Several Interpretations of Max-Sliced Mutual Information

Dor Tsur
School of Electrical Engineering
Ben-Gurion University of the Negev
Email: dortz@post.bgu.ac.il

Haim Permuter
School of Electrical Engineering
Ben-Gurion University of the Negev
Email: haimp@bgu.ac.il

Ziv Goldfeld
School of Electrical Engineering
Cornell University
Email: goldfeld@cornell.edu

2024 IEEE International Symposium on Information Theory (ISIT) | 979-8-3503-8284-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISIT57864.2024.10619532

***Abstract*—Max-sliced mutual information (mSMI) was recently proposed as a data-efficient measure of dependence. This measure extends popular correlation-based methods and proves useful in various machine learning tasks. In this paper, we extend the notion of mSMI to discrete variables and investigate its role in popular problems of information theory and statistics. We use mSMI to propose a soft version of the Gács-Körner common information, which, due to the mSMI structure, naturally extends to continuous domains and multivariate settings. We then characterize the optimal growth rate in a horse race with constrained side information. Additionally, we examine the error of independence testing under communication constraints. Finally, we study mSMI in communications. We characterize the capacity of discrete memoryless channels with constrained encoders and decoders, and propose an mSMI-based scheme to decode information obtained through remote sensing. These connections motivate the use of max-slicing in information theory, and benefit from its merits.**

I. Introduction

Information theory plays a key role in the characterization and analysis across a myriad of fields that involve probability and statistical inference. Specifically, mutual information (MI) is used to analyze the shared information between two dependent variables, being central to both classical [1], [2] and contemporary methodologies [3], [4]. For example, MI characterizes the utility of side-information, in compression [5] and investment [6]. When a rate constraint is imposed on the side information, or some statistical inference task is performed under communication constraints, MI can be used to characterize the performance in the resulting setting [7], [8].

Another objective of processing of dependent random variables is the extraction of common information [9]. This paper focuses on the Gács-Körner common information (GKCI) between a pair (X,Y), which seeks a maximum entropy variable that can be deterministically extracted from both X and Y. However, GKCI often results in trivial solutions due to restrictive conditions. To address this, approximations involve either extracting randomness from X with limited disagreement with Y [10], [11] or using a genie-aided mechanism for common message generation [12], both relaxing the functional constraint through the notion of controlled rates. However, such rates are implicitly controlled by external hyperparameters and can result in multiletter optimization terms. Additionally, GKCI and its generalizations do not seamlessly generalize to more than two users, and lack a proper setting for continuous spaces, making their adaptation to data-driven settings challenging. For more information on GKCI, see [13], which provides a comprehensive review and draws connections to additional settings.

In this paper, we consider max-sliced MI (mSMI) [14], which seeks the most informative projections of the pair (X,Y) within a given function class. mSMI benefits from well-defined structural properties, a closed form in the Gaussian case and sharp neural estimation bounds. Additionally, it was shown beneficial to contemporary machine learning tasks, encompassing independence testing, multi-view representation learning and generative modelling. By extending the notion of mSMI to discrete spaces, we employ it for the characterization and generalization of popular subjects in information theory and statistics. The main contributions of the paper are

1) We show that mSMI serves as a soft generalization of GKCI that focuses on the structural constraints of the feasible set, rather than an implicit rate constraint. This connection implies a possible extension of GKCI to continuous spaces (Section III).
2) Using the relationship between mSMI and GKCI, we propose a tractable multivariate extension of both classical and soft notions of GKCI (Section IV).
3) We use mSMI to characterize the optimal growth rate in a horse race under computational constraints, which relates to source coding under similar side-information constraints (Section V).
4) We use mSMI to characterize the type-II error exponent for the independence testing problem under constrained communications (Section VI).
5) In communications, we use mSMI to study the capacity under constrained, encoder decoder architectures. Additionally, we consider a constrained sensing problem, for which we propose a two-stage strategy from which we can decode the conveyed message in the channel only through constrained sensors (Section VII).

