{X})\right] \Big| X \to Y \to \hat{X} \right\},$$

since, by the assumption that $\mathcal{I}$ satisfies the DPI, $\mathcal{I}(X;\hat{X}) \leq \mathcal{I}(X;Y) \leq \theta$. When $\mathcal{I}(X;Y) = I(X;Y)$, $D_{I,d}(p_X, \theta)$ is the distortion-rate function [9, pg. 306]. When the distortion function d is the Hamming

---

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

distortion, $D_{\mathcal{I},d}(p_X,\theta)$ gives the smallest probability of error for estimating X given an observation Y that satisfies $\mathcal{I}(X;Y) \leq \theta$. This case will be of particular interest in this section, motivating the next definition.

Definition 10. Denoting the Hamming distortion metric as

$$d_H(x,y) \triangleq \begin{cases} 0, & x = y, \\ 1, & \text{otherwise,} \end{cases}$$

we define the *error-rate function*[4] for the dependence measure $\mathcal{I}$ as

$$e_{\mathcal{I}}(p_X,\theta) \triangleq D_{\mathcal{I},d_H}(p_X,\theta).$$

The definition of error-rate function directly leads to the following simple lemma.

Lemma 9. *For a given dependence measure $\mathcal{I}$ and any fixed $p_{X,Y}$ such that $\mathcal{I}(p_{X,Y}) \leq \theta$*

$$P_e(X|Y) \geq e_{\mathcal{I}}(p_X,\theta).$$

Proof. Observe that $P_e(X|Y) = \min_{X \to Y \to \hat{X}} \mathbb{E}\left[d_H(X,\hat{X})\right]$, where the minimum is over all distributions $p_{\hat{X}|X}$ that satisfy the Markov constraint $X \to Y \to \hat{X}$. Since $\mathcal{I}$ satisfies the DPI, then $\mathcal{I}(X;\hat{X}) \leq \mathcal{I}(X;Y) \leq \theta$, and the result follows from Definition 9. $\qquad\square$

The previous lemma shows that the characterization of $e_{\mathcal{I}}(p_X,\theta)$ 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) and Y is an adversary's observation (e.g. ciphertext). Knowing $e_{\mathcal{I}}$ allows us to translate an upper bound $\mathcal{I}(X;Y) \leq \theta$ into an estimation guarantee: regardless of an adversary's computational resources, given only access to Y he will not be able to estimate X with an average error probability $P_e(X|Y)$ smaller than $e_{\mathcal{I}}(p_X,\theta)$. Therefore, by simply estimating θ and calculating $e_{\mathcal{I}}(p_X,\theta)$ we are able to evaluate the security threat incurred by an adversary that has access to Y.

Example 3 (Error-rate function for mutual information.)**.** Using the expression for the rate-distortion function under Hamming distortion for mutual information ([75, (9.5.8)]), for $\mathcal{I}(X;Y) = I(X;Y)$ and $\mathcal{X} = [m]$, the error-rate function is given by $e_I(p_X,\theta) = d^*$, where d^* is the solution of

$$h_b(d^*) + d^* \log(m-1) = H(X) - \theta, \tag{45}$$

and $h_b(x) \triangleq -x \log x - (1-x) \log(1-x)$. Denoting $X \to Y \to \hat{X}$ and $p_e \triangleq P_e(X|Y)$, note that (45) implies Fano's inequality [9, 2.140]:

$$h_b(p_e) + p_e \log(m-1) \geq H(X) - I(X;Y) = H(X|Y). \tag{46}$$

4.2 A Lower Bound for Error Probability Based on the PICs

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

Definition 11. Let $\mathbf{\Lambda}(p_{X,Y})$ denote the vector of PICs of a joint distribution $p_{X,Y}$ sorted in decreasing order, i.e. $\mathbf{\Lambda}(p_{X,Y}) = (\lambda_1(X;Y),\ldots,\lambda_d(X;Y))$. We denote $\mathbf{\Lambda}(p_{X,Y}) \leq \widetilde{\mathbf{\lambda}} \triangleq (\widetilde{\lambda}_1,\ldots,\widetilde{\lambda}_d)$ if $\lambda_k(X;Y) \leq \widetilde{\lambda}_k$ for $k \in [d]$

$$\mathcal{R}(q,\widetilde{\mathbf{\lambda}}) \triangleq \left\{ p_{X,Y} | p_X = q \text{ and } \mathbf{\Lambda}(p_{X,Y}) \leq \widetilde{\mathbf{\lambda}} \right\}. \tag{47}$$

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

Theorem 6. *For $\mathbf{\lambda} = (\lambda_1,\ldots,\lambda_d)$ and fixed p_X, let*

$$k^* \triangleq \max \left\{ k \in [m] \;\middle|\; p_X(k) \geq \sum_{i \in [m]} p_X(i)^2 \right\}. \tag{48}$$

---

[4]The term *error-rate* function is used in the same sense as *distortion-rate* function in rate distortion theory [9, Chap. 10]. We adopt "error" instead of distortion here since we only consider Hamming distance as the distortion metric.

In addition, let $\mathbf{p} = (p_X(1), \ldots, p_X(m))$ and $\boldsymbol{\lambda}_{k^*} = (\lambda_1, \ldots, \lambda_{k^*}, \lambda_{k^*}, \lambda_{k^*+1}, \ldots, \lambda_{m-1})$ (where $\lambda_m \triangleq 0$ and $\boldsymbol{\lambda}_m = (\lambda_1, \ldots, \lambda_m)$). Defining

$$u(p_X, \boldsymbol{\lambda}) \triangleq \min_{0 \leq \beta \leq p_X(2)} \beta + \sqrt{\mathbf{p}^T \boldsymbol{\lambda}_{k^*} - \lambda_{k^*} \|\mathbf{p}\|_2^2 + \left\| [\mathbf{p} - \beta]^+ \right\|_2^2} \,,$$

then for any $(X,Y) \sim q_{X,Y} \in \mathcal{R}(p_X, \boldsymbol{\lambda})$,

$$P_e(X|Y) \geq 1 - u(p_X, \boldsymbol{\lambda}). \tag{49}$$

Proof. The proof of the theorem is presented in Appendix C. ☐

Remark 8. If $\lambda_i = 1$ for all $1 \leq i \leq d$, (49) reduces to $P_e(X|Y) \geq 0$. Furthermore, if $\lambda_i = 0$ for all $1 \leq i \leq d$, (49) simplifies to $P_e(X|Y) \geq 1 - p_X(1)$.

We now present a few direct but, as we shall show in the next section, useful corollaries of the result in Theorem 6. We note that a bound with the same square-root order dependence on χ^2-divergence as Eq. (50) below has appeared in the context of bounding the minmax decision risk in [76, Eq. (3.4)]. However, the proof technique used in [76] does not seem to lead to the general bound presented in Theorem 6.

Corollary 3. *If X is uniformly distributed in $[m]$, then*

$$P_e(X|Y) \geq 1 - \frac{1}{m} - \frac{\sqrt{(m-1)\chi^2(X;Y)}}{m} \,. \tag{50}$$

Furthermore, for $\rho_m(X;Y) = \sqrt{\lambda_1}$

$$P_e(X|Y) \geq 1 - \frac{1}{m} - \sqrt{\lambda_1}\left(1 - \frac{1}{m}\right)$$
$$= 1 - \frac{1}{m} - \rho_m(X;Y)\left(1 - \frac{1}{m}\right).$$

Corollary 4. *For any pair of variables (X,Y) with marginal distribution in X equal to p_X and maximal correlation (largest principal inertia) $\rho_m(X;Y)^2 = \lambda_1$, we have for all $\beta \geq 0$*

$$P_e(X|Y) \geq 1 - \beta - \sqrt{\lambda_1\left(1 - \sum_{i=1}^m p_X(i)^2\right) + \sum_{i=1}^m \left([p_X(i) - \beta]^+\right)^2} \,. \tag{51}$$

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

$$P_e(X|Y) \geq 1 - p_X(2) - \sqrt{\lambda_1\left(1 - \sum_{i=1}^m p_X(i)^2\right) + (p_X(1) - p_X(2))^2} \tag{52}$$

$$\geq 1 - p_X(1) - \rho_m(X;Y)\sqrt{\left(1 - \sum_{i=1}^m p_X(i)^2\right)} \,, \tag{53}$$

where (53) follows from (52) being decreasing in $p_X(2)$.

Remark 9. The bounds (51) and (53) are particularly helpful for showing how the error probability scales with the input distribution and the maximal correlation. For a given $p_{X,Y}$, recall that

$$\mathsf{Adv}(X|Y) \triangleq 1 - p_X(1) - P_e(X|Y),$$

defined in (6), 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 (53)

$$\mathsf{Adv}(X|Y) \leq \rho_m(X;Y)\sqrt{\left(1 - \sum_{i=1}^m p_X(i)^2\right)}$$

$$\leq \rho_m(X;Y) = \sqrt{\lambda_1(X;Y)}.$$

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

We present next results on the extremal properties of the error-rate function. This analysis will be particularly useful for determining how to bound the probability of error of estimating functions of a random variable.

4.3 Extremal Properties of the Error-Rate Function and Bounding the Estimation Error of Functions of a Hidden Random Variable

Owing 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. We will now prove that, for a fixed θ, $e_{\mathcal{I}}(p_X, \theta)$ is *Schur-concave* in p_X if $\mathcal{I}(p_X, p_{\hat{X}|X})$ is concave in p_X for a fixed $p_{\hat{X}|X}$. Ahlswede [18, 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 simpler and more general, and is based on a proof technique used by Ahlswede in [18, Theorem 1].

Theorem 7. *If $\mathcal{I}(p_X, p_{\hat{X}|X})$ is concave in p_X for a fixed $p_{\hat{X}|X}$, then $e_{\mathcal{I}}(p_X, \theta)$ is Schur-concave in p_X for a fixed θ.*

Proof. The proof is presented in Appendix C. □

For a given integer $1 \leq M \leq |\mathcal{X}|$, we define

$$\mathcal{F}_M \triangleq \{f : \mathcal{X} \to \mathcal{U} \mid f \text{ is surjective and } |\mathcal{U}| \geq M\} \tag{54}$$

and

$$P_{e,M}(X|Y) \triangleq \min_{f \in \mathcal{F}_M} P_e(f(X)|Y). \tag{55}$$

$P_{e,|\mathcal{X}|}(X|Y)$ is simply the error probability of estimating X from Y, i.e. $P_{e,|\mathcal{X}|}(X|Y) = P_e(X|Y)$. The surjectivity condition in the definition of $\mathcal{F}_M$ is mostly technical, and was added to (i) avoid the constant function being in $\mathcal{F}_M$ (which would render the estimation error trivial) and (ii) enable the use of Schur-concavity results to derive bounds on estimation error. Note that, in the discrete setting considered here, by varying M we span the set of all functions of X, so there is no loss of generality. Nevertheless, there are practical settings where this condition naturally arises. In classification problems, for example, the surjectivity condition would correspond to the number of classes used to classify X.

