Partial Information Decomposition via Deficiency for Multivariate Gaussians

Praveen Venkatesh*
praveen.venkatesh@alleninstitute.org
Allen Institute, Seattle, WA
University of Washington, Seattle, WA

Gabriel Schamberg*
gabes@mit.edu
Picower Institute for Learning and Memory
Massachusetts Institute of Technology, Cambridge, MA

Abstract—Bivariate partial information decompositions (PIDs) characterize how the information in a "message" random variable is decomposed between two "constituent" random variables in terms of *unique*, *redundant* and *synergistic* information components. These components are a function of the joint distribution of the three variables, and are typically defined using an optimization over the space of all possible joint distributions. This makes it computationally challenging to compute PIDs in practice and restricts their use to low-dimensional random vectors. To ease this burden, we consider the case of jointly Gaussian random vectors in this paper. This case was previously examined by Barrett [1], who showed that certain operationally well-motivated PIDs reduce to a *closed form* expression for *scalar messages*. Here, we show that Barrett's result does not extend to vector messages in general, and characterize the set of multivariate Gaussian distributions that reduce to closed-form. Then, for all *other* multivariate Gaussian distributions, we propose a convex optimization framework for approximately computing a specific PID definition based on the statistical concept of *deficiency*. Using simplifying assumptions specific to the Gaussian case, we provide an efficient algorithm to approximately compute the bivariate PID for multivariate Gaussian variables with tens or even hundreds of dimensions. We also theoretically and empirically justify the goodness of this approximation.

I. INTRODUCTION

Partial information decompositions (PIDs) provide a framework for characterizing the joint information content of three or more random variables. The three-variable case is usually discussed in terms of how *two* random variables, X and Y, contribute to the information about a "message" variable, M, and is hence referred to as the *bivariate* decomposition. More precisely, a bivariate PID decomposes the mutual information $I\big(M; (X, Y)\big)$ into four additive components: the information about M that is (i) unique to X; (ii) unique to Y; (iii) redundant in both X and Y; and (iv) synergistic, i.e., cannot be obtained from X or Y *individually*, but is present in their combination (X, Y). For an intuitive example, suppose $M = [M_1, M_2, M_3]$, $X = [M_1, M_2, M_3 \oplus Z]$, and $Y = [M_2, Z]$, where $M_1, M_2, M_3, Z \sim$ i.i.d. Ber(1/2) and $\oplus$ represents bitwise-XOR. Then, $I\big(M; (X, Y)\big)$ is 3 bits, and can be decomposed as follows: X has 1 bit of unique information about M (i.e., M_1) that is not contained in Y, while Y has zero unique information. X and Y have 1 bit of redundant information, i.e., M_2, which can be recovered from *either* X or Y. Finally, X and Y have 1 bit of synergistic information, i.e., M_3, which is *not* available in either X or Y *individually*, but can only be decoded when X and Y are taken *together*. Multiple competing definitions have been proposed for unique,

redundant and synergistic information [2]–[8], however their differences are only poorly understood (see [9] for a review).

With the development of these formal definitions, PIDs have begun to find uses in a wide variety of contexts. Many works, including ours, have used the PID to analyze biological and neural systems [11]–[19], e.g., to understand how neural activity and responses jointly encode stimuli. PID has also been used to characterize synergistic interactions in financial markets [20], and to define new measures of bias in machine learning [21]. Unfortunately, we find that only a few of these definitions have clear operational interpretations, and *these* definitions involve an optimization over the space of all joint distributions over (M, X, Y), making them hard to compute in practice. The absence of efficient methods to compute these PIDs, particularly for high dimensions, has been a bottleneck in their wider adoption.

Towards finding efficient ways of computing PIDs, this paper focuses on the bivariate PID for jointly Gaussian random vectors M, X and Y. Barrett [1] previously examined this problem for the case of *scalar messages* M, and showed that certain operationally well-motivated PID definitions can be reduced to a PID that has a *closed form* expression in the Gaussian case (Section II). This has enabled the efficient computation of these operationally sound PIDs for Gaussians with scalar M. However, the case of *vector* M has remained unresolved in the PID literature.

This paper has two main contributions concerning the bivariate PID for fully multivariate Gaussian random variables:[1] (1) In Section III-A, we show that for *vector* M, Barrett's result (and hence the reduction to closed-form) does not apply in general. We introduce a key concept underpinning Barrett's result called *Blackwell sufficiency* [22], using which we characterize the condition under which Barrett's result extends to fully multivariate Gaussians. (2) For cases where this condition *does not* hold, we provide a convex optimization framework to efficiently approximate a PID definition that is based on the statistical concept of *deficiency* [23] (Section III-B). We also provide results showing how this approximation bounds the true Gaussian PID and specific cases when it is guaranteed

[1]We use the term "fully *multivariate* Gaussian" here to refer to instances of a *bivariate* PID where M may also be a *vector* (in contrast to Barrett's work). This is not to be confused with *multivariate PIDs*, where information about M is decomposed among more than two variables.

to agree. In Section IV, we present numerical experiments that validate and go beyond our analytical results. Finally, we conclude with a discussion of open questions in Section V.

II. INTRODUCTION TO PID AND BARRETT'S RESULT

Let M, X and Y be random variables with sample spaces M, X and Y respectively, and joint density P_{MXY}. Then, a bivariate Partial Information Decomposition is loosely defined as a set of non-negative functions $UI(M : X \setminus Y)$, $UI(M : Y \setminus X)$, $RI(M : X; Y)$ and $SI(M : X; Y)$ that satisfy:

$$I\big(M; (X, Y)\big) = UI(M : X \backslash Y) + UI(M : Y \backslash X) \quad (1)$$
$$+ RI(M : X; Y) + SI(M : X; Y),$$
$$I(M; X) = UI(M : X \backslash Y) + RI(M : X; Y), \quad (2)$$
$$I(M; Y) = UI(M : Y \backslash X) + RI(M : X; Y). \quad (3)$$

Here, $UI(M : X \setminus Y)$ represents the information about M uniquely[2] present in X and not in Y, while $RI(M : X; Y)$ and $SI(M : X; Y)$ represent the redundant and synergistic information about M contained between X and Y. Usually, we also require that RI and SI are symmetric in X and Y (e.g., see axioms in [2], [3]).

For brevity, we may also write the terms in the RHS of (1) as UI_X, UI_Y, RI, and SI respectively. Since we have three equations and four unknowns, defining any one these four unknowns suffices to determine the other three.

We now state two PID definitions (which are non-identical, in general) in order to present the main result of Barrett [1]. Their interpretations are given later in Section III-A.

Definition 1 (MMI-PID [1]). *Let the redundant information about M present in both X and Y be the **M**inimum of their respective **M**utual **I**nformations with M (hence "MMI"):*

$$RI_{\mathrm{MMI}}(M : X; Y) \coloneqq \min\{I(M; X), I(M; Y)\}. \quad (4)$$

Along with equations (1)–(3), this fully determines the MMI-PID, i.e., $UI_{\mathrm{MMI}}(M : X \backslash Y)$, $UI_{\mathrm{MMI}}(M : Y \backslash X)$, and $SI_{\mathrm{MMI}}(M : X; Y)$ are now well-defined.

Remark 1. When P_{MXY} is jointly Gaussian, the MMI-PID can be written in closed form using the following well-known identity (derived in Appendix A for completeness):

$$I(M; Z) = \frac{1}{2} \log\left(\frac{\det(\Sigma_M)}{\det(\Sigma_M - \Sigma_{MZ}\Sigma_Z^{-1}\Sigma_{MZ}^{\mathsf{T}})}\right), \quad (5)$$

where Z could be X, Y, or their concatenation $[X^{\mathsf{T}}, Y^{\mathsf{T}}]^{\mathsf{T}}$, Σ_M and Σ_Z are the covariance matrices of M and Z, and Σ_{MZ} is their cross-covariance matrix. The closed-form expression follows by using (5) in (1)–(4).

Definition 2 ($\sim$-PID[3] [3]). *The unique information about M present in X and not in Y is given by*

$$\widetilde{UI}(M : X \setminus Y) \coloneqq \min_{Q \in \Delta_P} I_Q(M; X \mid Y), \quad (6)$$

where $\Delta_P \coloneqq \{Q_{MXY} : Q_{MX} = P_{MX}, Q_{MY} = P_{MY}\}$ and I_Q is the conditional mutual information over the joint distribution Q_{MXY}.

Theorem 1 (Barrett [1]). *If M, X and Y are jointly Gaussian and M is scalar, then the $\sim$-PID of Definition 2 reduces to the (simpler) MMI-PID of Definition 1.*[4]

Remark 2. Theorem 1 is significant because the $\sim$-PID is operationally well-motivated (as we will see), while the MMI-PID can be expressed in closed form for jointly Gaussian P_{MXY} as shown in Remark 1. This equivalence has therefore inspired usage of the MMI-PID in several studies that apply PIDs to real-world data [11]–[13], [20].

III. MAIN RESULTS

First we show that Barrett's result (Theorem 1) does not extend to vector M in general, so the closed-form expression from Remark 1 may not always apply.

Counterexample. Suppose $M = [M_1, M_2]$, $X = M_1 + Z_1$ and $Y = M_2 + Z_2$, with M_1, M_2, Z_1, $Z_2 \sim$ i.i.d. $\mathcal{N}(0, 1)$. Based on the intuition from the introduction, X and Y have equal amounts of *unique* information about M, and one can show that this holds for the $\sim$-PID. However, since $I(M; X) = I(M; Y)$, X and Y have only *redundant* information under the MMI-PID, and no unique information. Thus, the $\sim$-PID does not always reduce to the MMI-PID for vector M. $\square$

This counterexample is shown more formally in Appendix A.

A. When do Multivariate Gaussians have Closed-form PIDs?

To answer this question, we make some novel and important observations about Barrett's result by breaking it down into a few distinct steps:

A) Given *any* PID definition that satisfies the basic equations (1)–(3), if P_{MXY} is such that either $UI_X = 0$ or $UI_Y = 0$, then that PID reduces to the MMI-PID.
B) Barrett considers a few specific PID definitions (including the $\sim$-PID mentioned above). In fact, we can consider all PIDs that satisfy the following property: $UI_Y = 0$ if and only if X is *Blackwell sufficient* for M with respect to Y (formally defined below). We call such PIDs *Blackwellian*.[5]
C) When P_{MXY} is jointly Gaussian and M is scalar, we always have that either X is Blackwell sufficient for M w.r.t. Y, or Y is Blackwell sufficient w.r.t. X. However, this is not always the case when M is a vector.

Theorem 1 follows by reading these steps in reverse order. Thus, we observe that the concept of Blackwell sufficiency is the driving force in Barrett's result, although this is not explicitly recognized in that paper [1]. So we now proceed to formally define and understand Blackwell sufficiency.