II. Max-Sliced Mutual Information

This section introduces mSMI, which seeks MI maximizing functions of a given jointly distributed pair. Unless stated otherwise, we consider discrete random variables X and Y, defined over $\mathcal{X}$ and $\mathcal{Y}$, respectively. We adapt the definition of

mSMI to discrete spaces and characterize its properties. mSMI is formally given as follows

Definition 1 (Max-sliced mutual information)**.** *Let $(X,Y)\sim P_{XY}$, let $\mathcal{F}_k$ and $\mathcal{G}_k$ be non-empty classes of mapping from $\mathcal{X}$ and $\mathcal{Y}$ to $\mathcal{U}$ with $|\mathcal{U}|=k$, respectively, and let $\mathcal{H}_k=\mathcal{F}_k\times\mathcal{G}_k$. The mSMI between X and Y w.r.t. $\mathcal{H}_k$ is given by*

$$\overline{\mathsf{SI}}_{\mathcal{H}_k}(X,Y)=\sup_{(f,g)\in\mathcal{H}_k} I(f(X);g(Y)). \tag{1}$$

When $\mathcal{H}_k$ includes all mappings from $\mathcal{X}$ and $\mathcal{Y}$ to $\mathcal{U}$ we denote $\overline{\mathsf{SI}}_k(X,Y)=\overline{\mathsf{SI}}_{\mathcal{H}_k}(X,Y)$.

The authors of [14] investigated the structural properties of mSMI. We state two properties that we later use in this work

Proposition 1 (Structural properties (partial))**.** *Let $(X,Y)\sim P_{XY}$ and let $\mathcal{H}$ be an arbitrary class of nonconstant functions (f,g). The following properties hold:*

1) ***Bounds:*** *For $\mathcal{H}_1\subseteq\mathcal{H}_2$:*

$$\overline{\mathsf{SI}}_{\mathcal{H}_1}(X;Y)\le\overline{\mathsf{SI}}_{\mathcal{H}_2}(X;Y)\le\mathsf{I}(X;Y).$$

2) ***Tensorization:*** *For mutually independent $\{(X_i,Y_i)\}_{i=1}^n$,*

$$\overline{\mathsf{SI}}_k\left(X^n,Y^n\right)=\sum_{i=1}^n\overline{\mathsf{SI}}_k(X_i;Y_i)$$

Beyond the above properties, mSMI was shown to identify independence, to follow a sub-chain rule, and a KL representation thereof was given. We note that the definition of $\overline{\mathsf{SI}}$ readily extends to $\mathsf{range}(f)\neq\mathsf{range}(g)$.

In general, we could define the mSMI using stochastic mappings, i.e., $X\overset{P_{U|X}}{\to}U$ and $Y\overset{P_{V|Y}}{\to}V$. The corresponding mSMI is given by

$$\overline{\mathsf{SI}}^{\mathsf{st}}_k(X,Y):=\max_{P_{U|X},P_{V|Y}} I(U;V),$$

where the maximization is over all transitional kernels $P_{U|X}$ and $P_{V|Y}$ such that $|\mathcal{U}|=|\mathcal{V}|=k$. However, we can show that optimizing over deterministic functions is sufficient

Lemma 1. *Let $(X,Y)\sim P_{X,Y}$, then*

$$\overline{\mathsf{SI}}^{\mathsf{st}}_k(X,Y)=\overline{\mathsf{SI}}_k(X,Y). \tag{2}$$

Finally, we define the one-sided mSMI by $\overline{\mathsf{SI}}^X_{\mathcal{H}}$ and $\overline{\mathsf{SI}}^Y_{\mathcal{H}}$, which are achieved by taking $g(Y)=Y$ or $f(X)=X$ respectively in (1).

III. Soft Gaćs-Körner Common Information

The GKCI was proposed as a means for quantifying how much randomness is shared between two dependent variables $(X,Y)\sim P_{XY}$. Formally, it is given as follows

Definition 2 (Gaćs-Körner Common Information)**.** *Let $(X,Y)\sim P_{XY}$. The GKCI is given by the solution to the following optimization problem*

$$\mathsf{C}_{\mathsf{GK}}(X,Y):=\max_{f,g} H(U),\quad s.t.\quad U=f(X)=g(Y), \tag{3}$$

where $\mathcal{U}$ is the alphabet of U with $|\mathcal{U}|\le\min\{|\mathcal{X}|,|\mathcal{Y}|\}$ and the equality is with probability 1.