The next theorem shows that a lower bound for $P_{e,M}$ can be derived for any dependence measure $\mathcal{I}$ 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 8. *For a given $M \in [m]$ and p_X with $\mathcal{X} = [m]$ and $p_X(1) \geq p_X(2) \geq \ldots \geq p_X(m)$, let $U = g_M(X)$, where $g_M : \{1, \ldots, m\} \to \{1, \ldots, M\}$ 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 dependence measure $\mathcal{I}$, there exists a function $L_{\mathcal{I}}(\cdot, \cdot)$ such that for all distributions q_X and any θ, $e_{\mathcal{I}}(q_X, \theta) \geq L_{\mathcal{I}}(q_X, \theta)$. If $L_{\mathcal{I}}(p_X, \theta)$ is Schur-concave in p_X, then for $X \sim p_X$ and $\mathcal{I}(X; Y) \leq \theta$,

$$P_{e,M}(X|Y) \geq L_{\mathcal{I}}(p_U, \theta). \tag{56}$$

In addition[6], for any $S \to X \to Y$ such that p_U majorizes p_S,

$$P_e(S|Y) \geq L_{\mathcal{I}}(p_U, \theta). \tag{57}$$

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

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

---

[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 $k = 2, \ldots, M$.
[6] We thank Dr. Nadia Fawaz (nadia.fawaz@gmail.com) for pointing out this extension.

where (a) follows from the DPI, (b) follows from $e_{\mathcal{I}}(q_X, \theta)$, being decreasing in θ, (c) follows from $e_{\mathcal{I}}(q_X, \theta) \geq L_{\mathcal{I}}(q_X, \theta)$ for all q_X, and θ and (d) follows from the Schur-concavity of the lower bound and by observing that p_U majorizes $p_{f(X)}$ for every $f \in \mathcal{F}_M$. In the case of $P_e(S|X)$, the same inequalities hold with S playing the role of $f(X)$ in (a) and (b), and the last inequality also following from Schur-concavity of $L_{\mathcal{I}}(p_S, \theta)$ in p_S. □

Remark 10. The function $g_M(X) = U$ in Theorem 8 is formed by adding the most likely symbols of X, and, consequently, p_U majorizes any other distribution $p_{f(X)}$ for $f \in \mathcal{F}_M$. The function g_M can thus be regarded as the "least uncertain" function of X in $\mathcal{F}_M$ in the following sense: since Rényi entropy[7] is Schur-concave, $H_\alpha(g_M(X)) \leq H_\alpha(f(X))$ for all $f \in \mathcal{F}_M$ and $\alpha \geq 0$.

The following results illustrates how Theorem 8 can be used for mutual information and maximal correlation.

Corollary 5. *Let* $I(X;Y) \leq \theta$. *Then*

$$P_{e,M}(X|Y) \geq d^*$$

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. Let $R_I(p_X, \delta) \triangleq \min_{p_{\hat{X}|X}} \{I(X; \hat{X}) | \mathbb{E}[d_H(X, \hat{X})] \leq \delta\}$ be the well known rate-distortion function under Hamming distortion. Then $R_I(p_X, \delta)$ satisfies ([75, (9.5.8)]) $R_I(p_X, \delta) \geq H(X) - h_b(d^*) - d^* \log(m-1)$. The result follows from Theorem 7, since mutual information is concave in p_X. □

Corollary 6. *Let* $\mathcal{J}_1(X;Y) = \rho_m(X;Y) \leq \theta$. *Then*

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

where $P_{e,M}(X|Y)$ *is defined in* (55) *and* U *is defined as in Theorem 8.*

Proof. The proof follows directly from Theorems 2, 3 and Corollary 4, by noting that (53) is Schur-concave in p_X. □

The previous result leads to the next theorem, which states that the probability of guessing *any* function of a hidden variable X from an observation Y is upper bounded by the maximal correlation of X and Y.

Theorem 9. *Let* p_X *be fixed,* $|\mathcal{X}| < \infty$ *and* $\mathcal{F}_M$ *be given in* (54). *Define (cf.* (6))

$$\mathsf{Adv}_M(X|Y) \triangleq \max \left\{ 1 - \max_{k \in [M]} p_{f(X)}(k) - P_e(f(X)|Y) \mid f \in \mathcal{F}_M \right\}.$$

Then

$$\mathsf{Adv}_M(X|Y) \leq \rho_m(X;Y) \sqrt{1 - \frac{1}{M}} \leq \rho_m(X;Y). \tag{58}$$

Proof. For $f \in \mathcal{F}_M$

$$\mathsf{Adv}(f(X)|Y) \leq \rho_m(f(X);Y) \sqrt{1 - \sum_{i \in [M]} p_{f(X)}(i)^2}$$

$$\leq \rho_m(X;Y) \sqrt{1 - \sum_{i \in [M]} p_{f(X)}(i)^2}$$

$$\leq \rho_m(X;Y) \sqrt{1 - \frac{1}{M}},$$

where the first inequality follows from (53) and the definition (6), the second inequality follows by combining Theorem 3 (DPI for the PICs) and the fact that $\lambda_1(f(X); X) \leq 1$, which leads to $\rho_m(f(X);Y) \leq \rho_m(X;Y)$, and the last inequality follows from the fact that $\sum_{i \in [M]} p_{f(X)}(i)^2$ is minimized when $p_{f(X)}$ is uniform. The result follows by maximizing over all $f \in \mathcal{F}_M$. □

---

[7]The Rényi entropy of a discrete random variable X is given by $H_\alpha(X) \triangleq \frac{1}{1-\alpha} \log\left(\sum_{x \in \mathcal{X}} p_X(x)^\alpha\right)$.

The results presented in this section demonstrate that the PICs are a useful information measure that can shed light on fundamental limits of estimation. In particular, Theorem 9 connects the largest PIC, namely the maximal correlation, with the probability of correctly guessing *any* function of a hidden, discrete random variable. The PICs also provide a characterization of the functions of a hidden variable that can (or cannot) be estimated with small mean-squared error (Theorem 1). In the next section, we explore applications of the PICs to privacy and security.

5 Applications of the PICs to Security and Privacy

In this section, we present a few applications of the principal inertia components to problems in security and privacy. We adopt the privacy against statistical inference framework presented in [19]. This setup, called the *Privacy Funnel*, was introduced in [63]. Consider two communicating parties, namely Alice and Bob. Alice's goal is to disclose to Bob information about a set of measurement points, represented by the random variable X. Alice discloses this information in order to receive some utility from Bob. Simultaneously, Alice wishes to limit the amount of information revealed about a private random variable S that is dependent on X. For example, X may represent Alice's movie ratings, released to Bob in order to receive movie recommendations, whereas S may represent Alice's political preference or yearly income. Bob will try to extract the maximum amount of information about S from the data disclosed by Alice.

Instead of revealing X directly to Bob, Alice releases a new random variable, denoted by Y. This random variable is produced from X through a random mapping $p_{Y|X}$, called the *privacy-assuring mapping*. We assume that $p_{S,X}$ is fixed and known by both Alice and Bob, and $S \to X \to Y$. Alice's goal is to find a mapping $p_{Y|X}$ that minimizes $I(S;Y)$, while guaranteeing that the information disclosed about X is above a certain threshold t, i.e. $I(X;Y) \geq t$. We refer to the quantity $I(S;Y)$ as the *disclosed private information*, and $I(X;Y)$ as the *disclosed useful information*. As discussed in Section 1.1, when $I(S;Y) = 0$, we say that *perfect privacy* is achieved, i.e. Y does not reveal any information about S. We consider here the non-interactive, one-shot regime, where Alice discloses information once, and no additional information is released. We also assume that Bob knows the privacy-assuring mapping $p_{Y|X}$ chosen by Alice, and no side information is available to Bob about S besides Y.

5.1 The Privacy Funnel

We define next the privacy funnel function, which captures the smallest amount of disclosed private information for a given threshold on the amount of disclosed useful information. We then characterize properties of the privacy funnel function in the rest of this section.

Definition 12. For $0 \leq t \leq H(X)$ and a joint distribution $p_{S,X}$ over $\mathcal{S} \times \mathcal{X}$, we define the *privacy funnel function* $G_I(t, p_{S,X})$ as

$$G_I(t, p_{S,X}) \triangleq \inf \{I(S;Y)|I(X;Y) \geq t, S \to X \to Y\}, \tag{59}$$

where the infimum is over all mappings $p_{Y|X}$ such that $\mathcal{Y}$ is finite. For a fixed $p_{S,X}$ and $t \geq 0$, the set of pairs $\{(t, G_I(t, p_{S,X}))\}$ is called the *privacy region* of $p_{S,X}$.

We now enunciate a few useful properties of $G_I(t, p_{S,X})$ and the privacy region.

Lemma 10.

$$G_I(t, p_{S,X}) = \min_{p_{Y|X}} \{I(S;Y)|I(X;Y) \geq t, S \to X \to Y, \ |\mathcal{Y}| \leq |\mathcal{X}|+1\}. \tag{60}$$

In addition, for a fixed $p_{S,X}$, the mapping $t \mapsto \frac{G_I(t,p_{S,X})}{t}$ is non-decreasing, and $G_I(t, p_{S,X})$ is convex in t.

Proof. The proof is in Appendix D. $\qquad\qquad\square$

Lemma 11. *For $0 \leq t \leq H(X)$,*

$$\max\{t - H(X|S), 0\} \leq G_I(t, p_{S,X}) \leq \frac{tI(X;S)}{H(X)}. \tag{61}$$
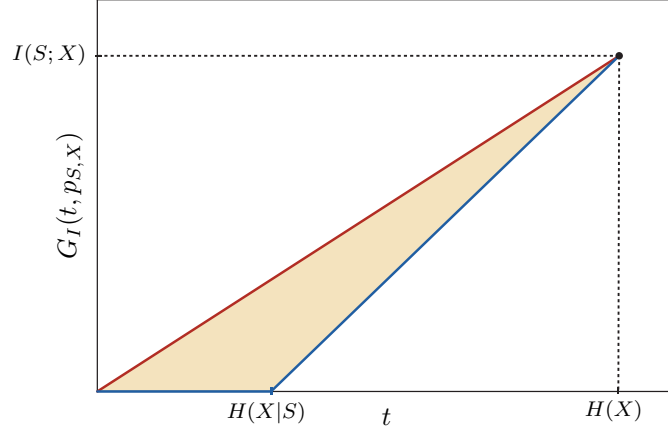
Figure 3: For a fixed $p_{S,X}$, the privacy region is contained within the shaded area. The red and the blue lines correspond, respectively, to the upper and lower bounds presented in Lemma 11.

Proof. Observe that $G_I(H(X), p_{S,X}) = I(X; S)$, since $I(X; Y) = H(X)$ implies that $p_{Y|X}$ is a one-to-one mapping of X. The upper bound then follows directly from (99).

Clearly $G_I(t, p_{S,X}) \geq 0$. In addition, for any $p_{Y|X}$,

$$
\begin{aligned}
I(S; Y) &= I(X; Y) - I(X; Y|S) \\
&\geq I(X; Y) - H(X|S) \\
&\geq t - H(X|S),
\end{aligned}
$$

proving the lower bound. $\qquad\square$

Figure 3 illustrates the bounds from Lemma 11. The privacy region is contained withing the shaded area. The next two examples illustrate that both the upper bound (red line) and the lower bound (blue line) of the privacy region can be achieved for particular instances of $p_{S,X}$.

Example 4. Let $X = (S, W)$, where $W \perp\!\!\!\perp S$. Then by setting $Y = W$, we have $I(S; Y) = 0$ and $I(X; Y) = H(W) = H(X|S)$. Consequently, from Lemmas 10 and 11, $G_I(t, p_{S,X}) = 0$ for $t \in [0, H(X|S)]$. By letting $Y = W$ w.p. λ and $Y = (S, W)$ w.p. $1 - \lambda$ for $\lambda \in [0, 1]$, the lower-bound $G_I(t, p_{S,X}) = t - H(X|S)$ can be achieved for $H(X|S) = H(W) \leq t \leq H(X)$. Consequently, the lower bound in (61) is sharp.

Example 5. Now let $X = f(S)$. Then $I(X; S) = H(X)$ and

$$
I(S; Y) = I(X; Y) - I(X; Y|S) = I(X; Y).
$$

Consequently, $G_I(t, p_{S,X}) = t$, and the upper bound in (61) is sharp.

5.2 The Optimal Privacy-Utility Coefficient and the Smallest PIC

We now study the smallest possible ratio between disclosed private and useful information, defined next.