Blackwell sufficiency. The idea of Blackwell sufficiency comes from the field of statistical decision theory [22], [24]. Suppose we are allowed to make inferences about M by observing either X or Y, but not both. Then, informally speaking, X is said to be *Blackwell sufficient* for M with

---

[2]Note that the $X \setminus Y$ in the argument of UI is purely notational, and does not represent a set difference. All quantities in (1) are functions of P_{MXY}.

[3]This PID is also sometimes referred to as the BROJA PID in the literature.

[4]Barrett's original theorem statement is actually slightly more general: all PIDs satisfying assumption ($*$) of Bertschinger et al. [3] reduce to the MMI-PID for jointly Gaussian P_{MXY} and scalar M. This includes a few other well-known definitions such as [2] and [4].

[5]Barrett actually considers all PIDs that satisfy Assumption ($*$) of [3]. While this is not incorrect, it is *incidental*, and considering Blackwellian PIDs is more precise. This difference is explained in Appendix B.

respect to Y if, on average, we can make equal or better inferences about M (w.r.t. *any* loss function) by observing X rather than Y. This forms a useful operational interpretation for Blackwellian PIDs, which satisfy $UI_Y = 0$ if and only if X is Blackwell sufficient for M with respect to Y. Thus, $UI_Y = 0$ if and only if Y can provide no better information than X for making inferences about M, which provides a concrete intuition for the meaning of *unique* information [3].

For simplicity, we define an equivalent [22] notion of Blackwell sufficiency, which is more amenable to our setup here. Given random variables M, X and Y with joint density P_{MXY}, Blackwell sufficiency is formally defined for the stochastic transformations $P_{X|M}$ and $P_{Y|M}$ (also called "channels" from M to X and from M to Y respectively):[6]

Definition 3 (Blackwell sufficiency: $\succcurlyeq_M$). *We say that a channel $P_{X|M}$ is* Blackwell sufficient *for another channel $P_{Y|M}$ (denoted $X \succcurlyeq_M Y$) if $\exists \, P_{Y'|X} \in \mathcal{C}(\mathsf{Y}|\mathsf{X})$ such that*

$$P_{Y'|X} \circ P_{X|M} = P_{Y|M}, \tag{7}$$

where $\mathcal{C}(\mathsf{Y}|\mathsf{X})$ is the set of all channels from X to Y, and $\circ$ represents channel composition, i.e. $\forall \, m \in \mathsf{M}, y \in \mathsf{Y}$,

$$(P_{Y'|X} \circ P_{X|M})(y|m) := \int P_{Y'|X}(y|x) P_{X|M}(x|m) \, dx. \tag{8}$$

Intuitively, $X \succcurlyeq_M Y$ means that we can generate a new random variable Y' from X (using the stochastic transformation $P_{Y'|X}$), so that the effective channel from M to Y' is equivalent to the original channel from M to Y (i.e., for any $M = m$, samples of Y' are statistically identical to samples of Y).

Definition 4 (Blackwellian PID). *We say that a bivariate PID of $I\big(M; (X,Y)\big)$ is* Blackwellian *if Y has zero unique information about M with respect to X if and only if $X \succcurlyeq_M Y$.*

Recalling our main objective: we wish to find efficient ways of computing the fully multivariate Gaussian PID. Based on step (C) in the list above, our first goal is to characterize the conditions under which $X \succcurlyeq_M Y$ for fully multivariate Gaussian P_{MXY}. Then, steps (B) and (A) would follow, ending with the closed form expression of the MMI-PID.

Characterizing Blackwell sufficiency for Gaussians. First, we parameterize the multivariate Gaussian distribution P_{MXY}. Let the sample spaces of M, X and Y respectively be $\mathsf{M} = \mathbb{R}^{d_M}$, $\mathsf{X} = \mathbb{R}^{d_X}$ and $\mathsf{Y} = \mathbb{R}^{d_Y}$. Let the joint distribution be given by $P_{MXY} = \mathcal{N}(0, \Sigma)$. Since Blackwell sufficiency ignores the dependence between X and Y conditional on M, we define:

$$M \sim P_M = \mathcal{N}(0, \Sigma_M) \tag{9}$$

$$X \mid M \sim P_{X|M} = \mathcal{N}(H_X M, \Sigma_{X|M}) \tag{10}$$

$$Y \mid M \sim P_{Y|M} = \mathcal{N}(H_Y M, \Sigma_{Y|M}) \tag{11}$$

Here, H_X, H_Y, $\Sigma_{X|M}$ and $\Sigma_{Y|M}$ can all be extracted from the full covariance matrix, Σ. Also, note that we *do not* assume that $X \perp\!\!\!\perp Y \mid M$; this dependence is captured in Σ.

Remark 3. Without loss of generality, we assume that $\mathbb{E}[M] = 0$. We also assume that the noise covariance matrices $\Sigma_{X|M}$ and $\Sigma_{Y|M}$ are full rank, which need not always be

true (e.g., in a noiseless channel). However, this assumption is required to keep mutual information values finite, as explained in Appendix A.

We can now characterize the conditions for Blackwell sufficiency in fully multivariate Gaussian distributions.

Theorem 2 (Blackwell Sufficiency for Multivariate Gaussians). *For jointly Gaussian random vectors M, X and Y, $X \succcurlyeq_M Y$ if and only if*

$$H_X^\mathsf{T} \Sigma_{X|M}^{-1} H_X \succcurlyeq H_Y^\mathsf{T} \Sigma_{Y|M}^{-1} H_Y, \tag{12}$$

where $A \succcurlyeq B$ means that $A - B$ is positive semidefinite (for positive semidefinite matrices A and B).

A proof of this theorem is provided in Appendix B. The result relies on a connection between concepts from two disparate fields: Blackwell sufficiency and the *stochastic degradedness* of broadcast channels.[7] We then leverage prior work from the literature analyzing the stochastic degradedness of MIMO Gaussian broadcast channels [26], [27] to prove our result. With this, we can formally state our extension of Barrett's result to fully multivariate Gaussians:

Corollary 3 (Extension of Barrett's Result). *If P_{MXY} is jointly Gaussian and satisfies (12), then any Blackwellian PID computed on P_{MXY} reduces to the MMI-PID and can be written in closed form as shown in Remark 1.*

Remark 4. Since $\Sigma_{X|M}$ and $\Sigma_{Y|M}$ are assumed full rank (see Remark 3), we can apply a whitening transform:

$$\widetilde{H}_X := \Sigma_{X|M}^{-\frac{1}{2}} H_X, \quad \widetilde{H}_Y := \Sigma_{Y|M}^{-\frac{1}{2}} H_Y. \tag{13}$$

This simplifies (12) to $\widetilde{H}_X^\mathsf{T} \widetilde{H}_X \succcurlyeq \widetilde{H}_Y^\mathsf{T} \widetilde{H}_Y$. We assume henceforth that H_X and H_Y have already been whitened, and hence $\Sigma_{X|M} = \Sigma_{Y|M} = I$.

B. Efficiently Computing the Gaussian PID in General

Next, we explore how the Gaussian PID can be computed (approximately) when P_{MXY} does *not* satisfy (12). For this, we consider another PID definition based on the statistical concept of *deficiency* [5], which we call the δ-PID.

Definition 5 (δ-PID [5]). *Let the (weighted output) deficiency[8] of X with respect to Y about M be defined as[9]*

$$\delta(M : Y \backslash X) := \inf_{P_{Y'|X} \, \in \, \mathcal{C}(\mathsf{Y}|\mathsf{X})} \mathbb{E}_{P_M}\big[D(P_{Y|M} \,\|\, P_{Y'|X} \circ P_{X|M})\big], \tag{14}$$

where $D(\cdot\|\cdot)$ is the KL-divergence. Then, define the deficiency-based redundant information about M in X and Y as:

$$RI_\delta(M : X; Y) := \min\big\{I(M; X) - \delta(M : X \backslash Y), \\ I(M; Y) - \delta(M : Y \backslash X)\big\}. \tag{15}$$

As with the MMI-PID, equations (1)–(3) fully determine the remaining components of the δ-PID—$UI_{\delta X}$, $UI_{\delta Y}$ and SI_δ.

---

[6]Note that Blackwell sufficiency ignores potential dependencies between X and Y given M, i.e., it does not depend upon $P_{XY|M}$ in its entirety; it only depends on the X- and Y-marginals of $P_{XY|M}$.

[7]This connection was mentioned in passing by Raginsky [25], but we prove the equivalence formally in Lemma 6 (see Appendix B). This connection has also never before been recognized in the context of PID.

[8]There are many ways to define deficiency; Raginsky [25] provides a number of these that consider the worst-case over M. We prefer an expectation over M, since M is a random variable in our setup.

[9]The reason for this notation is that the deficiency of X w.r.t. Y translates to the unique information present in Y and not in X.

Deficiency finds the channel $P_{Y'|X}$ that minimizes the expected divergence between $P_{Y|M}$ and $P_{Y'|X} \circ P_{X|M}$, where the expectation is over M. The divergence goes to zero if and only if $P_{Y|M} = P_{Y'|X} \circ P_{X|M}$, i.e. if $X \succcurlyeq_M Y$, thus the δ-PID is also Blackwellian [24]. Thus, deficiency directly measures how far P_{MXY} is from being Blackwell *sufficient* (as stated in (7)), and therefore forms a natural measure of unique information. Equation (15) plays the role of symmetrizing the redundant information, since $I(M;X) - \delta(M : X \setminus Y)$ and $I(M;Y) - \delta(M : Y \setminus X)$ are not necessarily equal.