The main drawback of the GKCI is that the equality constraint is too restrictive, resulting in $\mathsf{C}_{\mathsf{GK}}(X,Y)=0$ for most $P_{XY}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})$. Specifically, [13] proposes the following equivalence

Theorem 1 (Prop. 3.2.3, [13])**.** *Let $\bar{\rho}(X,Y)=\sup_{f,g}\mathsf{Cov}(f(X),g(Y))/\sqrt{\mathsf{Var}(f(X))\mathsf{Var}(g(Y))}$ For $(X,Y)\sim P_{XY}$, the following are equivalent*

1) $\bar{\rho}(X,Y)=1$
2) $\mathsf{C}_{\mathsf{GK}}(X,Y)>0$
3) There exist a pair of non-constant functions (f,g) such that $f(X)=g(Y)$ P_{XY}-a.s.

In this work we consider a finer granularity of the GKCI and define an alphabet-dependent version thereof. To this end, we denote the size of $\mathcal{U}$ with k and define the kth GKCI function class as

$$\mathcal{H}_{\mathsf{GK},k}(P_{XY})=\{(f,g):U=f(X)=g(Y),P_U\in\mathcal{P}(\mathcal{U})\},$$

i.e., the class of mappings to $\mathcal{U}$ that follow the GK constraint. Next, we omit the dependence on P_{XY} to avoid heavy notation. This allows us to define $\mathsf{C}_{GK,k}(X,Y):=\max_{(f,g)\mathcal{H}_{\mathsf{GK},k}\in}H(U)$. Consequently we have the following

Lemma 2. *The following hold*

1) If $k_1\le k_2$, then $\mathsf{C}_{\mathsf{GK},k_1}(X,Y)\le\mathsf{C}_{\mathsf{GK},k_2}(X,Y)$.
2) Let $k_{max}=\min\{|\mathcal{X}|,|\mathcal{Y}|\}$, then

$$\mathsf{C}_{\mathsf{GK}}(X,Y)=\mathsf{C}_{\mathsf{GK},k_{max}}(X,Y).$$

A. Proposed Generalization

The canonical example under which GKCI does not trivially nullify considers X and Y that decompose into (U,X') and (U,Y'), respectively, with $X'\perp\!\!\!\perp Y'$. In this case, the entire MI is captured by both GKCI and mSMI, i.e.,

$$\mathsf{C}_{\mathsf{GK}}(X;Y)=\overline{\mathsf{SI}}(X,Y)=I(X;Y)=H(U).$$

We can extend the relation between GKCI and mSMI to the general case through the following proposition.

Proposition 2 (mSMI generalizes GKCI)**.** *Let $(X,Y)\sim P_{XY}$, $k\le\min\{|\mathcal{X}|,|\mathcal{Y}|\}$, $\mathcal{F}_k=\{f:\mathcal{X}\to\mathcal{U}\}$, $\mathcal{G}_k=\{g:\mathcal{Y}\to\mathcal{U}\}$ and set $\mathcal{H}_k=\mathcal{F}_k\times\mathcal{G}_k$. Denote*

$$\mathcal{H}_{k,GK}:=\{(f,g)\in\mathcal{H}_k:f(X)=g(Y)\quad w.p.\quad 1\}.$$

Then,

$$\overline{\mathsf{SI}}_{\mathcal{H}_{GK}}(X,Y)=\mathsf{C}_{\mathsf{GK}}(X,Y), \tag{4}$$

and for any $\mathcal{H}_k\supseteq\mathcal{H}_{GK}$

$$\mathsf{C}_{\mathsf{GK}}(X,Y)\le\overline{\mathsf{SI}}_{\mathcal{H}_k}(X,Y)\le I(X;Y), \tag{5}$$

Proposition 2 poses mSMI as a generalization of GKCI. In this sense, we interpret mSMI as *soft* GKCI, in the sense that we relax the equivalence notion $f(X)=g(Y)$ into the notion of conditional entropy minimization, while both seek

entropy maximization. This can be seen from the following decomposition of mSMI

$$\overline{\mathsf{SI}}_{\mathcal{H}} = \max_{f\in\mathcal{F}}(H(f(X)) - \min_{g\in\mathcal{G}} H(f(X)|g(Y))),$$

Using mSMI, we can capture the entire set of possible values between $\mathsf{C}_{\mathsf{GK}}(X,Y)$ and $I(X;Y)$ by considering different function classes $\mathcal{H}$.