Definition 13. The *optimal privacy-utility coefficient* for a given distribution $p_{S,X}$ is given by

$$
v^*(p_{S,X}) \triangleq \inf_{p_{Y|X}} \frac{I(S; Y)}{I(X; Y)}. \tag{62}
$$

It follows directly from Lemma 10 that

$$
v^*(p_{S,X}) = \lim_{t \to 0} \frac{G_I(t, p_{S,X})}{t}. \tag{63}
$$

We show in Section 5.3 that the value of $v^*(p_{S,X})$ is related to the smallest PIC of $p_{S,X}$ (i.e. the smallest eigenvalue of the spectrum of the conditional expectation operator, defined below). We also prove that $v^*(p_{S,X}) = 0$ is a necessary and sufficient condition for achieving perfect privacy while disclosing a non-trivial amount of useful information. Before introducing these results, we present an alternative characterization of $v^*(p_{S,X})$ (Lemma 12), and introduce a measure based on the smallest PIC (Definition 14) and an auxiliary result (Lemma 13).

27

Remark 11. The proofs of Lemma 12 and Lemma 14 in this section are closely related to [23]. We acknowledge that their proof techniques inspired some of the results presented here.

The next result provides a characterization of the optimal privacy-utility coefficient.

Lemma 12. *Let q_S denote the distribution of S when $p_{S|X}$ is fixed and $X \sim q_X$. Then*

$$v^*(p_{S,X}) = \inf_{q_X \neq p_X} \frac{D(q_S||p_S)}{D(q_X||p_X)}. \tag{64}$$

Proof. The proof is in Appendix D. □

The smallest PIC is of particular interest for privacy, and upper bounds the value of $v^*(p_{S,X})$. In particular, we will be interested in the coefficient $\delta(p_{S,X})$, defined bellow

Definition 14. *Let $d \triangleq \min\{|\mathcal{S}|, |\mathcal{X}|\} - 1$, and $\lambda_d(S; X)$ the smallest PIC of $p_{S,X}$. We define*

$$\delta(p_{S,X}) \triangleq \begin{cases} \lambda_d(S; X) & \text{if } |\mathcal{X}| \leq |\mathcal{S}|, \\ 0 & \text{otherwise.} \end{cases} \tag{65}$$

The following lemma provides a useful characterization of $\delta(p_{S,X})$, related to the interpretation of the PICs as the spectrum of the conditional expectation operator given in Theorem 1. This result is a direct consequence of Theorem 1, and we present a self-contained proof in Appendix D.

Lemma 13. *For a given $p_{S,X}$,*

$$\delta(p_{S,X}) = \min \left\{ \|\mathbb{E}\left[f(X)|S\right]\|_2^2 \,\middle|\, f : \mathcal{X} \to \mathbb{R}, \mathbb{E}\left[f(X)\right] = 0, \|f(X)\|_2 = 1 \right\}. \tag{66}$$

5.3 Information Disclosure with Perfect Privacy

If $v^*(p_{S,X}) = 0$, then it may be possible to disclose some information about X without revealing any information about S. However, since $G_I(0, p_{X,S}) = 0$, it is not immediately clear that $v^*(p_{S,X}) = 0$ implies that there exists t strictly bounded away from 0 such $G_I(t, p_{X,S}) = 0$. This would represent the ideal privacy setting, since, from Lemma 10, there would exist a privacy-assuring mapping that allows the disclosure of some non-negligible amount of useful information while guaranteeing $I(S; Y) = 0$. This, in turn, would mean that perfect privacy is achievable with non-negligible utility *regardless of the specific privacy metric used*, since S and Y would be independent.

In this section, we prove that if the optimal privacy-utility coefficient is 0, then there indeed exists a privacy-assuring mapping that allows the disclosure of a non-trivial amount of useful information while guaranteeing perfect privacy. We also show that the value of $\delta(p_{S,X})$ is closely related to $v^*(p_{S,X})$. This relationship is analogous to the one between the hypercontractivity coefficient s^*, defined in [22] and [77], and the maximal correlation ρ_m. In particular, as shown in the next two lemmas, $v^*(p_{S,X}) \leq \delta(p_{S,X})$ and $v^*(p_{S,X}) = 0 \iff \delta(p_{S,X}) = 0$.

Lemma 14. *For any $p_{S,X}$ with finite support $\mathcal{S} \times \mathcal{X}$,*

$$v^*(p_{S,X}) \leq \delta(p_{S,X}). \tag{67}$$

and

$$\inf_{p_X} v^*(p_{S,X}) = \inf_{p_X} \delta(p_{S,X}). \tag{68}$$

Proof. The proof is in Appendix D. □

The next theorem proves that $\delta(p_{S,X})$ can serve as a proxy for perfect privacy, since the optimal privacy-utility coefficient is 0 if and only if $\delta(p_{S,X})$ is also 0.

Lemma 15. *Let $p_{S,X}$ be such that $H(X) > 0$ and $\mathcal{S}$ and $\mathcal{X}$ are finite. Then*[8]

$$v^*(p_{S,X}) = 0 \iff \delta(p_{S,X}) = 0. \tag{69}$$

Proof. The proof can be found in Appendix D. □

---

[8]If S is binary, then (69) implies that perfect privacy is achievable iff S and X are independent (since $\delta(p_{S,X}) = \rho_m(S; X)^2$), recovering [61, Thm. 2].

We are now ready to prove that a non-trivial amount of useful information can be disclosed without revealing any private information if and only if $v^*(p_{S,X}) = 0$ (or equivalently, $\delta(p_{S,X}) = 0$). This result follows naturally from Theorem 15, since $v^*(p_{S,X}) = 0$ implies that $\delta(p_{S,X}) = 0$, which means that the matrix $\mathbf{Q}$ and, consequently, $\mathbf{P}_{S|X}$, is either not full rank or has more columns than rows (i.e. $|\mathcal{X}| > |\mathcal{S}|$). This, in turn, can be exploited in order to find a mapping $p_{Y|X}$ such that Y reveals some information about X, but no information about S. This argument is made precise in the next theorem.

Remark 12. When $\mathbf{P}_{S|X}$ is not full rank or has more columns than rows, then S and X are weakly independent. As shown in [78, Thm. 4] and [61], this implies that a privacy-assuring mapping that achieves perfect privacy while disclosing a non-trivial amount of useful information can be found. Theorem 10 recovers this result in terms of the smallest PIC, and Corollary 8 provides an estimate of the amount of useful information that can be revealed with perfect privacy.

Theorem 10. *For a given $p_{S,X}$, there exists a privacy-assuring mapping $p_{Y|X}$ such that $S \to X \to Y$, $I(X;Y) > 0$ and $I(S;Y) = 0$ if and only if $\delta(p_{S,X}) = 0$ (equivalently $v^*(p_{S,X}) = 0$). In particular,*

$$\exists t_0 > 0 : G_I(t_0, p_{S,X}) = 0 \iff \delta(p_{S,X}) = 0. \tag{70}$$

Proof. The direct part of the theorem follows directly from the definition of $v^*(p_{S,X})$ and Lemma 15. Assume that $\delta(p_{S,X}) = 0$. Then, from Lemma 13, there exists $f : \mathcal{X} \to \mathbb{R}$ such that $\|f(X)\|_2 = 1$, $\mathbb{E}[f(X) = 0]$, and $\|\mathbb{E}[f(X)|S]\|_2 = 0$. Consequently, $\mathbb{E}[f(X)|S = s] = 0$ for all $s \in \mathcal{S}$.

Fix $\mathcal{Y} = [2]$, and, for $\epsilon > 0$ and ϵ appropriately small,

$$p_{Y|X}(y|x) = \begin{cases} \frac{1}{2} - \epsilon f(x), & y = 1, \\ \frac{1}{2} + \epsilon f(x), & y = 2. \end{cases}$$

Note that it is sufficient to choose $\epsilon = (2 \max_{x \in \mathcal{X}} |f(X)|)^{-1}$, so ϵ is strictly bounded away from 0. In addition, $p_Y(1) = 1/2$. Therefore,

$$I(X;Y) = 1 - \sum_{x \in \mathcal{X}} p_X(x) h_b \left(\frac{1}{2} + \epsilon f(x) \right) > 0. \tag{71}$$

Since $S \to X \to Y$,

$$\begin{aligned} p_{Y|S}(y|s) &= \sum_{x \in \mathcal{X}} p_{Y|X}(y|x) p_{X|S}(x|s) \\ &= \sum_{x \in \mathcal{X}} \left(\frac{1}{2} + (-1)^y \epsilon f(x) \right) p_{X|S}(x|s) \\ &= 1/2 + (-1)^y \epsilon \mathbb{E}[f(X)|S = s] \\ &= 1/2, \end{aligned}$$

and, consequently, S and Y are independent. Then $I(S;Y) = 0$, and the result follows. $\square$

The previous result proves that if either $|\mathcal{X}| > |\mathcal{S}|$ or the smallest principal inertia component of $p_{S,X}$ is 0 (i.e. $\delta(p_{S,X}) = 0$), then it is possible to achieve perfect privacy while disclosing some useful information. In particular, the value of t_0 in (100) is lower-bounded by the expression in (71). We note that this result would not necessarily hold if $\mathcal{S}$ and $\mathcal{X}$ are not finite sets.

Since $I(S;Y) = 0$ implies that S and Y are independent, Theorem 10 holds not only for mutual information, but also for *any* dependence measure $\mathcal{I}$, defined in Definition 8, that satisfies $\mathcal{I}(X;Y) = 0$ if and only if X and Y are independent. This leads to the following result.

Corollary 7. *Let $p_{S,X}$ be given, and $\mathcal{I}$ be a non-negative dependence measure (e.g. total variation or maximal correlation, cf. Definition 8) such that for any two random variable A and B, $\mathcal{I}(A;B) = 0 \iff A \perp\!\!\!\perp B$. Then there exists $p_{Y|X}$ such that $S \to X \to Y$, $\mathcal{I}(X;Y) > 0$ and $\mathcal{I}(S;Y) = 0$ if and only if $\delta(p_{S,X}) = 0$.*

Proof. This is a direct consequence of Theorem 10, since $\mathcal{I}(X;Y) > 0 \iff I(X;Y) > 0$ and $\mathcal{I}(S;Y) = 0 \iff I(S;Y) = 0$. $\square$

Remark 13. As long as privacy is measured in terms of statistical dependence (with perfect privacy implying statistical independence) and some utility can be derived when Y is not independent of X, then $\delta(p_{S,X})$ fully characterizes when perfect privacy is achievable with non-trivial utility.

29

We present next an explicit lower bound for the largest amount of useful information that can be disclosed while guaranteeing perfect privacy. The result follows directly from the construction used in the proof of Theorem 10, and is presented in Appendix D.

Corollary 8. *For fixed $p_{S,X}$, let*

$$\mathcal{F}_0 \triangleq \{f : \mathcal{X} \to \mathbb{R} | \mathbb{E}[f(X)] = 0, \ \|f(X)\|_2 = 1, \ \|\mathbb{E}[f(X)|S]\|_2 = 0\} \cup w_0,$$

where w_0 is the trivial function that maps $\mathcal{X}$ to $\{0\}$. Then $G_I(t, p_{S,X}) = 0$ for $t \in [0, t^*]$, where

$$t^* \geq 1 - \max_{f \in \mathcal{F}_0} \mathbb{E}\left[h_b\left(\frac{1}{2} + \frac{f(X)}{2\|f\|_\infty}\right)\right]. \tag{72}$$

Furthermore, the lower bound for t^* is sharp when $\delta(p_{S,X}) = 0$, i.e. there exists a $p_{S,X}$ such that $t^* > 0$ and $G_I(t, p_{S,X}) = 0$ if and only if $t \in [0, t^*]$.

The previous bound for t^* can be loose, especially if $|\mathcal{X}|$ is large. In addition, the right-hand side of (72) can be made arbitrarily small by decreasing $\min_{x \in \mathcal{X}} p_X(x)$. Nevertheless, (72) is an explicit bound on the amount of useful information that can be disclosed with perfect privacy.