Computing the deficiency as given by (14) is challenging because it requires an optimization over all continuous conditional distributions $P_{Y'|X}$. So, we consider the restricted problem where $P_{Y'|X}$ lies in the set of linear additive Gaussian noise channels $\mathcal{C}_G(\mathsf{Y}\,|\,\mathsf{X}) \subset \mathcal{C}(\mathsf{Y}\,|\,\mathsf{X})$. Then, $P_{Y'|X}$ can be parameterized in terms of its channel gain and noise covariance matrices, $T \in \mathbb{R}^{d_Y \times d_X}$ and $\Sigma_T \in \mathbb{R}^{d_Y \times d_Y}$, $\Sigma_T \succcurlyeq 0$, as $P_{Y'|X} \sim \mathcal{N}(TX, \Sigma_T)$. Thus, the *Gaussian deficiency* can be defined as follows:

$$\delta_G(M : Y \setminus X)$$
$$:= \inf_{P_{Y'|X} \in \mathcal{C}_G(\mathsf{Y}|\mathsf{X})} \mathbb{E}_{P_M}\left[D(P_{Y|M} \,\|\, P_{Y'|X} \circ P_{X|M}) \right] \quad (16)$$
$$= \inf_{T,\Sigma_T \succcurlyeq 0} \frac{1}{2}\Bigg[\, \mathbb{E}_{P_M}\left[\left\| (TH_X - H_Y)M \right\|_{\Sigma_T + TT^\mathsf{T}}^2 \right] \quad (17)$$
$$+ \operatorname{Tr}\left\{ (\Sigma_T + TT^\mathsf{T})^{-1} \right\}$$
$$+ \log \det(\Sigma_T + TT^\mathsf{T}) - d_Y \Bigg],$$

where $\|a\|_B^2 := aB^{-1}a^\mathsf{T}$ is the squared Mahalanobis distance, and (17) follows from the expression for the KL-divergence of normal distributions (derived in Appendix A) and using Remark 4. Unfortunately, we cannot prove that this problem is convex, since $\operatorname{Tr}\{(\Sigma_T + TT^\mathsf{T})^{-1}\}$ is the composition of a convex function ($\operatorname{Tr}\{(\cdot)^{-1}\}$) with a non-monotonic function of T. Therefore, we propose a convex approximation of (17) to find an *approximate* minimizer $\widehat{P}_{Y'|X} = \mathcal{N}(\widehat{T}X, \widehat{\Sigma}_T)$:

Definition 6 ($\widehat{\delta}_G$ and the $\widehat{\delta}_G$-PID)**.** *Let*
$$\widehat{T} := \operatorname*{argmin}_{T} \mathbb{E}_{P_M}\left[\left\| (TH_X - H_Y)M \right\|_{I + H_Y \Sigma_M H_Y^\mathsf{T}}^2 \right] \quad (18)$$
$$\text{s.t.} \quad I + H_Y \Sigma_M H_Y^\mathsf{T} - T(I + H_X \Sigma_M H_X^\mathsf{T})T^\mathsf{T} \succcurlyeq 0$$

$$\widehat{\Sigma}_T := I + H_Y \Sigma_M H_Y^\mathsf{T} - \widehat{T}(I + H_X \Sigma_M H_X^\mathsf{T})\widehat{T}^\mathsf{T} \quad (19)$$
Then, we define the approximate deficiency to be:
$$\widehat{\delta}_G(M : Y \setminus X) := \frac{1}{2}\Bigg[\mathbb{E}_{P_M}\left[\left\| (\widehat{T}H_X - H_Y)M \right\|_{\widehat{\Sigma}_T + \widehat{T}\widehat{T}^\mathsf{T}}^2 \right]$$
$$+ \operatorname{Tr}\left\{ (\widehat{\Sigma}_T + \widehat{T}\widehat{T}^\mathsf{T})^{-1} \right\}$$
$$+ \log \det(\widehat{\Sigma}_T + \widehat{T}\widehat{T}^\mathsf{T}) - d_Y \Bigg] \quad (20)$$

Substituting $\widehat{\delta}_G$ into (15), we obtain an approximation of RI_δ, and hence of the δ-PID, which we call the $\widehat{\delta}_G$-PID.

For brevity, we use $\widehat{UI}_X$, $\widehat{UI}_Y$, $\widehat{RI}$ and $\widehat{SI}$ to refer to the constituent atoms of the $\widehat{\delta}_G$-PID.

The derivation of the convex formulation in (18) has three main steps, which we summarize here and describe in detail in Appendix C.

1) First, we obtain a condition on Σ_T (in terms of T), which is locally optimal. This condition allows us to reduce the optimization problem from two variables (Σ_T and T) to just one (T alone). This step is *exact*, and is shown in Proposition 8 in Appendix C.
2) Next, we reinterpret and approximate the objective to significantly simplify its functional form, while attempting to minimize the same entities.
3) Lastly, as part of reducing the optimization problem to a single variable, the constraint $\Sigma_T \succcurlyeq 0$ is now replaced by a positive semidefiniteness constraint in terms of T. We approximate and simplify this constraint to make it more amenable for a convex program to handle, which yields the condition shown in (19).

This approximation can be shown to satisfy certain desirable properties. Firstly, $\widehat{\delta}_G$ is well defined, i.e., $\widehat{\Sigma}_T + \widehat{T}\widehat{T}^\mathsf{T}$ is guaranteed to be invertible when using the approximation of Definition 6. This is formally stated and proved in Lemma 9, which can be found in Appendix D.

Secondly, the estimate in (20) provably coincides with the true deficiency in the following circumstances:

Proposition 4. *For jointly Gaussian P_{MXY},*
$$\widehat{\delta}_G(M : Y \setminus X) = 0 \Leftrightarrow \delta(M : Y \setminus X) = 0 \quad (21)$$
$$\delta(M:Y\setminus X) = I(M;Y) \Rightarrow \widehat{\delta}_G(M:Y\setminus X) = I(M;Y) \quad (22)$$
A proof is given in Appendix E.

Thirdly, since $\widehat{P}_{Y'|X} \in \mathcal{C}_G(\mathsf{Y}\,|\,\mathsf{X}) \subset \mathcal{C}(\mathsf{Y}\,|\,\mathsf{X})$, we have:
$$\widehat{\delta}_G(M : Y \setminus X) \geq \delta_G(M : Y \setminus X) \geq \delta(M : Y \setminus X) \quad (23)$$
These inequalities show that the $\widehat{\delta}_G$-PID bounds the true δ-PID:

Proposition 5. *For jointly Gaussian random vectors M, X and Y, the $\widehat{\delta}_G$-PID bounds the δ-PID:*
$$\begin{aligned} \widehat{UI}_X &\geq UI_{\delta X} & \widehat{RI} &\leq RI_\delta \\ \widehat{UI}_Y &\geq UI_{\delta Y} & \widehat{SI} &\leq SI_\delta \end{aligned} \quad (24)$$

IV. EMPIRICAL RESULTS

Having established and theoretically justified a framework for approximating the δ-PID, we address three questions using simulations: (Q1) Does $\widehat{\delta}_G$ yield a *non-negative* PID for jointly Gaussian distributions? (Q2) Is the $\widehat{\delta}_G$-PID consistent with the results of Barrett [1] for the case of univariate M? (Q3) Does the $\widehat{\delta}_G$-PID meet our intuition for different d_M, d_X and d_Y? (Q4) How good is the approximation provided by the $\widehat{\delta}_G$-PID?

To address these questions, we sampled 80,000 joint covariance matrices $\Sigma \in \mathbb{R}^{d \times d}$ ($d = d_M + d_X + d_Y$) from a standard Wishart distribution, which fully characterized P_{MXY}. We sampled 20,000 matrices each from four sampling schemes:
(S1) $d_M \sim \operatorname{Unif}\{1 \ldots 10\}$ and $d_X = d_Y = d_M$
(S2) $d_M \sim \operatorname{Unif}\{1 \ldots 9\}$ and $d_X, d_Y \overset{\text{iid}}{\sim} \operatorname{Unif}\{d_M + 1 \ldots 10\}$
(S3) $d_M \sim \operatorname{Unif}\{2 \ldots 10\}$ and $d_X, d_Y \overset{\text{iid}}{\sim} \operatorname{Unif}\{1 \ldots d_M - 1\}$
(S4) $d_M \sim \operatorname{Unif}\{2 \ldots 9\}$, $d_X \sim \operatorname{Unif}\{1 \ldots d_M - 1\}$, and $d_Y \sim \operatorname{Unif}\{d_M + 1 \ldots 10\}$
Without loss of generality, we took $d_X \leq d_Y$ (i.e., we switched their values if $d_Y < d_X$). Simulations were performed on a 2.4 GHz 8-Core Intel Core i9 processor and took 188.8 minutes to

complete. Details of the experimental setup and implementation are provided in Appendices F and G and the code is available on Github [28].

First, we found that all estimated PID components were non-negative for *every one* of the 80,000 sampled distributions. This suggests an affirmative answer to (Q1). Secondly, we found that in every distribution with $d_M = 1$ ($N = 4277$), only one of X or Y had unique information. Therefore, the $\widehat{\delta}_G$-PID is consistent with Barrett's result in our experiments, answering (Q2) in the affirmative. Of the remaining 75,723 distributions for which $d_M > 1$, X and Y *both* had unique information in 93.6% of cases (70,876 distributions). Thus, the overwhelming majority of Gaussian distributions with vector M *do not* satisfy Blackwell sufficiency and do not reduce to the MMI-PID, justifying the need for our approximation-based approach.

To address (Q3), we visualize the distribution of unique, redundant, and synergistic information across all four sampling schemes in Figure 1. Since the scale of the PID components varies with $I(M; (X, Y))$, we consider the normalized PID quantities $\overline{UI}_X$, $\overline{UI}_Y$, $\overline{RI}$, and $\overline{SI}$ obtained by dividing each PID component by $I(M; (X, Y))$. These normalized values are all non-negative and satisfy $\overline{UI}_X + \overline{UI}_Y + \overline{RI} + \overline{SI} = 1$. In Figure 1 we represent each distribution by its location on a 3-simplex that characterizes the proportion of $I(M; (X, Y))$ that is accounted for by each PID component. Below the 3-simplexes for each sampling scheme, we also show the distribution of $\overline{UI}_X$, $\overline{UI}_Y$, $\overline{RI}$, and $\overline{SI}$ in the form of box plots.

Figure 1 shows how our approximate PID for multivariate Gaussians meets our intuitive expectations. Firstly, since $d_Y \geq d_X$, Y tends to have more unique information than X (see box plots), excepting (S1) where $d_X = d_Y$. Secondly, (S2) with $d_M < d_X, d_Y$ closely mimics the scalar-M case, rarely exhibiting unique information in both X and Y simultaneously. This is seen in the simplex plot, wherein the isosceles triangle forming the lower third of the simplex (with vertices at $\overline{UI}_X$, $\overline{UI}_Y$ and the centroid of the simplex) is almost completely devoid of points. Thirdly, (S3) and (S4) have large amounts of unique information in Y (as seen in the box plots), since d_M is large and provides more dimensions of M that Y can uniquely capture.

Synergy has a strong prevalence under all four sampling schemes (see box plots), with the greatest prevalence in (S2), when $d_M < d_X, d_Y$. This agrees with the intuition provided by Barrett [1], who mentions the prevalence of synergy for Gaussians with $d_M = 1$. Redundancy is never particularly prevalent—this may be due to how we sample random covariance matrices, which reduces the likelihood that X and Y capture the same dimensions of M. (S3) and (S4) have the least redundancy (as observed in the box plots), since X has fewer dimensions than M in both these cases.