B. Comparison with Approximate GKCI

The most related work to the proposed soft GKCI is [11], where an approximate version of GKCI was proposed. The approximation seeks a mapping $f:\mathcal{X}\to\mathcal{U}$ which maximizes $H(f(X))$, while relaxing the equality constraint into a disagreement constraint of $H(f(X)|Y)\leq\delta$. Equivalently, the proposed problem can be presented as

$$G_\lambda(X,Y) = \max_{f:\mathcal{X}\to\mathcal{U}} H(f(X)) - \lambda H(Y|f(X)),$$

with $\mathsf{range}(f)\leq\min\{\mathcal{X},\mathcal{Y}+1\}$. The parameter λ controls the trade-off between minimizing disagreement and maximizing entropy. The approximate GKCI coincides with GKCI through the limit

$$\lim_{\lambda\to\infty} G_\lambda(X,Y) = \mathsf{C}_{\mathsf{GK}}(X,Y).$$

The main difference between $\overline{\mathsf{SI}}(X,Y)$ and $G_\lambda(X,Y)$ is through the notion of relaxation. For $G_\lambda(X,Y)$, the relaxation represents the information rate which implicitly controls the resulting set of functions we optimize over. $\overline{\mathsf{SI}}(X,Y)$, on the other hand, explicitly considers computational constraint imposed on the system, which are manifested through the structure of $\mathcal{F}$ and $\mathcal{G}$.

C. Common Information in Continuous Spaces

In this section we consider absolutely continuous random variables (X,Y). In this case, finding functions f and g that form a P_{XY}-a.s. equality becomes significantly more challenging, implying a reduction in the set of distributions for which the GKCI does not nullify. Thus, only discrete distributions were considered in analysis of GKCI and its relaxations. In contrast to GKCI, mSMI is properly defined over continuous spaces [14]. As optimization over mappings between continuous spaces introduces more structure to our problem, we can choose the function classes to be the family of linear projections into some k-dimensional space[1], i.e.

$$\overline{\mathsf{SI}}_k(X;Y) := \max_{(\mathrm{A},\mathrm{B})\in\mathrm{St}(k,d_x)\times\mathrm{St}(k,d_y)} \mathsf{I}(\mathrm{A}^\intercal X;\mathrm{B}^\intercal Y),$$

where $\mathrm{St}(k,d)$ is the Stiefel manifold of $d\times k$ matrices with orthonormal columns.

Being the focus of [14], continuous mSMI was shown beneficial in several aspects. First, mSMI has a closed-form expression in the Gaussian case, which stems from an equivalence with canonical correlation analysis.

[1]The definition remains the same for general function spaces [14].

Proposition 3 (Prop. 2, [14]). *Let $X\sim\mathcal{N}(m_X,\Sigma_X)$ and $Y\sim\mathcal{N}(m_Y,\Sigma_Y)$ be d_x– and d_y–dimensional jointly Gaussian vectors with nonsingular covariance matrices and cross-covariance Σ_{XY}. For any $k\leq d_x\wedge d_y$, we have*

$$\overline{\mathsf{SI}}_k(X;Y) = -\frac{1}{2}\sum_{i=1}^{k}\log\left(1-\sigma_i(\mathrm{T}_{XY})^2\right), \tag{6}$$

where $\mathrm{T}_{XY}=\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}\in\mathbb{R}^{d_x\times d_y}$, and $\sigma_k(\mathrm{T}_{XY})\leq\ldots\leq\sigma_1(\mathrm{T}_{XY})\leq 1$ are the top k singular values of T_{XY}.

When the Gaussian assumption is violated, mSMI can be estimated with parametric models. Building upon the literature of neural estimation [15], a neural network-based estimator can be formulated, which benefits from sharp estimation error rates. In fact, every neural estimation methodology effectively estimated a max-sliced version of the corresponding measure [15], [16], as the limitation to specific neural network classes imposes a constrained family of estimators. Consequently, mSMI serves as the most viable generalization of the GKCI to continuous spaces.

IV. Multivariate Common Information via mSMI

In this section we leverage the relation between mSMI and GKCI to extend the notion of soft common information to the multivariate case. We define the multivariate GKCI of X^m by a simple generalization of C_{GK}, as the entropy maximizing variable which is given by a function of every X_i in X^m, i.e.,

$$\mathsf{C}^m_{\mathsf{GK}}(X^m) := \max_U H(U), \quad \text{s.t.} \quad U=f(X_i), \quad i=1,\ldots m.$$

To define a multivariate generalization of mSMI, we consider a popular multivariate notion of MI, termed total correlation (TC) [17], [18].