When $S^n = (S_1, \ldots, S_n)$ and $X^n = (X_1, \ldots, X_n)$, where $(S_i, X_i) \sim p_{S,X}$ are i.i.d. random variables, the next proposition states that $\delta(p_{S^n, X^n}) = \delta(p_{S,X})^n$. Consequently, as long as $\delta(p_{S,X}) < 1$, it is possible to disclose a non-trivial amount of useful information while disclosing an arbitrarily small amount of private information by making n sufficiently large. Loosely speaking, this is similar to hiding a needle in a haystack: As the number of available samples of S and X increases, we can use the additional randomness to better hide the private variables S_i.

Proposition 1. *Let $S^n = (S_1, \ldots, S_n)$ and $X^n = (X_1, \ldots, X_n)$, where $(S_i, X_i) \sim p_{S,X}$ are i.i.d. random variables. Then*

$$v^*(p_{S^n, X^n}) \leq \delta(p_{S^n, X^n}) = \delta(p_{S,X})^n. \tag{73}$$

Proof. The result is a direct consequence of the tensorization property of the principal inertia components, presented in Lemma 1. □

6 Final Remarks

The PICs are powerful information-theoretic metrics that provide both (i) a measure of dependence between two random variables X and Y, and (ii) a complete characterization of which functions of X can be reliably estimated (in terms of mean-squared error) given an observation of Y. As shown here, the PICs play can be used for deriving bounds on one-bit functions of a channel input given a channel output. Furthermore, in privacy applications, we proved that perfect privacy can be achieved if and only if the smallest PIC is zero. The PICs were also used to derive bounds on estimation error probability. In particular, we demonstrated that the largest PIC (equivalently, the maximal correlation $\rho_m(X;Y)$) plays a key role in estimation:

$$\mathsf{Adv}(f(X)|Y) \leq \rho_m(X;Y),$$

i.e. the advantage over a random guess of estimating any function of X given Y is at most $\rho_m(X;Y)$.

Information theoretic security and privacy applications provide fertile ground for the use of PICs, specially when privacy is measured in terms of how well an adversary can estimate a secret (private) variable. The principal functions (cf. Definition 1) provide a basis for the finite-variance functions of a random variable, and the PICs measure the MMSE of estimating each of these functions. Consequently, the PICs provide a characterization of which functions of X can or cannot be inferred reliably (in terms of MMSE) from an observation of Y. This property can be used in privacy applications: For example, in order to quantify how well an adversary can estimate a private function $S = f(X)$ given a disclosed variable Y, it is sufficient to express $f(X)$ in terms of the principal functions of $p_{X,Y}$. The adversary's ability of correctly estimating $f(X)$ is then entirely determined by the PICs of $p_{X,Y}$.

More precisely, for $f : \mathcal{X} \to \mathbb{R}$, the mean squared-error $f(X)$ given Y can be expressed as

$$\mathsf{mmse}(f(X)|Y) = \mathbb{E}\left[f(X)^2 - \mathbb{E}[f(X)|Y]^2\right]$$

$$= \|f(X)\|_2^2 \left(1 - \frac{\|\mathbb{E}[f(X)|Y]\|_2^2}{\|f(X)\|_2^2}\right), \tag{74}$$

Consequently, the MMSE depends on the spectrum of the conditional expectation operator $(T_Y f)(y) \triangleq \mathbb{E}[f(X)|Y = y]$ which, in turn, is composed by the principal inertia components (cf. Theorem 1). When

$\mathbb{E}[f(X)] = 0$ and $\mathbb{E}\left[f(X)^2\right] = 1$, one can determine functions $f_1, f_2, \ldots$ as in Theorem 1, such that f_i is given by

$$f_i = \operatorname{argmax}\left\{\|\mathbb{E}[f(X)|Y]\|_2^2 \,|\, f : \mathcal{X} \to \mathbb{R},\; \mathbb{E}[f(X)] = 0,\; \mathbb{E}\left[f(X)^2\right] = 1,\right.$$
$$\left. \mathbb{E}[f(X)f_j(X)] = 0 \text{ for } 1 \le j \le i - 1\right\}.$$

Then

$$\|\mathbb{E}[f_i(X)|Y]\|_2^2 = \lambda_i(X;Y).$$

It follows directly that, for any non-trivial function $f : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}[f(X)] = 0$,

$$\mathsf{mmse}(f(X)|Y) \ge \|f(X)\|_2^2\left(1 - \rho_m(X;Y)^2\right), \tag{75}$$

with equality if $f(X) = cf_1(X)$, where $c = \|f(X)\|_2$. Therefore, for a fixed variance c, $cf_1(X)$ is the function of X that can be most reliably estimated (in terms of mean-squared error) from all possible mappings $\mathcal{X} \to \mathbb{R}$. Furthermore,

$$\mathsf{mmse}(f(X)|Y) = \|f(X)\|_2^2\left(1 - \sum_i c_i^2 \lambda_i(X;Y)\right), \tag{76}$$

where $c_i \triangleq \mathbb{E}[f(X)f_i(X)]/\|f(X)\|_2$ and $\sum_i c_i^2 = 1$. Consequently, functions that are closely "aligned" with f_i for small i cannot be inferred with small mean squared-error.

In privacy applications with estimation constraints, this result sheds light on the nature of the fundamental tradeoff between privacy and utility. If X and Y correspond, respectively, to the input and output of a privacy-assuring mapping, then the PICs and corresponding principal functions of $p_{X,Y}$ will determine which functions (features) of X remain private. If, for example, the principal functions corresponding to small PICs also span functions of X that should be reliably estimated from Y for utility purposes, then the privacy-assuring mapping $p_{Y|X}$ will provide an unfavorable tradeoff between privacy an utility.

As another example, assume that we wish to design a privacy-assuring mapping where the secret $S = (h_1(X), \ldots, h_t(X))$ is composed by a certain set of functions (features) $h_1, \ldots, h_t$ of X that are supposed to remain private. The privacy-assuring mapping $p_{Y|X}$ should then assure that the principal functions that span the subspace formed by $(h_1(X), \ldots, h_t(X))$ must have small PICs. These examples, together with the results presented here, motivate the future use of PICs to drive the design of privacy-assuring mappings that achieve a favorable tradeoff between privacy and utility.

Acknowledgments

Appendix A   Proofs from Section 2

Lemma 2

Proof. Let $f \in \mathcal{L}_2(p_X)$, $\mathbb{E}[f(X)] = 0$ and $g \in \mathcal{L}_2(Z)$, $\mathbb{E}[g(Z)] = 0$, $\|g(Z)\|_2 = 1$. Then

$$\mathbb{E}[f(X)g(Z)] = \mathbb{E}[\mathbb{E}[f(X)g(Z)|Y]]$$
$$\overset{(a)}{=} \mathbb{E}[\mathbb{E}[f(X)|Y]\,\mathbb{E}[g(Z)|Y]]$$
$$\overset{(b)}{\le} \|\mathbb{E}[f(X)|Y]\|_2 \|\mathbb{E}[g(Z)|Y]\|_2$$
$$\overset{(c)}{\le} \sqrt{\lambda_1(Z;Y)}\|\mathbb{E}[f(X)|Y]\|_2,$$

where (a) follows from the assumption that $X \to Y \to Z$, (b) follows from the Cauchy-Schwarz inequality, and (c) follows from characterization (3) in Theorem 1. By choosing $g(z) = \mathbb{E}[f(X)|Z = z]/\|\mathbb{E}[f(X)|Z]\|_2$ and using the last inequality, we have

$$\mathbb{E}[f(X)g(Z)] = \mathbb{E}[\mathbb{E}[f(X)|Z]\,g(Z)] = \|\mathbb{E}[f(X)|Z]\|_2 \le \sqrt{\lambda_1(Z;Y)}\|\mathbb{E}[f(X)|Y]\|_2.$$

Squaring both sides, we arrive at (20). $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Appendix B   Proofs from Section 3

Lemma 5

Proof. Let $Y^n = X^n \oplus Z^n$ for some Z^n distributed over $\{-1,1\}^n$ and independent of X^n. Thus

$$\begin{aligned}
\mathbb{E}\left[\chi_{\mathcal{S}}(Y^n)|X^n\right] &= \mathbb{E}\left[\chi_{\mathcal{S}}(Z^n \oplus X^n) \mid X^n\right] \\
&= \mathbb{E}\left[\chi_{\mathcal{S}}(X^n)\chi_{\mathcal{S}}(Z^n) \mid X^n\right] \\
&= \chi_{\mathcal{S}}(X^n)\mathbb{E}\left[\chi_{\mathcal{S}}(Z^n)\right],
\end{aligned}$$

where the last equality follows from the assumption that $X^n \perp\!\!\!\perp Z^n$. By letting $c_{\mathcal{S}} = \mathbb{E}\left[\chi_{\mathcal{S}}(Z^n)\right]$, it follows that $p_{Y^n|X^n} \in \mathcal{A}_n$ and, consequently, $\mathcal{B}_n \subseteq \mathcal{A}_n$.

Now let y_n be fixed and $\delta_{y^n} : \{-1,1\}^n \to \{0,1\}$ be given by

$$\delta_{y^n}(x^n) = \begin{cases} 1, & x^n = y^n, \\ 0, & \text{otherwise.} \end{cases}$$

Since the function δ_{y^n} has Boolean inputs, it can be expressed in terms of its Fourier expansion [69, Prop. 1.1] as

$$\delta_{y^n}(x^n) = \sum_{\mathcal{S} \subseteq [n]} \widehat{d}_{\mathcal{S}}\chi_{\mathcal{S}}(x^n)$$

for some set of coefficients $\widehat{d}_{\mathcal{S}} \in \mathbb{R}$, $\mathcal{S} \subseteq [n]$. Now let $p_{Y^n|X^n} \in \mathcal{A}_n$. Observe that $p_{Y^n|X^n}(y^n|x^n) = \mathbb{E}\left[\delta_{y^n}(Y^n) \mid X^n = x^n\right]$ and, for $z^n \in \{-1,1\}^n$,

$$\begin{aligned}
p_{Y^n|X^n}(y^n \oplus z^n|x^n \oplus z^n) &= \mathbb{E}\left[\delta_{y^n \oplus z^n}(Y^n) \mid X^n = x^n \oplus z^n\right] \\
&= \mathbb{E}\left[\delta_{y^n}(Y^n \oplus z^n) \mid X^n = x^n \oplus z^n\right] \\
&= \mathbb{E}\left[\sum_{\mathcal{S} \subseteq [n]} \widehat{d}_{\mathcal{S}}\chi_{\mathcal{S}}(Y^n \oplus z^n) \mid X^n = x^n \oplus z^n\right] \\
&= \mathbb{E}\left[\sum_{\mathcal{S} \subseteq [n]} \widehat{d}_{\mathcal{S}}\chi_{\mathcal{S}}(Y^n)\chi_{\mathcal{S}}(z^n) \mid X^n = x^n \oplus z^n\right] \\
&\stackrel{(a)}{=} \sum_{\mathcal{S} \subseteq [n]} c_{\mathcal{S}}\widehat{d}_{\mathcal{S}}\chi_{\mathcal{S}}(x^n \oplus z^n)\chi_{\mathcal{S}}(z^n) \\
&= \sum_{\mathcal{S} \subseteq [n]} c_{\mathcal{S}}\widehat{d}_{\mathcal{S}}\chi_{\mathcal{S}}(x^n) \\
&\stackrel{(b)}{=} \mathbb{E}\left[\sum_{\mathcal{S} \subseteq [n]} \widehat{d}_{\mathcal{S}}\chi_{\mathcal{S}}(Y^n)|X^n = x^n\right] \\
&= \mathbb{E}\left[\delta_{y^n}(Y^n) \mid X^n = x^n\right] \\
&= p_{Y^n|X^n}(y^n|x^n).
\end{aligned}$$