To address (Q4), unfortunately, there are no alternative estimators of the δ-PID that can serve as potential baselines. Nor are there alternative estimators of Blackwellian PIDs for Gaussian variables (excepting the scalar-M case, for which Proposition 4 already provides guarantees). Therefore, we
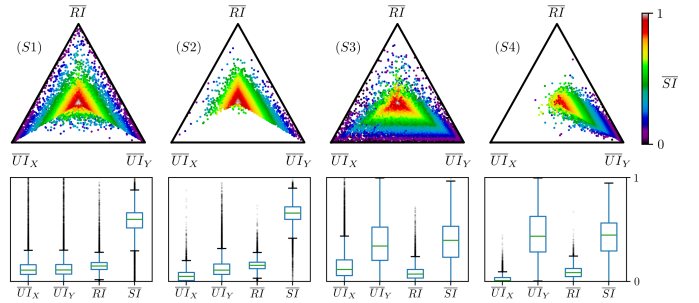

Fig. 1: Distribution of normalized unique, redundant, and synergistic information for Gaussian systems sampled from a standard Wishart distribution: (S1) $d_M = d_X = d_Y$; (S2) $d_M < d_X \leq d_Y$; (S3) $d_X \leq d_Y < d_M$; and (S4) $d_X < d_M < d_Y$. Top row: scatter plots of the computed $\overline{UI}_X$, $\overline{UI}_Y$, $\overline{RI}$ and $\overline{SI}$ on the 3-simplex (3D views in Appendix I). Each point is a single sampled Gaussian system. Bottom row: box plots showing the relative prevalence of each PID atom. The box shows the median and the first and third quartiles, while the whiskers extend to 1.5 times the interquartile range (difference of third and first quartiles).
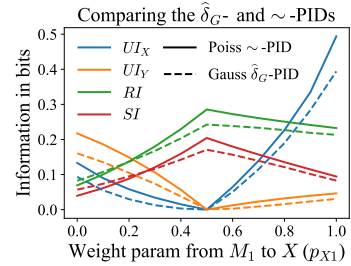

Fig. 2: A comparison of the $\sim$-PID evaluated [23] on a multivariate Poisson distribution against the $\widehat{\delta}_G$-PID evaluated on the covariance matrix of the same distribution. We take $M = [M_1, M_2]$, with $M_1, M_2 \sim$ iid. Poiss(2) and $X \sim \text{Binom}(M_1, p_{X1}) + \text{Binom}(M_2, p_{X2}) + \text{Poiss}(1)$. Y is defined similar to X. p_{X1} is varied between 0.0 and 1.0 in increments of 0.1, while p_{X2}, p_{Y1} and p_{Y2} are kept fixed at 0.5. Despite differences in the PID definitions, errors in approximating the Poisson distribution as a Gaussian, and errors in the two estimates, there is a reasonable agreement between the two PIDs.

compare the $\widehat{\delta}_G$-PID with the best available estimator [23] for the $\sim$-PID. These results are explained in the caption of Figure 2, with further details in Appendix H.

V. SUMMARY AND OPEN QUESTIONS

Barrett [1] showed that a Blackwellian PID (the $\sim$-PID) can be easily computed for Gaussian P_{MXY} with scalar M. Theorem 2 shows that this extends to vector M only under a specific condition, which likely holds only for a small minority of Gaussian distributions. Therefore, we provided a convex optimization framework to approximate the δ-PID, to enable applications of the PID on multivariate Gaussian data.

Our framework forms the first efficient method for computing Blackwellian PIDs (even approximately) for high-dimensional distributions. Although we have not theoretically proved that the $\widehat{\delta}_G$-PID is non-negative, we found no negative instances empirically. Several open questions remain: Can deficiency be computed exactly? When does our approximation agree with the true deficiency (beyond Proposition 4)? Can we characterize Blackwell sufficiency for general probability distributions (e.g., see [29])? Finally, applying the $\widehat{\delta}_G$-PID in practice also requires studying its statistical properties when estimating covariance matrices from data and providing statements of confidence.

REFERENCES

[1] A. B. Barrett, "Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems," *Physical Review E*, vol. 91, no. 5, p. 052802, 2015.

[2] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," *arXiv preprint arXiv:1004.2515*, 2010.

[3] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, "Quantifying unique information," *Entropy*, vol. 16, no. 4, pp. 2161–2183, 2014.

[4] M. Harder, C. Salge, and D. Polani, "Bivariate measure of redundant information," *Physical Review E*, vol. 87, no. 1, p. 012130, 2013.

[5] P. K. Banerjee, E. Olbrich, J. Jost, and J. Rauh, "Unique informations and deficiencies," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2018, pp. 32–38.

[6] X. Niu and C. J. Quinn, "A measure of synergy, redundancy, and unique information using information geometry," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 3127–3131.

[7] C. Finn and J. T. Lizier, "Pointwise partial information decomposition using the specificity and ambiguity lattices," *Entropy*, vol. 20, no. 4, p. 297, 2018.

[8] J. W. Kay and R. A. Ince, "Exact partial information decompositions for gaussian systems based on dependency constraints," *Entropy*, vol. 20, no. 4, p. 240, 2018.

[9] J. T. Lizier, N. Bertschinger, J. Jost, and M. Wibral, "Information decomposition of target effects from multi-source interactions: perspectives on previous, current and future work," *Entropy*, vol. 20, no. 4, p. 307, 2018.

[10] Full version of this paper with appendices. [Online]. Available: https://praveenv253.github.io/assets/doc/papers/2022--isit--full-paper.pdf

[11] N. Colenbier, F. Van de Steen, L. Q. Uddin, R. A. Poldrack, V. D. Calhoun, and D. Marinazzo, "Disambiguating the role of blood flow and global signal with partial information decomposition," *Neuroimage*, vol. 213, p. 116699, 2020.

[12] T. W. Boonstra, L. Faes, J. N. Kerkman, and D. Marinazzo, "Information decomposition of multichannel emg to map functional interactions in the distributed motor system," *NeuroImage*, vol. 202, p. 116093, 2019.

[13] J. Krohova, L. Faes, B. Czippelova, Z. Turianikova, N. Mazgutova, R. Pernice, A. Busacca, D. Marinazzo, S. Stramaglia, and M. Javorka, "Multiscale information decomposition dissects control mechanisms of heart rate variability at rest and during physiological stress," *Entropy*, vol. 21, no. 5, p. 526, 2019.

[14] N. M. Timme and C. Lapish, "A tutorial for information theory in neuroscience," *eneuro*, vol. 5, no. 3, 2018.

[15] G. Pica, E. Piasini, H. Safaai, C. Runyan, C. Harvey, M. Diamond, C. Kayser, T. Fellin, and S. Panzeri, "Quantifying how much sensory information in a neural code is relevant for behavior," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[16] I. Gat and N. Tishby, "Synergy and redundancy among brain cells of behaving monkeys," *Advances in neural information processing systems*, pp. 111–117, 1999.

[17] E. Schneidman, W. Bialek, and M. J. Berry, "Synergy, redundancy, and independence in population codes," *Journal of Neuroscience*, vol. 23, no. 37, pp. 11 539–11 553, 2003.

[18] N. Brenner, S. P. Strong, R. Koberle, W. Bialek, and R. R. d. R. v. Steveninck, "Synergy in a neural code," *Neural computation*, vol. 12, no. 7, pp. 1531–1552, 2000.

[19] P. Venkatesh, S. Dutta, and P. Grover, "Information flow in computational systems," *IEEE Transactions on Information Theory*, vol. 66, no. 9, pp. 5456–5491, 2020.

[20] T. Scagliarini, L. Faes, D. Marinazzo, S. Stramaglia, and R. N. Mantegna, "Synergistic information transfer in the global system of financial markets," *Entropy*, vol. 22, no. 9, p. 1000, 2020.

[21] S. Dutta, P. Venkatesh, P. Mardziel, A. Datta, and P. Grover, "An information-theoretic quantification of discrimination with exempt features," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3825–3833.

[22] D. Blackwell, "Equivalent comparisons of experiments," *The Annals of Mathematical Statistics*, pp. 265–272, 1953.

[23] P. K. Banerjee, J. Rauh, and G. Montúfar, "Computing the unique information," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 141–145.

[24] E. Torgersen, *Comparison of statistical experiments*. Cambridge University Press, 1991, vol. 36.

[25] M. Raginsky, "Shannon meets Blackwell and Le Cam: Channels, codes, and statistical experiments," in *2011 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2011, pp. 1220–1224.

[26] L. Gerdes, M. Riemensberger, and W. Utschick, "On the equivalence of degraded Gaussian MIMO broadcast channels," in *WSA 2015; 19th International ITG Workshop on Smart Antennas*. VDE, 2015, pp. 1–5.

[27] X. Shang and H. V. Poor, "Noisy-interference sum-rate capacity for vector Gaussian interference channels," *IEEE transactions on information theory*, vol. 59, no. 1, pp. 132–153, 2012.

[28] Multivariate Gaussian PID package. [Online]. Available: https://github.com/gabeschamberg/mvar-gauss-pid

[29] A. Makur and Y. Polyanskiy, "Comparison of channels: Criteria for domination by a symmetric channel," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5704–5725, 2018.

[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[31] K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook*. Technical University of Denmark, 2012, version: November 15, 2012. [Online]. Available: http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html

[32] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[33] S. Laue, M. Mitterreiter, and J. Giesen, "Computing higher order derivatives of matrix and tensor expressions," *Advances in neural information processing systems*, vol. 31, 2018.

[34] ——, "A simple and efficient tensor calculus," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 4527–4534.

[35] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[36] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd, "A rewriting system for convex optimization problems," *Journal of Control and Decision*, vol. 5, no. 1, pp. 42–60, 2018.

[37] F. Zhang, *The Schur complement and its applications*. Springer Science & Business Media, 2006, vol. 4.

[38] B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd, "Conic optimization via operator splitting and homogeneous self-dual embedding," *Journal of Optimization Theory and Applications*, vol. 169, no. 3, pp. 1042–1068, June 2016. [Online]. Available: http://stanford.edu/~boyd/papers/scs.html

[39] ——, "SCS: Splitting conic solver, version 2.1.3," https://github.com/cvxgrp/scs, Nov. 2019.

Partial Information Decomposition via Deficiency for Multivariate Gaussians

Praveen Venkatesh and Gabriel Schamberg

APPENDIX A
MISCELLANEOUS DERIVATIONS

A. Derivation of the Closed-form Expression for the Gaussian MMI-PID from Remark 1

Remark 1 establishes a closed-form expression for the MMI-PID of Gaussian random variables. First, we derive the expression for mutual information between two jointly Gaussian random vectors, as shown in equation (5).

$$I(M; Z) \stackrel{(a)}{=} h(M) - h(M \mid Z) \tag{25}$$

$$h(M) \stackrel{(b)}{=} \frac{1}{2} \log\big((2\pi e)^{d_M} \det(\Sigma_M)\big) \tag{26}$$

$$h(M \mid Z) \stackrel{(c)}{=} \frac{1}{2} \log\big((2\pi e)^{d_M} \det(\Sigma_{M|Z})\big) \tag{27}$$

where (a) is the basic formula for mutual information [30, Ch. 2] and (b) is the formula for the differential entropy of a Gaussian random vector [30, Thm. 8.4.1]. Here, $h(\cdot)$ represents differential entropy, while Σ_M is the covariance matrix of M. In (c), we have used the fact that $P_{M|Z}$ is also a Gaussian distribution for jointly Gaussian P_{MZ}, so the differential entropy follows the same formula as in (b), with the difference that we use the *conditional* covariance matrix $\Sigma_{M|Z}$ corresponding to the conditional distribution $P_{M|Z}$.