Definition 3. *Let $(X_1,\ldots,X_m)\sim P_{X^m}$ and $f^m=(f_1,\ldots,f_m)$. The TC of X^m is given by*

$$TC(X_1;\ldots;X_m) = \mathsf{D}_{\mathsf{KL}}\left(P_{X^m}\middle\|\bigotimes_{i=1}^m P_{X_i}\right),$$

where $\bigotimes_{i=1}^m P_i = P_1\otimes P_2\otimes\cdots\otimes P_m$.

Consequently, the max-sliced TC (mSTC), is given by

$$\begin{aligned}&\overline{\mathsf{TC}}_{\mathcal{H}}(X^m)\\ &:= \frac{1}{m-1}\max_{f^m\in\mathcal{H}} TC(f_1(X_1);\ldots;f_m(X_m))\\ &= \frac{1}{m-1}\max_{f^m\in\mathcal{H}} \mathsf{D}_{\mathsf{KL}}((f_1,\ldots,f_m)_\# P_{X^m}\|\bigotimes_{i=1}^m (f_{i,\#}P_{X_i})).\end{aligned}$$

Equipped with the above definitions, we can state a multivariate generalization of Proposition 2

Proposition 4. *Let $\mathcal{H}^m_{GK,k}$ be the m-GK function class, i.e.*

$$\mathcal{H}^m_{GK,k} = \{f_1,\ldots,f_m : f_i(X_i)=f_j(X_j)\quad\forall i,j=1,\ldots,m\}$$

Then, $\overline{\mathsf{TC}}_{\mathcal{H}^m_{GK,k}}(X^m) = TC_{GK}(X^m)$.

Note that the proposed relation holds for both discrete and continuous settings. Furthermore, a neural estimation of mSTC is attainable via [19].

V. Gambling With Limited Side Information

In this section, following the setting from [8], we characterize the optimal growth rate in a horse race with constrained side information using one-sided mSMI. An example for such model is an investor, which has access to the entire internet as side information, but has limited computational capabilities in processing such information through their hardware.

Formally, we consider a horse race, which comprises a variable $X \sim p$, where $p(i)$ is the winning probability of the ith horse, a simplex vector b, termed portfolio, and a payoff $o \in \mathbb{R}_{\geq 0}^{|\mathcal{X}|}$, such that the investor receives o_i dollars for each dollar invested in horse i. We are interested in the optimal growth rate of the gambler, given by $W^\star = \max_b W(b,p)$, with $W(b,p) = \mathbb{E}\left[\log b^\mathsf{T} X'\right]$, where $X' = (0,\ldots,0,o_i,0,\ldots,0)$ is a one-hot encoding of X', thus following the same distribution.

Kelly [6] showed that $W^\star = \sum_{i=1}^{|\mathcal{X}|} p_i \log o_i - H(X)$ and the optimal portfolio is given by $b^\star = p$. Additionally, in the presence of side information Y, the *increase* in growth rate is given by $\Delta = I(X;Y)$. We consider a constrained version of the side information, given by $g(Y)$ with $g \in \mathcal{G}$, and we look for the best optimal growth rate w.r.t. $\mathcal{G}$. We have the following result

Proposition 5. *Let V be the side information in the horse-race, where the portfolio b is calculated from $g(Y)$. Then, the optimal growth rate is given by*

$$\Delta_{\mathcal{G}} = \overline{\mathsf{SI}}^{Y}_{\mathcal{H}}(X;Y).$$

This result follows directly from the work of Erkip [8], which showed that under rate constrained side information the optimal growth rate is given by the maximization of the set of rate constrained random variables.

Building on the relation to the result of [8], we can attain further interpretation of mSMI. As studied by [8], Gambling in a horse race with constrained side information has similar characteristics to the problem of source coding with constrained side information. Specifically, the *minimum descriptive complexity* of X given a description $g(Y)$ with rate $R \geq H(g(Y))$ is given by

$$C(\mathcal{G}) = \min_{g \in \mathcal{G}} H(X|g(Y)) \tag{7}$$

Consequently, as noted by [8],

$$H(X) - C(\mathcal{G}) = \Delta(\mathcal{G}) = \overline{\mathsf{SI}}^{Y}_{\mathcal{H}}(X,Y),$$

where $\mathcal{G}$ is the set of all g whose entropy is upper bounded by R.