Equalities (a) and (b) follow from the definition of $\mathcal{A}_n$. By defining the distribution of Z^n as $p_{Z^n}(z^n) \triangleq p_{Y^n|X^n}(z^n|\mathbf{1}^n)$, where $\mathbf{1}^n$ is the vector with all entries equal to 1, it follows that $Z^n = X^n \oplus Y^n$, $Z^n \perp\!\!\!\perp X^n$ and $p_{Y^n|X^n} \subseteq \mathcal{B}_n$. $\qquad\square$

Lemma 7

Proof. Let $\mathbf{x} \in \mathcal{C}^m(a, \mathbf{P}^T)$ and $\mathbf{y} \in \mathcal{C}^n(b, \mathbf{P})$. Then, for $\mathbf{P}$ decomposed as $\mathbf{P} = \mathbf{D}_X^{1/2}\mathbf{Q}\mathbf{D}_Y^{1/2}$ where $\mathbf{Q}$ given in (14) and denoting $\mathbf{\Sigma}^- = \text{diag}(0, \sigma_1, \ldots, \sigma_d)$,

$$\begin{aligned}
\mathbf{x}^T\mathbf{P}\mathbf{y} &= ab + \mathbf{x}^T\mathbf{D}_X^{1/2}\mathbf{U}\mathbf{\Sigma}^-\mathbf{V}^T\mathbf{D}_Y^{1/2}\mathbf{y} \\
&= ab + \hat{\mathbf{x}}^T\mathbf{\Sigma}^-\hat{\mathbf{y}}, \tag{77}
\end{aligned}$$

32

where $\hat{\mathbf{x}} \triangleq \mathbf{U}^T \mathbf{D}_X^{1/2} \mathbf{x}$ and $\hat{\mathbf{y}} \triangleq \mathbf{V}^T \mathbf{D}_Y^{1/2} \mathbf{y}$. Since $\hat{x}_1 = \|\hat{\mathbf{x}}\|_2 = a$ and $\hat{y}_1 = \|\hat{\mathbf{y}}\|_2 = b$, then

$$
\begin{aligned}
\hat{\mathbf{x}}^T \boldsymbol{\Sigma}^- \hat{\mathbf{y}} &= \sum_{i=2}^{d+1} \sigma_{i-1} \hat{\mathbf{x}}_i \hat{\mathbf{y}}_i \\
&\leq \sigma_1 \sqrt{\left(\|\hat{\mathbf{x}}\|_2^2 - \hat{\mathbf{x}}_1^2\right)\left(\|\hat{\mathbf{y}}\|_2^2 - \hat{\mathbf{y}}_1^2\right)} \\
&= \sigma_1 \sqrt{(a - a^2)(b - b^2)}.
\end{aligned}
$$

The result follows by noting that $\sigma_1 = \rho_m(X; Y)$. $\qquad\square$

Appendix C   Proofs from Section 4

Theorem 6

Consider the matrix $\mathbf{Q} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ given in (14), and define

$$
\widetilde{\mathbf{A}} \triangleq \mathbf{D}_X^{1/2}\mathbf{U}, \ \widetilde{\mathbf{B}} \triangleq \mathbf{D}_Y^{1/2}\mathbf{V}.
$$

Then

$$
\mathbf{P} = \widetilde{\mathbf{A}}\boldsymbol{\Sigma}\widetilde{\mathbf{B}}^T, \tag{78}
$$

where $\widetilde{\mathbf{A}}^T \mathbf{D}_X^{-1}\widetilde{\mathbf{A}} = \widetilde{\mathbf{B}}^T \mathbf{D}_Y^{-1}\widetilde{\mathbf{B}} = \mathbf{I}$.

It follows directly from Theorem 1 that $\widetilde{\mathbf{A}}$, $\widetilde{\mathbf{B}}$ and $\boldsymbol{\Sigma}$ have the form

$$
\widetilde{\mathbf{A}} = [\mathbf{p}_X \ \ \mathbf{A}], \ \widetilde{\mathbf{B}} = [\mathbf{p}_Y \ \ \mathbf{B}], \ \boldsymbol{\Sigma} = \operatorname{diag}\left(1, \sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d}\right), \tag{79}
$$

and, consequently, the joint distribution can be written as

$$
p_{X,Y}(x, y) = p_X(x)p_Y(y) + \sum_{k=1}^d \sqrt{\lambda_k} b_{y,k} a_{x,k}, \tag{80}
$$

where $a_{x,k}$ and $b_{y,k}$ are the entries of $\mathbf{A}$ and $\mathbf{B}$ in (79), respectively.

Theorem 6 follows directly from the next two lemmas.

Lemma 16. *Let the marginal distribution $\mathbf{p}_X$ and the PICs $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ 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 $0 \leq \beta \leq p_X(2)$*

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

where

$$
\begin{aligned}
f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) &= \sum_{i=2}^{d+1} p_X(i)(\lambda_{i-1} + c_i - c_{i-1}) \\
&\quad + p_X(1)(c_1 + \alpha) - \alpha \mathbf{p}_X^T \mathbf{p}_X , \tag{81}
\end{aligned}
$$

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 $\mathbf{P}$ 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 $\mathbf{F}$, where the (i, j)-th entry is the probability $p_{\hat{X}|Y}(j|i)$. The probability of correct estimation P_c is then

$$
P_c = 1 - P_e(X|Y) = \operatorname{tr}\left(\mathbf{P}_{X,\hat{X}}\right),
$$

where $\mathbf{P}_{X,\hat{X}} \triangleq \mathbf{PF}$.

The matrix $\mathbf{P}_{X,\hat{X}}$ can be decomposed according to (78), resulting in

$$P_c = \operatorname{tr}\left(\mathbf{D}_X^{1/2}\mathbf{U}\widetilde{\boldsymbol{\Sigma}}\mathbf{V}^T\mathbf{D}_{\hat{X}}^{1/2}\right) = \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}\mathbf{V}^T\mathbf{D}_{\hat{X}}^{1/2}\mathbf{D}_X^{1/2}\mathbf{U}\right), \tag{82}$$

where

$$\mathbf{U} = \begin{bmatrix} \mathbf{p}_X^{1/2} & \mathbf{u}_2 & \cdots & \mathbf{u}_m \end{bmatrix},$$

$$\mathbf{V} = \begin{bmatrix} \mathbf{p}_{\hat{X}}^{1/2} & \mathbf{v}_2 & \cdots & \mathbf{v}_m \end{bmatrix},$$

$$\widetilde{\boldsymbol{\Sigma}} = \operatorname{diag}\left(1, \sqrt{\widetilde{\lambda}_1}, \ldots, \sqrt{\widetilde{\lambda}_d}\right),$$

$$\mathbf{D}_{\hat{X}} = \operatorname{diag}\left(\mathbf{p}_{\hat{X}}\right),$$

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

$$P_c = \mathbf{p}_X^T\mathbf{p}_{\hat{X}} + \sum_{k=2}^{m}\sum_{i=1}^{m}\left(\widetilde{\lambda}_{k-1}p_X(i)p_{\hat{X}}(i)\right)^{1/2}u_{k,i}v_{k,i}$$

$$= \mathbf{p}_X^T\mathbf{p}_{\hat{X}} + \sum_{k=2}^{m}\sum_{i=1}^{m}\widetilde{\lambda}_{k-1}^{1/2}\widetilde{u}_{k,i}\widetilde{v}_{k,i}$$

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

$$P_c \leq \mathbf{p}_X^T\mathbf{p}_{\hat{X}} + \sum_{i=1}^{m}\left(\sum_{k=2}^{m}\widetilde{v}_{k,i}^2\right)^{1/2}\left(\sum_{k=2}^{m}\widetilde{\lambda}_{k-1}\widetilde{u}_{k,i}^2\right)^{1/2}$$

$$= \mathbf{p}_X^T\mathbf{p}_{\hat{X}} + \sum_{i=1}^{m}\left(p_{\hat{X}}(i)(1 - p_{\hat{X}}(i))\sum_{k=2}^{m}\widetilde{\lambda}_{k-1}\widetilde{u}_{k,i}^2\right)^{1/2}$$

$$\leq \mathbf{p}_X^T\mathbf{p}_{\hat{X}} + \left(1 - \sum_{i=1}^{m}p_{\hat{X}}(i)^2\right)^{1/2}\left(\sum_{i=1}^{m}\sum_{k=2}^{m}\widetilde{\lambda}_{k-1}\widetilde{u}_{k,i}^2\right)^{1/2}. \tag{83}$$

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

$$\sum_{i=1}^{m}\sum_{k=2}^{m}\widetilde{\lambda}_{k-1}\widetilde{u}_{k,i}^2 = \operatorname{tr}\left(\boldsymbol{\Sigma}\overline{\mathbf{U}}^T\mathbf{D}_X\overline{\mathbf{U}}\right)$$

$$\leq \sum_{k=1}^{d}\sigma_k\widetilde{\lambda}_k,$$

$$\leq \sum_{k=1}^{d}\sigma_k\lambda_k. \tag{84}$$

where σ_k are the singular values of $\overline{\mathbf{U}}^T\mathbf{D}_X\overline{\mathbf{U}}$. The first inequality follows from the application of Von-Neumman's trace inequality [24, Thm. 7.4.1.1] and the fact that $\overline{\mathbf{U}}^T\mathbf{D}_X\overline{\mathbf{U}}$ is symmetric and positive semi-definite. The second inequality follows by observing that the PICs satisfy the DPI and, therefore, $\widetilde{\lambda}_k \leq \lambda_k$.

We will now find an upper bound for (84) by bounding the eigenvalues σ_k. First, note that $\overline{\mathbf{U}}\,\overline{\mathbf{U}}^T = \mathbf{I} - \mathbf{p}_X^{1/2}\left(\mathbf{p}_X^{1/2}\right)^T$ and consequently

$$\sum_{k=1}^{d}\sigma_k = \operatorname{tr}\left(\overline{\mathbf{U}}^T\mathbf{D}_X\overline{\mathbf{U}}\right)$$

$$= \operatorname{tr}\left(\mathbf{D}_X\left(\mathbf{I} - \mathbf{p}_X^{1/2}\left(\mathbf{p}_X^{1/2}\right)^T\right)\right)$$

$$= 1 - \sum_{i=1}^{m}p_X(i)^2. \tag{85}$$

Second, note that $\overline{\mathbf{U}}^T \mathbf{D}_X \overline{\mathbf{U}}$ is a principal submatrix of $\mathbf{U}^T \mathbf{D}_X \mathbf{U}$, formed by removing the first row and columns of $\mathbf{U}^T \mathbf{D}_X \mathbf{U}$. It then follows from Cauchy's interlacing theorem [24, Theorem 4.3.17] that

$$p_X(m) \leq \sigma_{m-1} \leq p_X(m-1) \leq \ldots \leq p_X(2) \leq \sigma_1 \leq p_X(1). \tag{86}$$

Combining (85) and (86), an upper bound for (84) can be found by solving the following linear program

$$\max_{s_i} \quad \sum_{i=1}^{d} \lambda_i s_i \tag{87}$$

$$\text{subject to} \quad \sum_{i=1}^{d} s_i = 1 - \mathbf{p}_X^T \mathbf{p}_X,$$

$$p_X(i+1) \leq s_i \leq p_X(i), \ i = 1, \ldots, d \ .$$

Let $\delta_i \triangleq p_X(i) - p_X(i+1)$ and $\gamma_i \triangleq \lambda_i p_X(i+1)$. The dual of (87) is

$$\min_{y_i, \mu} \quad \mu \left(p_X(1) - \mathbf{p}_X^T \mathbf{p}_X \right) + \sum_{i=1}^{m-1} \delta_i y_i + \gamma_i \tag{88}$$

$$\text{subject to} \quad y_i \geq [\lambda_i - \mu]^+ , \ i = 1, \ldots, d \ .$$

For any given value of μ, the optimal values of the dual variables y_i in (88) are

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

Therefore the linear program (88) is equivalent to

$$\min_{\mu} f_0(\mu, \mathbf{p}_X, \boldsymbol{\lambda}), \tag{89}$$

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

Denote the solution of (87) by $f_P^*(\mathbf{p}_X, \boldsymbol{\lambda})$ and of (88) by $f_D^*(\mathbf{p}_X, \boldsymbol{\lambda})$. It follows that (84) can be bounded

$$\sum_{k=1}^{d} \sigma_k \lambda_k \leq f_P^*(\mathbf{p}_X, \boldsymbol{\lambda}) = f_D^*(\mathbf{p}_X, \boldsymbol{\lambda}) \leq f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) \ \forall \ \alpha \in \mathbb{R}. \tag{90}$$

We may consider $0 \leq \alpha \leq 1$ in (90) without loss of generality.