Next, we observe that $\Sigma_{M|Z}$ can be written as:

$$\Sigma_{M|Z} = \Sigma_M - \Sigma_{MZ} \Sigma_Z^{-1} \Sigma_{MZ}^\mathsf{T} \tag{28}$$

where Σ_{MZ} is the cross-covariance matrix between M and Z (e.g., see [31, Sec. 8.1.3]).

Thus, the expression for mutual information reduces to

$$I(M; Z) = \frac{1}{2} \log\big((2\pi e)^{d_M} \det(\Sigma_M)\big) - \frac{1}{2} \log\big((2\pi e)^{d_M} \det(\Sigma_{M|Z})\big) \tag{29}$$

$$= \frac{1}{2} \log\left(\frac{\det(\Sigma_M)}{\det(\Sigma_{M|Z})} \right) \tag{30}$$

$$= \frac{1}{2} \log\left(\frac{\det(\Sigma_M)}{\det(\Sigma_M - \Sigma_{MZ} \Sigma_Z^{-1} \Sigma_{MZ}^\mathsf{T})} \right). \tag{31}$$

A different closed-form expression for the mutual information between two Gaussian random variables appears later, in Section B-B of the Appendix. That expression has the more familiar form of $1/2 \log(1 + SNR)$, but requires a more explicit channel parameterization, which we avoid here.

We can now write out the MMI-PID for jointly Gaussian P_{MXY} using the above expression. First, we individually substitute X and Y for Z in the RHS below, to evaluate the MMI-redundancy:

$$RI_{\text{MMI}}(M : X; Y) = \min\{I(M; X), I(M; Y)\}. \tag{32}$$

Then, we can compute the MMI-unique informations by subtracting out the redundancy from the respective mutual information:

$$UI_{\text{MMI}}(M : X \setminus Y) = I(M; X) - RI_{\text{MMI}}(M : X; Y), \tag{33}$$

$$UI_{\text{MMI}}(M : Y \setminus X) = I(M; Y) - RI_{\text{MMI}}(M : X; Y). \tag{34}$$

And finally, the synergy can be computed by subtracting each of the aforementioned terms from the total mutual information:

$$SI_{\text{MMI}}(M : X; Y) = I(M; X, Y) - UI_{\text{MMI}}(M : X \setminus Y) - UI_{\text{MMI}}(M : Y \setminus X) - RI_{\text{MMI}}(M : X; Y). \tag{35}$$

Here, the total mutual information $I(M; X, Y)$ is computed by substituting the concatenated vector, $[X^\mathsf{T}, Y^\mathsf{T}]^\mathsf{T}$, in place of Z in (31).

B. Details of the Counterexample to Barrett's Result for Vector M

Recall the objective of the counterexample: we wish to show that when M is a vector, the $\sim$-PID of Definition 2 does not always reduce to the MMI-PID of Definition 1. Consider again the setup of the counterexample: $M = [M_1, M_2]$, $X = M_1 + Z_1$ and $Y = M_2 + Z_2$, with $M_1, M_2, Z_1, Z_2 \sim$ i.i.d. $\mathcal{N}(0,1)$.

First, we show what the MMI-PID looks like in this case:

$$I(M; X) = I(M_1, M_2; X) = I(M_1; X) + I(M_2; X \mid M_1) \tag{36}$$
$$= I(M_1; X) + 0, \tag{37}$$

since $(X, M_1) \perp\!\!\!\perp M_2$. Similarly, $I(M; Y) = I(M_2; Y)$. Now, by symmetry, we have that

$$I(M_1; X) = I(M_1; M_1 + Z_1) \overset{(a)}{=} I(M_2; M_2 + Z_2) = I(M_2; Y), \tag{38}$$

where (a) follows from the fact that M_1, M_2, Z_1 and Z_2 are all independent and identically distributed. Thus,

$$RI_{\text{MMI}}(M : X; Y) = \min\{I(M; X), I(M; Y)\} = I(M_1; X) = I(M_2; Y). \tag{39}$$

Consequently, (2) and (3) imply that $UI_{\text{MMI}}(M : X \setminus Y) = 0$ and $UI_{\text{MMI}}(M : Y \setminus X) = 0$.

On the other hand, the $\sim$-PID satisfies:

$$X \perp\!\!\!\perp Y \quad \text{and} \quad X \perp\!\!\!\perp Y \mid M. \tag{40}$$

The latter follows because

$$I(X; Y \mid M) = I(M_1 + Z_1; M_2 + Z_2 \mid M_1, M_2) \tag{41}$$
$$= h(M_1 + Z_1 \mid M_1, M_2) - h(M_1 + Z_1 \mid M_1, M_2, M_2 + Z_2) \tag{42}$$
$$= h(Z_1) - h(Z_1) = 0 \tag{43}$$

Now, [3, Thm. 20] and [3, Lem. 21] state that when the two conditions in (40) hold, then $\widetilde{RI}(M : X; Y) = 0$. Therefore, it follows that $\widetilde{UI}_X$ and $\widetilde{UI}_Y$ are equal to $I(M; X)$ and $I(M; Y)$ respectively.

Thus, we see that the $\sim$-PID does not always reduce to the MMI-PID for vector M.

C. Explanation of Remark 3

The mutual information between two random variables is invariant under invertible (and measurable) transforms applied to each of the individual variables (see, e.g., [32, Appendix]). Changing the means of M, X and Y constitutes an invertible transformation and would not affect their mutual information. Thus, we can assume without loss of generality that $\mathbb{E}[M] = 0$.

Further, we explicitly make the assumption that $\Sigma_{X|M}$ and $\Sigma_{Y|M}$ are full rank in order to ensure that mutual informations remain finite. For example, if M and X are scalar, and X is the output of a unity-gain noiseless channel with M as input, then we would have $X = M$ and $\Sigma_{X|M} = \sigma_{X|M}^2 = 0$. Thus,

$$I(M; X) = h(M) - h(M \mid X) = h(M) - h(M \mid M) = \infty, \tag{44}$$

because the conditional differential entropy of M given itself is negative infinity. Intuitively, this happens because M is a continuous random variable which contains an infinite number of (non-repeating) unknown bits in its decimal expansion. "Giving" M tells us all of these bits, and infinitely reduces our uncertainty about M.

Extending this concept to vector channels, to keep mutual information finite, we require that $\Sigma_{X|M}$ and $\Sigma_{Y|M}$ must *not* be noiseless in any direction (i.e., $p_{X|M}$ and $p_{Y|M}$ cannot be degenerate), and hence must be full rank. Technically, it suffices that $\Sigma_{X|M}$ and $\Sigma_{Y|M}$ have no rank deficiency along *those* dimensions of X and Y that *interact* with M, i.e.,

$$v \in \mathbb{R}^{d_X} \text{ s.t. } \|v\| = 1, \, v^\mathsf{T} \Sigma_{X|M} v = 0 \quad \Rightarrow \quad v^\mathsf{T} X \perp\!\!\!\perp M. \tag{45}$$

However, for simplicity, we assume that $\Sigma_{X|M}$ and $\Sigma_{Y|M}$ are entirely full rank.

D. Derivation of the Expression for Gaussian Deficiency in Equation (17)

First, we derive the expression for the KL-divergence between two n-dimensional multivariate normal distributions. For simplicity, suppose the distributions are given by $p \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $q \sim \mathcal{N}(\mu_2, \Sigma_2)$. Then,

$$D(p(x) \| q(x)) = \int p(x) \log \frac{p(x)}{q(x)} dx = \mathbb{E}_p\left[\log \frac{p(X)}{q(X)}\right] \tag{46}$$

$$\overset{(a)}{=} \mathbb{E}_p\left[\log\left(\frac{1}{(2\pi)^{n/2} \det(\Sigma_1)^{1/2}} \exp\left(-\frac{1}{2}\|X - \mu_1\|_{\Sigma_1}^2\right)\right) - \log\left(\frac{1}{(2\pi)^{n/2} \det(\Sigma_2)^{1/2}} \exp\left(-\frac{1}{2}\|X - \mu_2\|_{\Sigma_2}^2\right)\right)\right] \tag{47}$$

$$\stackrel{(b)}{=} \mathbb{E}_p\left[\log\frac{\det(\Sigma_2)^{1/2}}{\det(\Sigma_1)^{1/2}} - \frac{1}{2}\|X-\mu_1\|^2_{\Sigma_1} + \frac{1}{2}\|X-\mu_2\|^2_{\Sigma_2}\right] \tag{48}$$

$$\stackrel{(c)}{=} \frac{1}{2}\mathbb{E}_p\left[\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - (X-\mu_1)^\mathsf{T}\Sigma_1^{-1}(X-\mu_1) + (X-\mu_2)^\mathsf{T}\Sigma_2^{-1}(X-\mu_2)\right] \tag{49}$$

$$\stackrel{(d)}{=} \frac{1}{2}\mathbb{E}_p\left[\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \mathrm{Tr}\{(X-\mu_1)^\mathsf{T}\Sigma_1^{-1}(X-\mu_1)\} + \mathrm{Tr}\{(X-\mu_2)^\mathsf{T}\Sigma_2^{-1}(X-\mu_2)\}\right] \tag{50}$$

$$\stackrel{(e)}{=} \frac{1}{2}\mathbb{E}_p\left[\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \mathrm{Tr}\{\Sigma_1^{-1}(X-\mu_1)(X-\mu_1)^\mathsf{T}\} + \mathrm{Tr}\{\Sigma_2^{-1}(X-\mu_2)(X-\mu_2)^\mathsf{T}\}\right] \tag{51}$$

$$\stackrel{(f)}{=} \frac{1}{2}\left[\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \mathrm{Tr}\{\Sigma_1^{-1}\mathbb{E}_p[(X-\mu_1)(X-\mu_1)^\mathsf{T}]\} + \mathrm{Tr}\{\Sigma_2^{-1}\mathbb{E}_p[(X-\mu_2)(X-\mu_2)^\mathsf{T}]\}\right] \tag{52}$$

$$\stackrel{(g)}{=} \frac{1}{2}\left[\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \mathrm{Tr}\{\Sigma_1^{-1}\Sigma_1\} + \mathrm{Tr}\{\Sigma_2^{-1}(\mathrm{Var}_p[X-\mu_2] + \mathbb{E}_p[X-\mu_2]\mathbb{E}_p[X-\mu_2]^\mathsf{T})\}\right] \tag{53}$$

$$\stackrel{(h)}{=} \frac{1}{2}\left[\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \mathrm{Tr}\{I\} + \mathrm{Tr}\{\Sigma_2^{-1}(\Sigma_1 + (\mu_1-\mu_2)(\mu_1-\mu_2)^\mathsf{T})\}\right] \tag{54}$$

$$\stackrel{(i)}{=} \frac{1}{2}\left[\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - n + \mathrm{Tr}\{\Sigma_2^{-1}\Sigma_1 + \Sigma_2^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)^\mathsf{T}\}\right] \tag{55}$$

$$\stackrel{(j)}{=} \frac{1}{2}\left[\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - n + \mathrm{Tr}\{\Sigma_2^{-1}\Sigma_1\} + (\mu_1-\mu_2)^\mathsf{T}\Sigma_2^{-1}(\mu_1-\mu_2)\}\right] \tag{56}$$

$$\stackrel{(k)}{=} \frac{1}{2}\left[\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - n + \mathrm{Tr}\{\Sigma_2^{-1}\Sigma_1\} + \|\mu_1-\mu_2\|_{\Sigma_2}\}\right] \tag{57}$$

where in the above steps, we have:

(a) Expanded out the Gaussian distribution
(b) Expanded out the logs and canceled terms
(c) Expanded the Mahalanobis distance, $\|a\|^2_B := aB^{-1}a^\mathsf{T}$
(d) Used the fact that a scalar is equal to the trace of itself
(e) Used the fact that the trace is invariant under cyclic permutations
(f) Used linearity of trace to take the expectation inside the operator
(g) Noted that $\mathbb{E}_p[(X-\mu_1)(X-\mu_1)^\mathsf{T}]$ is simply Σ_1, and used the formula for variance in the last term.
(h) Resolved the variance and expectation, taken with respect to p
(i) Used the fact that the trace of the identity matrix is equal to the dimension of the matrix, n
(j) Used the cyclic-permutation invariance of trace again
(k) Used the formula for Mahalanobis distance again

The expression for Gaussian deficiency follows by observing that $p = \mathcal{N}(H_Y M, \Sigma_{Y|M})$ and $q = \mathcal{N}(TH_X M, T\Sigma_{X|M}T^\mathsf{T}+\Sigma_T)$. Further, we make use of Remark 4 to set $\Sigma_{X|M} = I$ and $\Sigma_{Y|M} = I$, so that

$$\mu_1 = H_Y M \qquad\qquad \Sigma_1 = I \tag{58}$$
$$\mu_2 = TH_X M \qquad\qquad \Sigma_2 = \Sigma_T + TT^\mathsf{T} \tag{59}$$

Finally, substituting into (57), we get

$$\mathbb{E}_{P_M}\left[D(P_{Y|M} \| P_{Y'|X} \circ P_{X|M})\right]$$
$$= \frac{1}{2}\left[\mathbb{E}_{P_M}\left[\|(TH_X - H_Y)M\|^2_{\Sigma_T+TT^\mathsf{T}}\right] + \mathrm{Tr}\left\{(\Sigma_T + TT^\mathsf{T})^{-1}\right\} + \log\det(\Sigma_T + TT^\mathsf{T}) - d_Y\right], \tag{60}$$

where we have used the fact that $\det(I) = 1$ and n here is d_Y.

APPENDIX B
PROOF OF THEOREM 2

The proof of Theorem 2 consists of two steps:

1) First, we show that the concept of Blackwell sufficiency (Definition 3) is identical to another concept known as stochastic degradedness (defined formally below) in the literature on broadcast channels. We show this in Lemma 6.

2) Then, we leverage previous work from the literature characterizing the stochastic degradedness of Gaussian MIMO broadcast channels.

Finally, we provide an explanation of what is expressed in footnote 5 in a subsection at the end of this section of the Appendix.

A. The Equivalence of Blackwell Sufficiency and Stochastic Degradedness

Definition 7 (Stochastic degradedness). *We say that a channel $P_{Y|M}$ is stochastically degraded with respect to another channel $P_{X|M}$ if there exists a random variable X' with sample space X such that $P_{X'|M} = P_{X|M}$ and M—X'—Y is a Markov chain.*

Lemma 6 (Equivalence of Blackwell sufficiency and stochastic degradedness). *$P_{X|M}$ is Blackwell sufficient for $P_{Y|M}$ if and only if $P_{Y|M}$ is stochastically degraded with respect to $P_{X|M}$.*

Proof. Throughout this proof, we drop the arguments of probability distributions. When we equate two distributions, we mean that they are identical at all points in their shared domain.[10]

($\Leftarrow$) Suppose $P_{Y|M}$ is stochastically degraded w.r.t. $P_{X|M}$. Then, $\exists\, X'$ such that

$$P_{X'|M} = P_{X|M} \text{ and } P_{YX'|M} = P_{Y|X'} P_{X'|M}. \tag{61}$$

Let $P_{Y'|X}(y\,|\,x) := P_{Y|X'}(y\,|\,x)\ \forall\, x \in \mathsf{X},\ y \in \mathsf{Y}$. Then, (61) implies that

$$P_{Y'|X} P_{X|M} = P_{Y|X'} P_{X'|M}. \tag{62}$$

Therefore,

$$\int P_{Y'|X} P_{X|M}\, dx = \int P_{Y|X'} P_{X'|M}\, dx \tag{63}$$

$$= \int P_{YX'|M}\, dx \;=\; P_{Y|M}, \tag{64}$$

which proves that $P_{X|M}$ is Blackwell sufficient for $P_{Y|M}$.

($\Rightarrow$) Suppose $P_{X|M}$ is Blackwell sufficient for $P_{Y|M}$. Then, $\exists\, P_{Y'|X}$ such that

$$\int P_{Y'|X} P_{X|M}\, dx = P_{Y|M}. \tag{65}$$

In other words,

$$P_{Y'|M} = P_{Y|M} \text{ and hence } P_{Y'M} = P_{YM}. \tag{66}$$

Let X' be defined through a stochastic transformation of Y and M: $P_{X'|YM}(x|y,m) := P_{X|Y'M}(x|y,m)\ \forall\, x \in \mathsf{X},\ y \in \mathsf{Y},$ $m \in \mathsf{M}$. Then, using (66), we find

$$P_{X'YM} = P_{X'|YM} P_{YM} \tag{67}$$

$$= P_{X|Y'M} P_{Y'M} = P_{XY'M}. \tag{68}$$

This in turn implies

$$P_{X'Y|M} = P_{XY'|M} \stackrel{(a)}{=} P_{Y'|X} P_{X|M} \tag{69}$$

$$\stackrel{(b)}{=} P_{Y|X'} P_{X'|M}, \tag{70}$$

which proves that M—X'—Y is a Markov chain. In the above equation, (a) follows from how $P_{Y'|X}$ is defined, while (b) follows from (68). $\qquad\square$

Remark 5. The equivalence stated in Lemma 6 was also mentioned in passing by Raginsky [25], though without proof. On the other hand, Cover and Thomas [30, Sec. 15.6.2] *define* stochastic degradedness similar to how we have defined Blackwell sufficiency. However, since we refer to the work of Gerdes et al. [26] in what follows, we have used their definition of stochastic degradedness, and proved the equivalence formally for the sake of completeness.

---

[10]This can be extended to be more measure-theoretically accurate, to mean that they are equal μ-almost everywhere under some measure μ that is absolutely continuous with respect to the two distributions being equated.

B. Characterizing the Stochastic Degradedness of Broadcast Channels

Before proceeding to the proof of Theorem 2, we introduce a lemma borrowed from Shang and Poor [27, Lemma 5] that provides an equivalent characterization of the condition in Theorem 2.

Lemma 7. *The condition $H_X^\mathsf{T} \Sigma_{X|M}^{-1} H_X \succcurlyeq H_Y^\mathsf{T} \Sigma_{Y|M}^{-1} H_Y$ holds if and only if*

$$\exists\, T \quad s.t. \quad H_Y = T H_X \quad and \quad T \Sigma_{X|M} T^\mathsf{T} \preccurlyeq \Sigma_{Y|M}. \tag{71}$$

Proof. Consider the whitened form of the channels:

$$\widetilde{H}_X := \Sigma_{X|M}^{-\frac{1}{2}} H_X, \quad \widetilde{H}_Y := \Sigma_{Y|M}^{-\frac{1}{2}} H_Y, \tag{72}$$

as mentioned in Remark 4. Then, we need to show $\widetilde{H}_X^\mathsf{T} \widetilde{H}_X \succcurlyeq \widetilde{H}_Y^\mathsf{T} \widetilde{H}_Y$ if and only if

$$\exists\, T \quad \text{s.t.} \quad \widetilde{H}_Y = T \widetilde{H}_X \quad \text{and} \quad T T^\mathsf{T} \preccurlyeq I. \tag{73}$$

The remainder of the proof follows from [27, Lemma 5]. $\qquad\square$

Proof of Theorem 2. The statement of Theorem 2 specifies a necessary and sufficient condition for $X \succcurlyeq_M Y$ for multivariate Gaussian P_{MXY}. However, thanks to Lemma 6, the same condition is necessary and sufficient for $P_{Y|M}$ being stochastically degraded with respect to $P_{X|M}$, and so this is what we proceed to show. This proof is derived in large part from the work of Gerdes et al. [26].

($\Leftarrow$) Suppose that $H_X^\mathsf{T} \Sigma_{X|M}^{-1} H_X \succcurlyeq H_Y^\mathsf{T} \Sigma_{Y|M}^{-1} H_Y$. Then, by Lemma 7, $\exists\, T$ such that $H_Y = T H_X$ and $T \Sigma_{X|M} T^\mathsf{T} \preccurlyeq \Sigma_{Y|M}$. In other words, we may write

$$Y' = TX + N, \tag{74}$$

where $N \sim \mathcal{N}(0, \Sigma_T)$, with $\Sigma_T := \Sigma_{Y|M} - T \Sigma_{X|M} T^\mathsf{T}$ (the fact that $T \Sigma_{X|M} T^\mathsf{T} \preccurlyeq \Sigma_{Y|M}$ ensures that Σ_T is positive semidefinite, and hence a valid covariance matrix). Therefore, $\exists\, Y'$ generated by a stochastic transformation $P_{Y'|X} = \mathcal{N}(TX, \Sigma_T)$, such that

$$P_{Y'|M} = \mathcal{N}(T H_X M, T \Sigma_{X|M} T^\mathsf{T} + \Sigma_T) = \mathcal{N}(H_Y M, \Sigma_{Y|M}) = P_{Y|M}. \tag{75}$$

Hence, by Definition 3, $X \succcurlyeq_M Y$.