VI. Doubly Constrained Independence Testing

This section considers a hypothesis test against independence under communications constraints [7]. The hypothesis test is given by

$$\begin{aligned} \mathsf{H}_0: &\quad (X,Y) \sim P_{XY} \\ \mathsf{H}_1: &\quad (X,Y) \sim P_X \otimes P_Y, \end{aligned} \tag{8}$$

i.e., we are interested in testing whether (X,Y) are sampled from the joint distribution, or the product of marginals. Due to Stein [20], the exponent of the type-II error, given by

$$\beta(n,e) := \min_{A \in \mathcal{X}^n \times \mathcal{Y}^n} \left\{ (P_X \otimes P_Y)^{\otimes n}(A) | P_{XY}^{\otimes n}(A) \geq 1-\epsilon \right\},$$

behaves as

$$\lim_{n\to\infty} \log \beta(n,\epsilon) = -I(X;Y),$$

for any $\epsilon \in (0,1)$. Ahlswede and Csiszár [7] studied the hypothesis test (8) under communication constraints, i.e., when X^n is processed through $f(X^n)$ whose range is upper bounded with some $R \geq 0$, i.e. $f \in \mathcal{F}_R$, such that

$$\mathcal{F}_R := \left\{ f : \frac{1}{n} \log\left(|\mathsf{range}(f)|\right) \leq R \right\}$$

To this end, define the function dependent exponent as

$$\begin{aligned} \beta(n,e,f) := \min_{A \in f(\mathcal{X}^n)\times\mathcal{Y}^n} \{ & P_{f(X^n)} \otimes P_Y^{\otimes n}(A) : \\ & P_{f(X^n),Y^n}(A) \geq 1-\epsilon \}, \end{aligned}$$

and let $\beta_R(n,e) = \inf_{f \in \mathcal{F}_{R_1}} \beta(n,e,f)$ be the range-dependent exponent. Under this setting, [7] showed that

$$\begin{aligned} &\lim_{n\to\infty} \log \beta_R(n,\epsilon) \\ &\qquad = \sup_{k \in \mathbb{N}, f \in \mathcal{F}_{R_1}} \left(\frac{1}{k} \mathsf{D}_{\mathsf{KL}}(P_{f(X^k)Y^k} \| P_{f(X^k)} P_{Y^k})\right). \end{aligned}$$

Furthermore, [7] showed that this relation extends to doubly-constrained communications, i.e., when both X^n and Y^n are processed through $f(X^n)$ and $g(Y^n)$, The resulting exponent is given by the following KL term,

$$\begin{aligned} &\lim_{n\to\infty} \log \beta_{R_1,R_2}(n,\epsilon) \\ &= - \sup_{\substack{(f,g) \in \mathcal{F}_{R_1} \times \mathcal{G}_{R_2}, \\ k \in \mathbb{N}}} \left(\frac{1}{k} \mathsf{D}_{\mathsf{KL}}(P_{f(X^k)g(Y^k)} \| P_{f(X^k)} P_{g(Y^k)}) \right), \end{aligned} \tag{9}$$

Unfortunately, the expression (9) is a multiletter expression and was not further analysed. In fact, the authors of [7] note that reducing the above term into a single-letter expression is mathematically challenging.

Due to the tensorization of mSMI (Proposition 1), we can reduce (9) into a single-letter expression under appropriate functional constraints. We have the following

Lemma 3. *Let $\beta_{\mathcal{H}}$ be the type-II error of the hypothesis test (8) under doubly constrained communications. Then. for any $\epsilon \in (0,1)$, we have*

$$\lim_{n\to\infty} \log \beta_{\mathcal{H}}(n,\epsilon) = -\overline{\mathsf{SI}}_{\mathcal{H}}(X,Y). \tag{10}$$

As discussed in [7] application of Stein's lemma does not depend on the constraints imposed on the sets of $\mathcal{F}$ and $\mathcal{G}$.