Using (90) to bound (83), we find

$$P_c \leq \mathbf{p}_X^T \mathbf{p}_{\hat{X}} + \left[f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) \left(1 - \sum_{i=1}^{m} p_{\hat{X}}(i)^2 \right) \right]^{1/2} \tag{91}$$

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

$$\max_{x_i} \quad \left[f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) \left(1 - \sum_{i=1}^{m} x_i^2 \right) \right]^{1/2} + \sum_{i=1}^{m} p_X(i) x_i \tag{92}$$

$$\text{subject to} \quad \sum_{i=1}^{m} x_i = 1,$$

$$x_i \geq 0, i = 1, \ldots, m \ .$$

The dual function of (92) over the constraint $\sum_{i=1}^{m} x_i = 1$ is

$$L(\beta) = \max_{x_i \geq 0} \quad \beta + \left[f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) \left(1 - \sum_{i=1}^{m} x_i^2 \right) \right]^{1/2}$$

$$+ \sum_{i=1}^{m} (p_X(i) - \beta) x_i$$

$$= \beta + \sqrt{f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) + \sum_{i=1}^{m} \left([p_X(i) - \beta]^+\right)^2}. \tag{93}$$

Since $L(\beta)$ is an upper bound of (92) for any β and, therefore, is also an upper bound of (91), then

$$P_c \leq \beta + \sqrt{f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) + \sum_{i=1}^{m} \left([p_X(i) - \beta]^+\right)^2}. \tag{94}$$

Note that we can consider $0 \leq \beta \leq p_X(2)$ in (94), since $L(\beta)$ is increasing for $\beta > p_X(2)$. Taking $P_e(X|Y) = 1 - P_c$, the result follows. □

The next result tightens the bound introduced in Lemma 16 by optimizing over all values of α.

Lemma 17. *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* (48). *Then*

$$f_0^*(\mathbf{p}_X, \boldsymbol{\lambda}) = \sum_{i=1}^{k^*} \lambda_i p_X(i) + \sum_{i=k^*+1}^{m} \lambda_{i-1} p_X(i) - \lambda_{k^*} \mathbf{p}_X^T \mathbf{p}_X , \tag{95}$$

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 for ease of notation $\lambda_0 \triangleq 1$ and $\lambda_m \triangleq 0$ (recall that the PICs correspond to $\lambda_1, \ldots, \lambda_{m-1}$). Then $c_i = \lambda_i - \alpha$ for $1 \leq i \leq k-1$ and $c_i = 0$ for $k \leq i \leq d$ in (81). Therefore

$$f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) = \sum_{i=1}^{k-1} \lambda_i p_X(i) + \alpha p_X(k) + \sum_{i=k+1}^{m} \lambda_{i-1} p_X(i) - \alpha \mathbf{p}_X^T \mathbf{p}_X. \tag{96}$$

Note that (96) 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 $\alpha = 0$. Theorem 6 follows directly. □

Theorem 7

Proof. Consider two probability distributions p_X and q_X defined over $\mathcal{X} = \{1, \ldots, m\}$, and 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 [79], and can be written as $q_X = \sum_{i=1}^{l} a_i(p_X \circ \pi_i)$ for some $l \geq 1$, where $a_i \geq 0$, $\sum a_i = 1$ and π_i is a permutation of p_X, i.e. $p_X \circ \pi_i = p_{\pi_i(X)}$. Hence, for a fixed $p_{\hat{X}|X}$:

$$\mathcal{I}(q_X, p_{\hat{X}|X}) = \mathcal{I}\left(\sum_{i=1}^{l} a_i(p_X \circ \pi_i), p_{\hat{X}|X}\right)$$

$$\geq \sum_{i=1}^{l} a_i \mathcal{I}(p_X \circ \pi_i, p_{\hat{X}|X}),$$

$$= \sum_{i=1}^{l} a_i \mathcal{I}(p_X, \pi_i \circ p_{\hat{X}|X}), \tag{97}$$

where the inequality follows from the concavity assumption and the last equality from $\mathcal{I}(X; \hat{X})$ being invariant to one-to-one mappings of X and $\hat{X}$. Consequently, from the definition of error-rate function in Defn. 10,

$$e_{\mathcal{I}}(q_X, \theta) = \inf_{p_{\hat{X}|X}} \left\{ \sum_{x, x' \in [m]} d_H(x, x') q_X(x) p_{\hat{X}|X}(x'|x) \,\middle|\, \mathcal{I}(q_X, p_{\hat{X}|X}) \leq \theta \right\}$$

$$\overset{(a)}{\geq} \inf_{p_{\hat{X}|X}} \left\{ \sum_{i \in [l]} a_i \sum_{x, x' \in [m]} d_H(\pi_i(x), x') p_X(x) p_{\hat{X}|X}(x'|\pi_i(x)) \,\middle|\, \sum_{i \in [l]} a_i \mathcal{I}(p_X, \pi_i \circ p_{\hat{X}|X}) \leq \theta \right\}$$

36

$$\stackrel{(b)}{=} \inf_{p_{\hat{X}|X}} \left\{ \sum_{i\in[l]} a_i \sum_{x,x'\in[m]} d_H(\pi_i(x),\pi_i(x'))p_X(x)p_{\hat{X}|X}(\pi_i(x')|\pi_i(x)) \right.$$

$$\left. \left| \sum_{i\in[l]} a_i \mathcal{I}(p_X,\pi_i \circ p_{\hat{X}|X} \circ \pi_i) \le \theta \right. \right\}$$

$$\stackrel{(c)}{\ge} \inf_{p^1_{\hat{X}|X},\dots,p^l_{\hat{X}|X}} \left\{ \sum_{i\in[l]} a_i \sum_{x,x'\in[m]} d_H(x,x')p_X(x)p^i_{\hat{X}|X}(x|x') \left| \sum_{i\in[l]} a_i \mathcal{I}(p_X,p^i_{\hat{X}|X}) \le \theta \right. \right\}$$

$$\stackrel{(d)}{=} \inf_{\theta_1,\dots,\theta_l \ge 0} \left\{ \sum_{i=1}^{l} a_i e_{\mathcal{I}}(p_X,\theta_i) \left| \sum_{i=1}^{l} a_i\theta_i = \theta \right. \right\}$$

$$\stackrel{(e)}{\ge} \inf_{\theta_1,\dots,\theta_l \ge 0} \left\{ e_{\mathcal{I}}\left(p_X,\sum a_i\theta_i\right) \left| \sum_{i=1}^{l} a_i\theta_i = \theta \right. \right\}$$

$$= e_{\mathcal{I}}(p_X,\theta),$$

where inequality (a) follows from (97), (b) follows from the fact that the infimum is taken over all mapping $p_{\hat{X}|X}$ and that $\mathcal{I}(X;\hat{X})$ is invariant to one-to-one mappings of X and $\hat{X}$, (c) follows by allowing a mapping $p^i_{\hat{X}|X}$ to be independently minimized for each i (as opposed to minimizing the same mapping $p_{\hat{X}|X}$ for all i), (d) is obtained by noting that the optimal choice of $p^i_{\hat{X}|X}$ is the one that minimizes the Hamming distortion d_H for a given upperbound on $\mathcal{I}(p_X,p^i_{\hat{X}|X})$, and (e) follows from the convexity of $e_I(p_X,\theta)$ in θ. Since this holds for any q_X that is majorized by p_X, $e_{\mathcal{I}}(p_X,\theta)$ is Schur-concave. $\quad\square$

Appendix D   Proofs from Section 5

Lemma 10

Proof. Let $p_{S,X}$ and $p_{Y|X}$ be given, with $S \to X \to Y$. Denote by $\mathbf{w}_i$ the vector in the $|\mathcal{X}|$-simplex with entries $p_{X|Y}(\cdot|i)$. Furthermore, let $a_i \triangleq H(X) - H(X|Y=i)$, and $b_i \triangleq H(S) - H(S|Y=i)$. Therefore

$$\sum_{i=1}^{|\mathcal{Y}|} p_Y(i)\left[\mathbf{w}_i, a_i, b_i\right] = [\mathbf{p}_X, I(X;Y), I(S;Y)]. \tag{98}$$

Since $\mathbf{w}_i$ belongs to the $|\mathcal{X}|$-simplex, the vector $[\mathbf{w}_i, a_i, b_i]$ is taken from a connected, compact $|\mathcal{X}|+1$ dimensional space. Then, from Fenchel-Eggleston strengthening of Carathéodory's theorem [80, Theorem 18, pg. 35], the point $[\mathbf{p}_X, I(X;Y), I(S;Y)]$ can also be achieved by at most $|\mathcal{X}|+1$ non-zero values of $p_Y(i)$. It follows directly that it is sufficient to consider $|\mathcal{Y}| \le |\mathcal{X}|+1$ for the mappings that approach the infimum $G_I(t,p_{S,X})$ in (59). The set of all mappings $p_{Y|X}$ for $|\mathcal{Y}| \le |\mathcal{X}|+1$ is compact, and both $p_{Y|X} \to I(S;Y)$ and $p_{Y|X} \to I(X;Y)$ are continuous and bounded when S, X and Y have finite support. Consequently, the infimum in (59) is attainable.