($\Rightarrow$) Next, suppose that $X \succcurlyeq_M Y$. Then, $\exists\, Y'$ generated by $P_{Y'|X}$, such that $P_{Y'|M} = P_{Y|M}$. Since M—X—Y' is a Markov chain,

$$I(M; X) \overset{(a)}{\geq} I(M; Y') \overset{(b)}{=} I(M; Y), \tag{76}$$

where (a) follows from the Data Processing Inequality [30, Ch. 2], and (b) follows because $P_{Y'M} = P_{YM}$. For a Gaussian channel $P_{X|M}$, the mutual information is given by [30, Thm. 8.4.1]

$$I(M; X) = h(X) - h(X \mid M) \tag{77}$$

$$= \frac{1}{2} \log \det\big((2\pi e)(\Sigma_{X|M} + H_X \Sigma_M H_X^\mathsf{T})\big) - \frac{1}{2} \log \det(2\pi e \Sigma_{X|M}) \tag{78}$$

$$= \frac{1}{2} \log \left(\frac{\det(\Sigma_{X|M} + H_X \Sigma_M H_X^\mathsf{T})}{\det(\Sigma_{X|M})} \right) \tag{79}$$

$$= \frac{1}{2} \log \left(\frac{\det(\Sigma_{X|M}) \det(I + \Sigma_{X|M}^{-1} H_X \Sigma_M H_X^\mathsf{T})}{\det(\Sigma_{X|M})} \right) \tag{80}$$

$$= \frac{1}{2} \log \det(I + \Sigma_{X|M}^{-1} H_X \Sigma_M H_X^\mathsf{T}). \tag{81}$$

Now, suppose for the sake of contradiction that $H_X^\mathsf{T} \Sigma_{X|M}^{-1} H_X \not\succcurlyeq H_Y^\mathsf{T} \Sigma_{Y|M}^{-1} H_Y$. Then, by the definition of positive semidefiniteness, $\exists\, c \in \mathbb{R}^{d_M}$ such that

$$c^\mathsf{T} H_X^\mathsf{T} \Sigma_{X|M}^{-1} H_X c < c^\mathsf{T} H_Y^\mathsf{T} \Sigma_{Y|M}^{-1} H_Y c. \tag{82}$$

Since $\log \det(I + AB) = \log \det(I + BA)$, $\log(\cdot)$ is an increasing function, and the determinant of a scalar is equal to itself, we have that

$$\log \det(1 + c^\mathsf{T} H_X^\mathsf{T} \Sigma_{X|M}^{-1} H_X c) < \log \det(1 + c^\mathsf{T} H_Y^\mathsf{T} \Sigma_{Y|M}^{-1} H_Y c) \tag{83}$$

$$\Rightarrow \log \det(I + \Sigma_{X|M}^{-1} H_X c c^\mathsf{T} H_X^\mathsf{T}) < \log \det(I + \Sigma_{Y|M}^{-1} H_Y c c^\mathsf{T} H_Y^\mathsf{T}) \tag{84}$$

Now, since $c c^\mathsf{T} \succcurlyeq 0$, it is a valid covariance matrix. So if we set $\Sigma_M := c c^\mathsf{T}$, we get

$$I(M; X) = \log \det(I + \Sigma_{X|M}^{-1} H_X \Sigma_M H_X^\mathsf{T}) \tag{85}$$

$$< \log \det(I + \Sigma_{Y|M}^{-1} H_Y \Sigma_M H_Y^\mathsf{T}) \tag{86}$$

$$= I(M;Y). \tag{87}$$

However, this contradicts (76), which holds no matter what Σ_M is. Therefore, we must have $H_X^\mathsf{T} \Sigma_{X|M}^{-1} H_X \succcurlyeq H_Y^\mathsf{T} \Sigma_{Y|M}^{-1} H_Y$. $\quad\square$

C. Explanation of Footnote 5: The Difference Between PIDs Satisfying Assumption ($*$) *and Blackwellian PIDs*

In footnote 5, we noted that Barrett's result applied to all PIDs satisfying Assumption ($*$) from Bertschinger et al. [3]. On the other hand, we consider all Blackwellian PIDs when showing the extension of Barrett's result. Here, we discuss the commonalities and differences between these two sets of PID definitions.

First, we formally restate Assumption ($*$) as given in Bertschinger et al. [3]:

$$UI(M : X\backslash Y) \text{ is a function of only } P_M, P_{X|M} \text{ and } P_{Y|M}, \tag{$*$}$$

as opposed to being a function of the complete joint distribution P_{MXY}. Barrett's main result showed that for jointly Gaussian P_{MXY} with scalar M, all PIDs that satisfied the above Assumption ($*$) reduced to the MMI-PID. However, we show a seemingly different result: Theorem 2 states that for fully multivariate Gaussian P_{MXY} satisfying equation (12), all *Blackwellian* PIDs reduce to the MMI-PID. A natural question that arises is: what is the relationship between Blackwellian PIDs and PIDs that satisfy Assumption ($*$)?

In fact, PIDs that satisfy Assumption ($*$) and PIDs that are Blackwellian form two distinct and unrelated sets. All Blackwellian PIDs need not satisfy Assumption ($*$), and all PIDs that satisfy Assumption ($*$) need not be Blackwellian.

By virtue of how it is defined, $\widetilde{UI}$ upper bounds all PIDs that satisfy Assumption ($*$) (see [3, Lem. 3]). Thus, if $UI^{(*)}$ is an arbitrary PID that satisfies Assumption ($*$), then whenever $\widetilde{UI}_X$ goes to zero, $UI_X^{(*)}$ must also go to zero. Crucially, however, the converse does not hold: $\widetilde{UI}_X > 0$ does *not* imply that $UI_X^{(*)}$ must also be *non*-zero.[11] This feature is what distinguishes PIDs satisfying Assumption ($*$) from Blackwellian PIDs, as explained below.

By definition, all Blackwellian PIDs have UI_X going to zero *if and only if* $Y \succcurlyeq_M X$, and thus they all have UI_X going to zero *together*. Since $\widetilde{UI}$ is Blackwellian [3], for any arbitrary Blackwellian PID UI^{Bw}, whenever $\widetilde{UI}$ goes to zero, UI^{Bw} also goes to zero, and further, whenever $\widetilde{UI}$ is non-zero, UI^{Bw} is also non-zero.

We can summarize these relationships as:

$$UI_X^{\mathsf{Bw}} = 0 \quad \Leftrightarrow \quad \widetilde{UI}_X = 0 \quad \Rightarrow \quad UI_X^{(*)} = 0 \tag{88}$$

Therefore, our extension of Barrett's result (Corollary 3) can also be stated for PIDs satisfying Assumption ($*$) rather than Blackwellian PIDs (since Corollary 3 is effectively a one-way implication). However, while Theorem 2 can also be written (for jointly Gaussian P_{MXY}) as

$$UI_Y^{\mathsf{Bw}} = 0 \quad \Leftrightarrow \quad H_X^\mathsf{T} \Sigma_{X|M}^{-1} H_X \succcurlyeq H_Y^\mathsf{T} \Sigma_{Y|M}^{-1} H_Y, \tag{89}$$

we cannot write the same with $UI_Y^{(*)} = 0$ in place of $UI_Y^{\mathsf{Bw}} = 0$, since the forward implication ($\Rightarrow$) would not always hold.

APPENDIX C
JUSTIFICATION FOR OUR CONVEX APPROXIMATION FORMULATION

First, recall the expression for Gaussian deficiency from equation (17), which we are trying to approximate through a convex objective:

$$\delta_G(M : Y\backslash X) = \inf_{T, \Sigma_T \succcurlyeq 0} \frac{1}{2}\Bigg[\mathbb{E}_{P_M}\Big[\big\| (TH_X - H_Y)M \big\|_{\Sigma_T + TT^\mathsf{T}}^2 \Big] \tag{90}$$
$$+ \operatorname{Tr}\big\{ (\Sigma_T + TT^\mathsf{T})^{-1} \big\} + \log \det(\Sigma_T + TT^\mathsf{T}) - d_Y \Bigg]$$

As stated in the main text, the derivation of our proposed convex approximation has three main steps:

1) First, we obtain a condition on Σ_T (in terms of T), which is locally optimal. This condition allows us to reduce the optimization problem from two variables (Σ_T and T) to just one (T alone). This step is *exact*, and is shown in Proposition 8.
2) Next, we reinterpret and approximate the objective to significantly simplify its functional form, while attempting to minimize the same entities. This objective is designed to specifically to ensure that the approximate deficiency recovers the true Gaussian deficiency at its extremal values (i.e. when δ_G is equal to zero, or equal to the mutual information).
3) Lastly, as part of reducing the optimization problem to a single variable, the constraint $\Sigma_T \succcurlyeq 0$ is now replaced by a positive semidefiniteness constraint in terms of T. We approximate and simplify this constraint to make it more amenable for a convex program to handle.

---

[11]A simple example might be the MMI-PID, which satisfies Assumption ($*$), and for which at least one of UI_X or UI_Y must be zero. However, there are always instances when $\widetilde{UI}_X > 0$ and $\widetilde{UI}_Y > 0$, such as the counterexample shown in Appendix A-B.

We first state Proposition 8, which encapsulates segments of the simplification process that are *exact*. The parts of the simplification that actually constitute approximations are described immediately afterward.

Proposition 8. *The Gaussian deficiency, as given by equation* (90)*, finds a local minimum at*

$$\Sigma_T = I - TT^\mathsf{T} + (H_Y - TH_X)\Sigma_M(H_Y - TH_X)^\mathsf{T}. \tag{91}$$

Substituting this expression back into the optimization problem significantly simplifies the expression for the Gaussian deficiency:

$$
\begin{aligned}
\delta_G(M : Y\backslash X) \;=\; &\inf_T \; \log\det(I + (H_Y - TH_X)\Sigma_M(H_Y - TH_X)^\mathsf{T}) \\
&\text{s.t.} \quad I - TT^\mathsf{T} + (H_Y - TH_X)\Sigma_M(H_Y - TH_X)^\mathsf{T} \succcurlyeq 0
\end{aligned} \tag{92}
$$

Proof. The expression in equation (90) can be simplified into a more manageable form, starting with the expectation:

$$
\begin{aligned}
\mathbb{E}_{P_M}&\left[\left\|(H_Y - TH_X)M\right\|^2_{\Sigma_T + TT^\mathsf{T}}\right] \\
&\overset{(a)}{=} \mathbb{E}_{P_M}\left[M^\mathsf{T}(H_Y - TH_X)^\mathsf{T}(\Sigma_T + TT^\mathsf{T})^{-1}(H_Y - TH_X)M\right] \tag{93}\\
&\overset{(b)}{=} \mathbb{E}_{P_M}\left[\mathrm{Tr}\{M^\mathsf{T}(H_Y - TH_X)^\mathsf{T}(\Sigma_T + TT^\mathsf{T})^{-1}(H_Y - TH_X)M\}\right] \tag{94}\\
&\overset{(c)}{=} \mathbb{E}_{P_M}\left[\mathrm{Tr}\{(\Sigma_T + TT^\mathsf{T})^{-1}(H_Y - TH_X)MM^\mathsf{T}(H_Y - TH_X)^\mathsf{T}\}\right] \tag{95}\\
&\overset{(d)}{=} \mathrm{Tr}\left\{(\Sigma_T + TT^\mathsf{T})^{-1}(H_Y - TH_X)\mathbb{E}_{P_M}\left[MM^\mathsf{T}\right](H_Y - TH_X)^\mathsf{T}\right\} \tag{96}\\
&\overset{(e)}{=} \mathrm{Tr}\left\{(\Sigma_T + TT^\mathsf{T})^{-1}(H_Y - TH_X)\Sigma_M(H_Y - TH_X)^\mathsf{T}\right\}, \tag{97}
\end{aligned}
$$

where in (a), we have simply expanded the Mahalanobis norm; in (b), we have used the fact that a scalar is equal to the trace of itself; in (c), we have relied on the property that the trace is invariant under cyclic permutations; in (d), we have moved the expectation inside, which is possible since all these operations are linear; and finally in (e), we have equated $\mathbb{E}[MM^\mathsf{T}]$ to Σ_M.