Thus the proof immediately follows from [7]. A similar result holds for one-sided mSMI holds by taking $\mathcal{F}$ (or $\mathcal{G}$) that contains the identity mapping in (10).

Remark 1. *Even though mSMI primarily focuses on functional constraints, there are cases where understanding the underlying rate constraint is beneficial. This may be advantageous on the block-level analysis, i.e., when we study mappings $f : \mathcal{X}^n \to \mathcal{X}^k$. We note that when mSMI is generalized to act on blocks X^n, Y^n with range $k \leq n$, it no longer yields a single-letter formula. However, the resulting class $\mathcal{H}$ can be considered a subset of the rate constrained set $(f, g) : H(f(X^n)), H(g(Y^n)) \leq R$ with $R = k/n$.*

VII. mSMI in Communications

In this section we consider two operational problems in communications. These problems consider some constraints on the computational capabilities of some of the users in the system. We will show that the communication rates in these cases can be characterized with mSMI.

A. Constrained Communications

We consider a discrete memoryless channel, with a constrained encoder decoder pair, whose constraints are realized by $f \in \mathcal{F}$ and $g \in \mathcal{G}$, respectively. One example is resolution reduction of the channel input due to computational restrictions of the encoder and decoder units hardware, or amplifier specifications. The communication scheme is described in Figure 1. In this case, we can consider the effective channel inputs and outputs as $(X_i, V_i) = (f(U_i), g(Y_i))$, respectively, where $U^n(M)$ is the encoder outputs. In this setting we have the following result

Proposition 6. *Fix $\mathcal{H}$. The capacity of the communication constrained channel is given by $\mathsf{C}_{\mathcal{H}} := \sup_{P_U} \overline{\mathsf{SI}}_{\mathcal{H}}(U;Y)$.*

Note that the constraints $f(U)$ and $g(Y)$ induce a new DMC. Consequently, the proof follows by standard random coding arguments for the induced DMC. We note that as a byproduct, a similar result holds when only the encoder or decoder are constrained, by taking the appropriate one-sided mSMI term in (6).

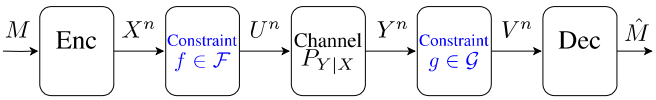

Fig. 1: Constrained communications. Encoder output enters $f : \mathcal{X} \to \mathcal{U}$ and the channel output enters $g : \mathcal{Y} \to \mathcal{V}$.

B. Constrained Sensing of Remote Communications

Consider a remote communication channel described by a fixed (encoder, channel, decoder) triplet. The encoder conveys a message $M \in [1 : 2^{nR}]$, thus sending $X^n(M)$, which is transformed into $Y \sim P_{Y|X}$. This setting is considered remote in the sense that we can only sample the channel input and outputs through sensors $f \in \mathcal{F}$ and $g \in \mathcal{G}$. This setting is visualized in Figure 2. It is motivated by sensing problems of biological channels, e.g., measuring information exchange in the brain through an external electrode apparatus.

To successfully recover the transmitted message, we propose a two-stage supervised strategy. First, we learn a codebook by observing a set of messages M and their corresponding sensed outputs $U^n(M)$. Then, equipped with the modified codebook, we decode the message from the sensed channel outputs, i.e. $\hat{M} = \hat{M}(V^n)$. These two stages can be conceptualized as a single-step process involving a training phase and inference. We propose the following

Proposition 7. *Under the aforementioned strategy, any rate $R \leq I(f(X); g(Y))$ is achievable.*

This result follows directly from the achievability of the corresponding DMC $P_{g(Y)|f(X)}$.

By selecting $(f^\star, g^\star) \in \operatorname{argmax}_{(f,g)\in\mathcal{H}} I(f(X); g(Y))$, we can extend the range of achievable rates. However, we note that changing f induces a new codebook. Additionally, we note that, under the random coding scheme, we require approximately $2^{n(I(f(X);g(Y)))}$ message observations. Finally, note that universal decoding schemes such as maximum (sliced) MI decoding [21] may be used and therefore obviate the need to know the latent channel distribution $P_{Y|X}$.