For $0 < t \le H(X)$ and $p_{S,X}$ fixed, let $G_I(t,p_{S,X}) = \alpha$. From the discussion above, there exists $p_{Y|X}$ that achieves $I(S;Y) = \alpha$ for $I(X;Y) \ge t$. Now consider $p_{\widetilde{Y}|X}$ where $\widetilde{\mathcal{Y}} = [|\mathcal{Y}|+1]$ and, for $0 < \lambda \le 1$,

$$p_{\widetilde{Y}|X}(y|x) = (1-\lambda)\mathbf{1}_{\{y=|\mathcal{Y}|+1\}} + \lambda\mathbf{1}_{\{y \ne |\mathcal{Y}|+1\}}p_{Y|X}(y|x).$$

Note that $\widetilde{Y}$ can be understood as an erased version of Y, with the erasure symbol being $|\mathcal{Y}|+1$. It follows (cf. [9, Sec. 7.1.5]) that $I(S;\widetilde{Y}) = \lambda I(S;Y) = \lambda\alpha$ and $I(X;\widetilde{Y}) = \lambda I(X;Y) \ge \lambda t$. We have thus explicitly constructed a new mapping $p_{\widetilde{Y}|X}$ that satisfies $S \to X \to \widetilde{Y}$ and achieves $I(S;\widetilde{Y}) = \lambda\alpha$ and $I(X;\widetilde{Y}) \ge \lambda t$. Therefore, from the definition of G_I in (59), $G_I(\lambda t, p_{S,X}) \le \lambda\alpha = \lambda I(S;Y)$. Consequently,

$$\frac{G_I(\lambda t, p_{S,X})}{\lambda t} \le \frac{\lambda I(S;Y)}{\lambda t} = \frac{G_I(t,p_{S,X})}{t}. \tag{99}$$

Since this holds for any $0 < \lambda \le 1$, then $\frac{G_I(t,p_{S,X})}{t}$ is non-decreasing in t. Finally, for a fixed $p_{S,X}$, the set of points $(\mathbf{w}_i, a_i, b_i) \in \mathbb{R}^{|\mathcal{X}|+2}$ that satisfies (98) is convex, and thus, for a fixed $\mathbf{p}_X$, it's lower-boundary, which corresponds to the graph of $(t, G_I(t,p_{S,X}))$, is convex. $\quad\square$

Lemma 12

Proof. For fixed $p_{Y|X}$ and $p_{S,X}$, and assuming $I(X;Y) > 0$,

$$\frac{I(S;Y)}{I(X;Y)} = \frac{\sum_{y \in \mathcal{Y}} p_Y(y) D(p_{S|Y=y}||p_S)}{\sum_{y \in \mathcal{Y}} p_Y(y) D(p_{X|Y=y}||p_X)}$$

$$\geq \min_{\substack{y \in \mathcal{Y}: \\ D(p_{X|Y=y}||p_X) > 0}} \frac{D(p_{S|Y=y}||p_S)}{D(p_{X|Y=y}||p_X)}$$

$$\geq \inf_{q_X \neq p_X} \frac{D(q_S||p_S)}{D(q_X||p_X)}.$$

Now let d^* be the infimum in the right-hand side of (64), and q_X satisfy

$$\frac{D(q_Y||p_Y)}{D(q_X||p_X)} = d^* + \delta,$$

where $\delta > 0$. For $\epsilon > 0$ and sufficiently small, let $p_{Y|X}$ be such that $\mathcal{Y} = [2]$, $p_Y(1) = \epsilon$, $p_{X|Y}(x|1) = q_X(x)$ and

$$p_{X|Y}(x|2) = \frac{1}{1-\epsilon} p_X(x) - \frac{\epsilon}{1-\epsilon} q_X(x).$$

Since for any distribution r_X with support $\mathcal{X}$ we have $D((1-\epsilon)p_X + \epsilon r_X || p_X) = o(\epsilon)$, we find

$$I(S;Y) = \epsilon D(p_{S|Y=1}||p_S) + (1-\epsilon) D(p_{S|Y=0}||p_S)$$
$$= \epsilon D(q_S||p_S) + o(\epsilon),$$

and equivalently, $I(X;Y) = \epsilon D(q_X||p_X) + o(\epsilon)$. Consequently,

$$\frac{I(S;Y)}{I(X;Y)} = \frac{\epsilon D(q_S||p_S) + o(\epsilon)}{\epsilon D(q_X||p_X) + o(\epsilon)} \to d^* + \delta,$$

where the limit is taken as $\epsilon \to 0$. Since this holds for any $\delta > 0$, then $v^*(p_{S,X}) \leq d^*$, proving the result. $\square$

Lemma 13

Proof. Let $f : \mathcal{X} \to \mathbb{R}$, $\mathbb{E}[f(X)] = 0$ and $\|f(X)\|_2^2 = 1$, and $\mathbf{f} \in \mathbb{R}^{|\mathcal{X}|}$ be a vector with entries $f_i = f(i)$ for $i \in \mathcal{X}$. Observe that

$$\|\mathbb{E}[f(X)|S]\|_2^2 = \sum_{s \in \mathcal{S}} p_S(s) \mathbb{E}[f(X)|S=s]^2$$
$$= \mathbf{f}^T \mathbf{P}_{X|S}^T \mathbf{D}_S \mathbf{P}_{X|S} \mathbf{f}^T$$
$$= \mathbf{f}^T \mathbf{D}_X^{1/2} \mathbf{Q}^T \mathbf{Q} \mathbf{D}_X^{1/2} \mathbf{f}$$
$$\geq \delta(p_{S,X}),$$

where the last inequality follows by noting that $\mathbf{x} \triangleq \mathbf{f}^T \mathbf{D}_X^{1/2}$ satisfies $\|\mathbf{x}\|_2 = 1$ and that $\delta(p_{S,X})$ is the smallest eigenvalue of the positive semi-definite matrix $\mathbf{Q}^T \mathbf{Q}$, where $\mathbf{Q}$ was defined in Definition 1 as $\mathbf{Q} \triangleq \mathbf{D}_S^{-1/2} \mathbf{P}_{X,S} \mathbf{D}_X^{-1/2}$. $\square$

Lemma 14

Proof. Let $p_{S|X}$ be fixed, and define

$$g_\lambda(p_X) \triangleq H(S) - \lambda H(X),$$

where $H(S)$ and $H(X)$ are the entropy of S and X, respectively, when $(S, X) \sim p_{S|X} p_X$. For $0 < \epsilon \ll 1$, let

$$p_\epsilon(i) \triangleq p_X(i)(1 + \epsilon f(i))$$

be a perturbed version of p_X, where $\mathbb{E}[f(X)] = 0$ and, w.l.o.g., $\|f(X)\|_2 = 1$. The second derivative of $g_\lambda(p_\epsilon)$ at $\epsilon = 0$ is[9]

$$\left.\frac{\partial^2 g_\lambda(p_\epsilon)}{\partial \epsilon^2}\right|_{\epsilon=0} = \log_2(e)\left(-\|\mathbb{E}[f(X)|S]\|_2^2 + \lambda\|f(X)\|_2^2\right)$$
$$= \log_2(e)\left(-\|\mathbb{E}[f(X)|S]\|_2^2 + \lambda\right). \tag{100}$$

Thus, from Lemma 13, if $\lambda \leq \delta(p_{S,X})$ then for any sufficiently small perturbation of p_X, (100) will be non-positive. Conversely, if $\lambda > \delta(p_{S,X})$, then we can find a perturbation $f(X)$ such that (100) is positive. Therefore, $g_\lambda(p_X)$ has a negative semi-definite Hessian if and only if $0 \leq \lambda \leq \delta(p_{S,X})$.

For any $S \to X \to Y$, we have $I(S;Y)/I(X;Y) \geq v^*(p_{S,X})$, and, consequently, for $0 \leq \lambda^\dagger \leq v^*(p_{S,X})$,

$$g_{\lambda\dagger}(p_X) \geq H(S|Y) - \lambda^\dagger H(X|Y),$$

and $g_{\lambda\dagger}(p_X)$ touches the upper-concave envelope of $g_{\lambda\dagger}$ at p_X. Since a function must be concave at the points where it matches its concave envelope, $g_{\lambda\dagger}$ has a negative semi-definite Hessian at p_X and, from (100), $\lambda^\dagger \leq \delta(p_{S,X})$. Since this holds for any $0 \leq \lambda^\dagger \leq v^*(p_{S,X})$, we find $v^*(p_{S,X}) \leq \delta(p_{S,X})$.

For a fixed $p_{S|X}$, the function $g_\lambda(p_X)$ is concave when $\lambda = 0$ and convex when $\lambda = 1$. Consequently, the maximum λ for which $g_\lambda(p_X)$ has a negative Hessian at p_X is $\delta(p_{S,X})$. Furthermore, Lemma 12 implies that a value λ_1 for which $g_\lambda(p_X)$ touches its lower concave envelope at p_X for all $\lambda_1 \geq \lambda$ is $v^*(p_{S,X})$. Therefore, both $\inf_{p_X} v^*(p_{S,X})$ and $\inf_{p_X} \delta(p_{S,X})$ equal the maximum value of λ such that the function $g_\lambda(p_X)$ is concave at all values of p_X. Therefore, we established that for a given $p_{S|X}$,

$$\inf_{p_X} v^*(p_{S,X}) = \inf_{p_X} \delta(p_{S,X}).$$

$\square$

Lemma 15

Proof. Theorem 14 immediately gives $\delta(p_{S,X}) = 0 \Rightarrow v^*(p_{S,X}) = 0$. Let $v^*(p_{S,X}) = 0$. Then, since $D(q_X\|p_X) \leq -\min_{i\in\mathcal{X}} \log_2 p_X(i)$ and $\mathcal{X}$ is finite, Lemma 12 implies that for any $\epsilon > 0$ there exists q_X and $0 < \delta \leq -\min_{i\in\mathcal{X}} \log_2 p_X(i)$ such that

$$D(q_X\|p_X) \geq \delta > 0$$

and

$$D(q_S\|p_S) < \epsilon.$$

We can then construct a sequence $q_X^1, q_X^2, q_X^3, \ldots$ such that $q_X^i \neq p_X$, $D(q_S^k\|p_S) \leq \epsilon_k$ and

$$\lim_{k\to\infty} \epsilon_k = 0.$$

Let $\mathbf{q}_S^k$ be a vector whose entries are $q_S^k(\cdot)$. Then, from Pinsker's inequality,

$$\epsilon_k \geq \frac{1}{2}\|\mathbf{q}_S^k - \mathbf{p}_S\|_1^2 \geq \frac{1}{2}\|\mathbf{q}_S^k - \mathbf{p}_S\|_2^2. \tag{101}$$

Defining $\mathbf{x}^k = \mathbf{q}_X^k - \mathbf{p}_X$, observe that $0 < \|\mathbf{x}^k\|_2^2 \leq 2$ and, from (101), $\|\mathbf{P}_{S|X}\mathbf{x}^k\|_2 \leq \sqrt{2\epsilon_k}$. Hence,

$$\lim_{k\to\infty} \frac{\|\mathbf{P}_{S|X}\mathbf{x}^k\|_2^2}{\|\mathbf{x}^k\|_2^2} = 0. \tag{102}$$

In addition, denoting $s_m \triangleq \min_{s\in\mathcal{S}} p_S(s)$ and $x_M \triangleq \min_{x\in\mathcal{X}} p_X(x)$, for each k we have

$$\frac{\|\mathbf{P}_{S|X}\mathbf{x}^k\|_2^2}{\|\mathbf{x}^k\|_2^2} \geq \min_{\|\mathbf{y}\|_2^2>0} \frac{\|\mathbf{P}_{S|X}\mathbf{y}\|_2^2}{\|\mathbf{y}\|_2^2}$$
$$= \min_{\|\mathbf{y}\|_2^2>0} \frac{\|\mathbf{P}_{S,X}\mathbf{D}_X^{-1/2}\mathbf{y}\|_2^2}{\|\mathbf{D}_X^{1/2}\mathbf{y}\|_2^2} \tag{103}$$
$$\geq \min_{\|\mathbf{y}\|_2^2>0} \frac{s_m\|\mathbf{D}_S^{-1/2}\mathbf{P}_{S,X}\mathbf{D}_X^{-1/2}\mathbf{y}\|_2^2}{x_M\|\mathbf{y}\|_2^2} \tag{104}$$

---

[9]This was observed in [23] and [77], and follows directly from $-\frac{\partial^2}{\partial\epsilon^2}a(1+b\epsilon)\log_2 a(1+b\epsilon) = -b^2 a\log_2(e)$.

$$= \frac{s_m}{x_M} \min_{\|\mathbf{y}\|_2^2 > 0} \frac{\|\mathbf{Q}\mathbf{y}\|_2^2}{\|\mathbf{y}\|_2^2} \tag{105}$$

$$= \frac{s_m \delta(p_{S,X})}{x_M}. \tag{106}$$

In the derivation above, (103) follows from $\mathbf{D}_X$ being invertible (by definition), (104) is a direct consequence of $\|\mathbf{D}_S^{-1/2}\mathbf{y}\|_2^2 \leq s_m^{-1}\|\mathbf{y}\|_2^2$ and $\|\mathbf{D}_X^{1/2}\mathbf{y}\|_2^2 \leq x_M\|\mathbf{y}\|_2^2$ for any $\mathbf{y}$, and (105) and (106) follow from the definition of $\mathbf{Q}$ and $\delta(p_{S,X})$, respectively. Combining (106) with (102), it follows that $\delta(p_{S,X}) = 0$, proving the desired result. $\square$

Corollary 8

Proof. If $\delta(p_{S,X}) = 0$, then the lower bound for t^* follows from the construction used in (71) and, more specifically, by (i) maximizing the right-hand side of (71) across all functions in $\mathcal{F}_0$ and (ii) observing that the maximum value of ϵ such that $p_{Y|X}$ is non-negative is $\epsilon = 1/2\|f\|_\infty$. If $\delta(p_{S,X}) > 0$, then $\mathcal{F}_0$ is singular (i.e. $\mathcal{F}_0 = \{w_0\}$), and the lower bound (72) reduces to the trivial bound $t^* \geq 0$.