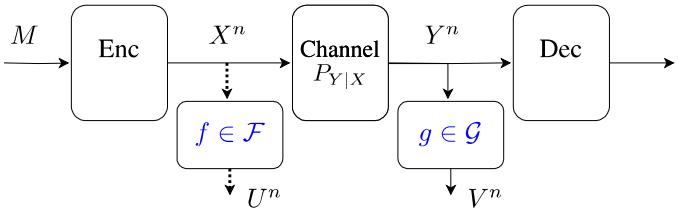

Fig. 2: Constrained sensing. node $U^n = f(X^n)$ (dashed) is used to learn a codebook on the first stage, and node $V^n = g(Y^n)$ is used to decode the sensed output given the learned codebook.

VIII. Conclusion

This paper explored various interpretations of mSMI in popular problems of information theory and statistics. Specifically, we used mSMI to generalize the notion of GKCI, providing us with a new continuous generalization and a multivariate extension thereof. Then, we showed the role of mSMI in gambling and compression under constrained side information. Finally, we showed the utility of mSMI to hypothesis testing, channel coding, and remote sensing under communication constraints. These connections provide new insights and capabilities to known problems, while benefiting from the well-established properties and implementations of mSMI. Future work consists of expanding the set of problems mSMI characterizes and drawing a connection with the Hirschfeld–Gebelein–Rényi (HGR) maximum correlation [22]–[24], which is a popular and intimately related measure of dependence.

REFERENCES

[1] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[2] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

[4] Ravid Shwartz-Ziv and Yann LeCun. To compress or not to compress–self-supervised learning and information theory: A review. *arXiv preprint arXiv:2304.09355*, 2023.

[5] Thomas M Cover and A Joy Thomas. *Elements of Information Theory*. Wiley, New-York, 2nd edition, 2006.

[6] John L Kelly. A new interpretation of information rate. *the bell system technical journal*, 35(4):917–926, 1956.

[7] Rudolf Ahlswede and Imre Csiszár. Hypothesis testing with communication constraints. *IEEE transactions on information theory*, 32(4):533–542, 1986.

[8] Elza Erkip and Thomas M Cover. The efficiency of investment information. *IEEE Transactions on information theory*, 44(3):1026–1040, 1998.

[9] Aaron Wyner. The common information of two dependent random variables. *IEEE Transactions on Information Theory*, 21(2):163–179, 1975.

[10] Salman Salamatian, Asaf Cohen, and Muriel Médard. Maximum entropy functions: Approximate gacs-korner for distributed compression. *arXiv preprint arXiv:1604.03877*, 2016.

[11] Salman Salamatian, Asaf Cohen, and Muriel Médard. Approximate gács-körner common information. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2234–2239. IEEE, 2020.

[12] Vinod M Prabhakaran and Manoj M Prabhakaran. Assisted common information with an application to secure two-party sampling. *IEEE Transactions on Information Theory*, 60(6):3413–3434, 2014.

[13] Lei Yu, Vincent YF Tan, et al. Common information, noise stability, and their extensions. *Foundations and Trends® in Communications and Information Theory*, 19(2):107–389, 2022.

[14] Dor Tsur, Ziv Goldfeld, and Kristjan Greenewald. Max-sliced mutual information. *arXiv preprint arXiv:2309.16200*, 2023.

[15] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540. PMLR, 2018.

[16] Dor Tsur, Ziv Aharoni, Ziv Goldfeld, and Haim Permuter. Neural estimation and optimization of directed information over continuous spaces. *IEEE Transactions on Information Theory*, 2023.

[17] Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

[18] Milan Studenỳ and Jirina Vejnarová. The multiinformation function as a tool for measuring stochastic dependence. *Learning in graphical models*, pages 261–297, 1998.

[19] Ke Bai, Pengyu Cheng, Weituo Hao, Ricardo Henao, and Larry Carin. Estimating total correlation with mutual information estimators. In *International Conference on Artificial Intelligence and Statistics*, pages 2147–2164. PMLR, 2023.

[20] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

[21] Valery D Goppa. Nonprobabilitistic mutual information without memory. *Problems of Control and Inform., Theory*, 4:97–102, 1975.

[22] Hermann O Hirschfeld. A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520–524, 1935.

[23] Hans Gebelein. Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941.

[24] Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3-4):441–451, 1959.

[25] Cheuk Ting Li and Abbas El Gamal. Strong functional representation lemma and applications to coding theorems. *IEEE Transactions on Information Theory*, 64(11):6967–6978, 2018.