In order to prove that the lower bound is sharp, consider S being an unbiased bit, drawn from $\{1, 2\}$, and X the result of sending S through an erasure channel with erasure probability $1/2$ and $\mathcal{X} = \{1, 2, 3\}$, with 3 playing the role of the erasure symbol. Let

$$f(x) \triangleq \begin{cases} 1, & x \in \{1, 2\}, \\ -1 & x = 3. \end{cases}$$

Then $f \in \mathcal{F}_0$, $h_b\left(\frac{1}{2} + \frac{f(x)}{2\|f\|_\infty}\right) = 0$ for $x \in \mathcal{X}$ and $t^* = 1$. But, from Lemma 11, $t^* \leq H(X|S) = 1$. The result follows. $\square$

References

[1] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AMLBook, Mar. 2012.

[2] L. Sweeney, "K-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002.

[3] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *TCC*, 2006.

[4] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*. Springer, 2006, vol. 4052, pp. 1–12.

[5] S. Salamatian, A. Zhang, F. P. Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft, "How to hide the elephant-or the donkey-in the room: Practical privacy against statistical inference for large data," *IEEE GlobalSIP*, 2013.

[6] S. Salamatian, A. Zhang, F. P. Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft, "Managing your private and public data: Bringing down inference attacks against your privacy," *IEEE J. Sel. Topics Signal Process.*, October 2015.

[7] S. Bhamidipati, N. Fawaz, B. Kveton, and A. Zhang, "PriView: Personalized Media Consumption Meets Privacy against Inference Attacks," *IEEE Software*, vol. 32, no. 4, pp. 53–59, Jul. 2015.

[8] C. E. Shannon, "Communication theory of secrecy systems," *Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949.

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

[10] J. C. Duchi and M. J. Wainwright, "Distance-based and continuum fano inequalities with applications to statistical estimation," *arXiv preprint arXiv:1311.2669*, 2013.

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

[12] L. Breiman and J. H. Friedman, "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, vol. 80, no. 391, pp. 580–598, Sep. 1985.

[13] A. Rényi, "On measures of dependence," *Acta mathematica hungarica*, vol. 10, no. 3, pp. 441–451, 1959.

[14] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004.

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

[16] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Upper Saddle River: Pearson, Aug. 2009.

[17] G. R. Kumar and T. A. Courtade, "Which Boolean functions maximize information of noisy inputs?" *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4515–4525, Aug. 2014.

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

[19] F. P. Calmon and N. Fawaz, "Privacy against statistical inference," in *Proc. 50th Ann. Allerton Conf. Commun., Contr., and Comput.*, 2012, pp. 1401–1408.

[20] A. Zhang, S. Bhamidipati, N. Fawaz, and B. Kveton, "PriView: Media Consumption and Recommendation Meet Privacy Against Inference Attacks," in *IEEE Web 2.0 Security and Privacy Workshop*, 2014.

[21] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Ann. Allerton Conf. Commun., Contr., and Comput.*, 1999, pp. 368–377.

[22] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the markov operator," *Ann. Probab.*, vol. 4, no. 6, pp. 925–939, Dec. 1976.

[23] 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.

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

[25] I. Csiszár, *Information Theory And Statistics: A Tutorial*. Now Publishers Inc, 2004.

[26] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," in *Proc. 48th Ann. Allerton Conf. Commun., Contr., and Comput.*, 2010, pp. 1327–1333.

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

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

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

[30] 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.

[31] O. Sarmanov, "Maximum correlation coefficient (nonsymmetric case)," *Selected Translations in Mathematical Statistics and Probability*, vol. 2, pp. 207–210, 1962.

[32] H. S. Witsenhausen, "On sequences of pairs of dependent random variables," *SIAM J. on Appl. Math.*, vol. 28, no. 1, pp. 100–113, Jan. 1975.

[33] Y. Polyanskiy, "Hypothesis testing via a comparator," in *Proc. 2012 IEEE Int. Symp. on Inf. Theory*, Jul. 2012, pp. 2206–2210.

[34] M. Raginsky, "Logarithmic Sobolev inequalities and strong data processing theorems for discrete channels," in *Proc. 2013 IEEE Int. Symp. on Inf. Theory*, Jul. 2013, pp. 419–423.

[35] F. P. Calmon, M. Varia, M. Médard, M. Christiansen, K. Duffy, and S. Tessaro, "Bounds on inference," in *Proc. 51st Ann. Allerton Conf. Commun., Contr., and Comput.*, Oct. 2013, pp. 567–574.

[36] A. Makur and L. Zheng, "Bounds between Contraction Coefficients," *arXiv:1510.01844 [cs, math]*, Oct. 2015.

[37] J. Liu, T. A. Courtade, P. Cuff, and S. Verdú, "Brascamp-Lieb Inequality and Its Reverse: An Information Theoretic View," *arXiv:1605.02818 [cs, math]*, May 2016.

[38] S.-L. Huang, C. Suh, and L. Zheng, "Euclidean information theory of networks," *IEEE Trans. on Info. Theory*, vol. 61, no. 12, pp. 6795–6814, 2015.

[39] A. Buja, "Remarks on Functional Canonical Variates, Alternating Least Squares Methods and Ace," *The Annals of Statistics*, vol. 18, no. 3, pp. 1032–1069, Sep. 1990.

[40] A. Makur, F. Kozynski, S.-L. Huang, and L. Zheng, "An efficient algorithm for information decomposition and extraction," in *Proceedings of the 53rd Annual Allerton Conference on Communication, Control and Computing, Allerton House, UIUC, Illinois, USA*, 2015.

[41] 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.

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

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

[44] 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.

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

[46] G. Kindler, R. O'Donnell, and D. Witmer, "Remarks on the Most Informative Function Conjecture at fixed mean," *arXiv:1506.03167*, 2015.

[47] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and the mutual information between Boolean functions," in *Proc. 51st Ann. Allerton Conf. Commun., Contr., and Comput.*, Oct. 2013, pp. 13–19.

[48] O. Ordentlich, O. Shayevitz, and O. Weinstein, "An Improved Upper Bound for the Most Informative Boolean Function Conjecture," *arXiv:1505.05794 [cs, math]*, May 2015.

[49] V. Chandar and A. Tchamkerten, "Most informative quantization functions," in *Proc. ITA Workshop, San Diego, CA, USA*, 2014.

[50] A. Samorodnitsky, "The "Most informative Boolean function" conjecture holds for high noise," *arXiv:1510.08656 [cs, math]*, Oct. 2015.

[51] I. S. Reed, "Information theory and privacy in data banks," in *Proc. of the National Computer Conference and Exposition*, ser. AFIPS '73. New York, NY, USA: ACM, June 1973, pp. 581–587.

[52] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer, "From t-closeness-like privacy to postrandomization via information theory," *IEEE Trans. on Knowledge and Data Engineering*, vol. 22, no. 11, pp. 1623–1636, Nov. 2010.

[53] L. Sankar, S. Rajagopalan, and H. Poor, "Utility-Privacy Tradeoffs in Databases: An Information-Theoretic Approach," *IEEE Trans. on Inf. Forensics and Security*, vol. 8, no. 6, pp. 838–852, Jun. 2013.

[54] R. Tandon, L. Sankar, and H. Poor, "Discriminatory lossy source coding: Side information privacy," *IEEE Transactions on Information Theory*, vol. 59, no. 9, pp. 5665–5677, Sep. 2013.

[55] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proceedings of the twenty-second ACM Symposium on Principles of Database Systems*, New York, NY, USA, 2003, pp. 211–222.

[56] A. Makhdoumi and N. Fawaz, "Privacy-utility tradeoff under statistical uncertainty," in *Proc. 51th Annual Allerton Conference on Communication, Controrl, and Computation*, 2013, pp. 1627–1634.

[57] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *arXiv:1405.3629 [cs, math]*, May 2014.

[58] C. T. Li and A. E. Gamal, "Maximal correlation secrecy," *arXiv:1412.5374 [cs, math]*, Dec. 2014.

[59] F. P. Calmon, M. Varia, and M. Médard, "On information-theoretic metrics for symmetric-key encryption and privacy," in *Proc. 52nd Annual Allerton Conference on Communication, Control, and Computing*, 2014.

[60] S. Chakraborty, N. Bitouze, M. Srivastava, and L. Dolecek, "Protecting data against unwanted inferences," in *2013 IEEE Information Theory Workshop (ITW)*, Sep. 2013, pp. 1–5.

[61] S. Asoodeh, F. Alajaji, and T. Linder, "Notes on information-theoretic privacy," in *Proc. 52nd Ann. Allerton Conf. Commun., Contr., and Comput.*, Sep. 2014, pp. 1272–1278.

[62] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Information Extraction Under Privacy Constraints," *arXiv preprint arXiv:1511.02381*, 2015.

[63] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Medard, "From the information bottleneck to the privacy funnel," in *IEEE Inf. Theory Workshop (ITW)*, 2014, pp. 501–505.

[64] A. Makhdoumi, F. P. Calmon, and M. Médard, "Forgot your password: Correlation dilution," in *International Symp. on Info. Theory*, 2015, pp. 2944–2948.

[65] S. Beigi and A. Gohari, "On the duality of additivity and tensorization," in *International Symp. on Info. Theory*. IEEE, 2015, pp. 2381–2385.

[66] C. R. J. Roger A. Horn, *Topics in Matrix Analysis*. Cambridge University Press, 1994.

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

[68] 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.

[69] R. O'Donnell, "Some topics in analysis of Boolean functions," in *Proc. 40th ACM Symp. on Theory of Computing*, 2008, pp. 569–578.

[70] M. Raginsky, J. G. Silva, S. Lazebnik, and R. Willett, "A recursive procedure for density estimation on the binary hypercube," *Electron. J. Statist.*, vol. 7, pp. 820–858, 2013.

[71] R. O'Donnell, *Analysis of Boolean Functions*, 1st ed. New York, NY: Cambridge University Press, Jun. 2014.

[72] F. P. Calmon, M. Médard, L. Zeger, J. Barros, M. M. Christiansen, and K. R. Duffy, "Lists that are smaller than their parts: A coding approach to tunable secrecy," in *Proc. 50th Annual Allerton Conf. on Commun., Control, and Comput.*, 2012.

[73] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv:physics/0004057 [physics.data-an]*, Apr. 2000.

[74] S. Goldwasser and S. Micali, "Probabilistic encryption," *Journal of Computer and System Sciences*, vol. 28, no. 2, pp. 270–299, Apr. 1984.

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

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

[77] S. Kamath and V. Anantharam, "Non-interactive simulation of joint distributions: The Hirschfeld-Gebelein-Rényi maximal correlation and the hypercontractivity ribbon," in *Proc. 50th Ann. Allerton Conf. Commun., Contr., and Comput.* IEEE, 2012, pp. 1057–1064.

[78] T. Berger and R. Yeung, "Multiterminal source encoding with encoder breakdown," *IEEE Trans. on Inf. Theory*, vol. 35, no. 2, pp. 237–244, Mar. 1989.

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

[80] H. G. Eggleston, *Convexity*, 1st ed. Cambridge England: Cambridge University Press, Jan. 2009.