Now that we have resolved the expectation, we can substitute this expression back into (90), and simplify the expression further:

$$
\begin{aligned}
\delta_G(M : Y\backslash X) = \inf_{T,\Sigma_T \succcurlyeq 0} \frac{1}{2}\Bigg[&\mathrm{Tr}\Big\{(\Sigma_T + TT^\mathsf{T})^{-1}(H_Y - TH_X)\Sigma_M(H_Y - TH_X)^\mathsf{T}\Big\} \\
&+ \mathrm{Tr}\left\{(\Sigma_T + TT^\mathsf{T})^{-1}\right\} + \log\det(\Sigma_T + TT^\mathsf{T}) - d_Y \Bigg] \tag{98}\\
= \inf_{T,\Sigma_T \succcurlyeq 0} \frac{1}{2}\Bigg[&\mathrm{Tr}\Big\{(\Sigma_T + TT^\mathsf{T})^{-1}\big(I + (H_Y - TH_X)\Sigma_M(H_Y - TH_X)^\mathsf{T}\big)\Big\} \\
&- d_Y + \log\det(\Sigma_T + TT^\mathsf{T}) \Bigg] \tag{99}
\end{aligned}
$$

From this, it might already be intuitively apparent that we would like

$$(\Sigma_T + TT^\mathsf{T})^{-1}\big(I + (H_Y - TH_X)\Sigma_M(H_Y - TH_X)^\mathsf{T}\big) = I \tag{100}$$

$$\Rightarrow \quad \Sigma_T = I - TT^\mathsf{T} + (H_Y - TH_X)\Sigma_M(H_Y - TH_X)^\mathsf{T}, \tag{101}$$

since this would allow the trace term to perfectly cancel d_Y in (99) (as the matrix within the trace has dimension $d_Y \times d_Y$). However, it is possible to show that (101) in fact gives a local optimum for Σ_T. This can be done by examining the partial derivative of the objective in (99) with respect to Σ_T, which is what we proceed to do next.[12]

First, we lay out the following identities for matrices A and B (refer [31, Sec. 2, 2.4 and 9.7]):

$$
\begin{aligned}
\partial(AB) = (\partial A)B + A(\partial B) \qquad & \partial\,\mathrm{Tr}\{A\} = \mathrm{Tr}\{\partial A\} \qquad && \frac{\partial A}{\partial A_{ij}} = J^{ij} \\
\partial(A^{-1}) = -A^{-1}(\partial A)A^{-1} \qquad & \partial(\log\det A) = \mathrm{Tr}\{A^{-1}\partial A\} \qquad && \mathrm{Tr}\{AJ^{ij}\} = A^\mathsf{T}|_{ij} = A_{ji}
\end{aligned} \tag{102}
$$

where J^{ij} is the single-entry matrix, which contains 0's everywhere except at the (i, j)-th location, where it contains a 1. These identities apply only when taking derivatives with respect to matrices that have no special structure (e.g., symmetry). Since Σ_T

---

[12]The matrix derivatives in this section were computed by hand (as shown in the body of the text), but they were also checked against an online symbolic matrix differentiation tool, https://www.matrixcalculus.org/ [33], [34].

13

is a symmetric matrix, we make use of some additional results [31, Sec. 2.8.2]. For a symmetric matrix S and some scalar function f that depends on S,

$$\left.\frac{\partial f}{\partial S}\right|_S = \frac{\partial' f}{\partial' S} + {\frac{\partial' f}{\partial' S}}^{\mathsf{T}} - \frac{\partial' f}{\partial' S} \odot I \tag{103}$$

where $\partial'/\partial' S$ is used to denote that the partial derivative is taken *as if S had no structure* (so the identities in (102) may be applied), and $\odot$ represents the element-wise multiplication of matrices (also called a Hadamard product).

Using (102) and (103), we can write out the partial derivative of the objective in (99) with respect to Σ_T, and equate it to zero. To simplify notation, let $A := \Sigma_T + TT^\mathsf{T}$ and $B := I + (H_Y - TH_X)\Sigma_M(H_Y - TH_X)^\mathsf{T}$, so that the objective in (99) can be written as:

$$f(\Sigma_T) := \frac{1}{2}\Big[\mathrm{Tr}\{A^{-1}B\} - d_Y + \log\det(A)\Big] \tag{104}$$

We begin by taking a partial derivative of f with respect to the (i,j)-th element of Σ_T, which we denote $\Sigma_T|_{ij}$:

$$\left.\frac{\partial' f}{\partial' \Sigma_T|_{ij}}\right|_{\Sigma_T} \overset{(a)}{=} \frac{1}{2}\left[\frac{\partial'}{\partial' \Sigma_T|_{ij}}\mathrm{Tr}\{A^{-1}B\} + \frac{\partial'}{\partial' \Sigma_T|_{ij}}\log\det(A)\right]\Bigg|_{\Sigma_T} \tag{105}$$

$$\overset{(b)}{=} \frac{1}{2}\left[\mathrm{Tr}\Big\{\frac{\partial' A^{-1}}{\partial' \Sigma_T|_{ij}}B\Big\} + \mathrm{Tr}\Big\{A^{-1}\frac{\partial' A}{\partial' \Sigma_T|_{ij}}\Big\}\right]\Bigg|_{\Sigma_T} \tag{106}$$

$$\overset{(c)}{=} \frac{1}{2}\left[\mathrm{Tr}\Big\{-A^{-1}\frac{\partial' A}{\partial' \Sigma_T|_{ij}}A^{-1}B\Big\} + \mathrm{Tr}\Big\{A^{-1}\frac{\partial' A}{\partial' \Sigma_T|_{ij}}\Big\}\right]\Bigg|_{\Sigma_T} \tag{107}$$

$$\overset{(d)}{=} \frac{1}{2}\left[\mathrm{Tr}\Big\{-A^{-1}\frac{\partial' \Sigma_T}{\partial' \Sigma_T|_{ij}}A^{-1}B\Big\} + \mathrm{Tr}\Big\{A^{-1}\frac{\partial' \Sigma_T}{\partial' \Sigma_T|_{ij}}\Big\}\right]\Bigg|_{\Sigma_T} \tag{108}$$

$$\overset{(e)}{=} \frac{1}{2}\left[\mathrm{Tr}\{-A^{-1}J^{ij}A^{-1}B\} + \mathrm{Tr}\{A^{-1}J^{ij}\}\right] \tag{109}$$

$$\overset{(f)}{=} \frac{1}{2}\left[\mathrm{Tr}\{-A^{-1}BA^{-1}J^{ij}\} + \mathrm{Tr}\{A^{-1}J^{ij}\}\right] \tag{110}$$

$$\overset{(g)}{=} \frac{1}{2}\left[-\big(A^{-1}BA^{-1}\big)_{ij}^\mathsf{T} + \big(A^{-1}\big)_{ij}^\mathsf{T}\right] \tag{111}$$

$$\overset{(h)}{=} \frac{1}{2}\big(-A^{-1}BA^{-1} + A^{-1}\big)_{ij} \tag{112}$$

$$\Rightarrow \quad \left.\frac{\partial' f}{\partial' \Sigma_T}\right|_{\Sigma_T} \overset{(i)}{=} \frac{1}{2}\big(-A^{-1}BA^{-1} + A^{-1}\big), \tag{113}$$

where in the above labelled steps, we have:

(a) Taken the partial derivative of f w.r.t. $\Sigma_T|_{ij}$ *as if Σ_T had no special structure*, and evaluated it at Σ_T
(b) Used the identities in (102), noting that B is constant w.r.t. Σ_T
(c) Used the identities in (102)
(d) Used the fact that $A = \Sigma_T + TT^\mathsf{T}$, and that the partial derivative of T with respect to $\Sigma_T|_{ij}$ is zero
(e) Used the identities in (102)
(f) Made use of the fact that the trace is invariant under cyclic permutations of its argument
(g) Used the identities in (102)
(h) Relied on the fact that A and B are symmetric, so that $(A^{-1}BA^{-1})^\mathsf{T} = A^{-1}BA^{-1}$, and $A^{-\mathsf{T}} = A^{-1}$
(i) Assembled the full partial derivative with respect to the *matrix* Σ_T

Hence, we can write out the complete partial derivative of the objective in (99) with respect to Σ_T, and set it to zero:

$$0 = \left.\frac{\partial f}{\partial \Sigma_T}\right|_{\Sigma_T} = \left[\frac{\partial' f}{\partial' \Sigma_T} + {\frac{\partial' f}{\partial' \Sigma_T}}^\mathsf{T} - \frac{\partial' f}{\partial' \Sigma_T}\odot I\right]\Bigg|_{\Sigma_T} \tag{114}$$

$$= \frac{1}{2}\Big[\big(-A^{-1}BA^{-1} + A^{-1}\big) + \big(-A^{-1}BA^{-1} + A^{-1}\big)^\mathsf{T} - \big(-A^{-1}BA^{-1} + A^{-1}\big)\odot I\Big] \tag{115}$$

$$= \big(-A^{-1}BA^{-1} + A^{-1}\big) - \frac{1}{2}\big(-A^{-1}BA^{-1} + A^{-1}\big)\odot I, \tag{116}$$

where in the last step, we have again relied on the symmetry of A and B.

A sufficient condition for the above equation to hold is

$$-A^{-1}BA^{-1} + A^{-1} = 0 \tag{117}$$

$$\Rightarrow \quad A^{-1}BA^{-1} = A^{-1} \tag{118}$$

$$\Rightarrow \quad BA^{-1} = I \tag{119}$$

$$\Rightarrow \quad B = A \tag{120}$$

Substituting the original expressions for A and B,

$$\Sigma_T + TT^\mathsf{T} = I + (H_Y - TH_X)\Sigma_M(H_Y - TH_X)^\mathsf{T} \tag{121}$$

$$\Rightarrow \quad \Sigma_T = I - TT^\mathsf{T} + (H_Y - TH_X)\Sigma_M(H_Y - TH_X)^\mathsf{T} \tag{122}$$

Finally, we can substitute this expression back into the objective in (99) to cancel the $\mathrm{Tr}\{\cdot\}$ and d_Y terms, leaving only the $\log\det(\cdot)$ term:

$$\delta_G(M : Y \setminus X) = \inf_T \ \log\det(\Sigma_T + TT^\mathsf{T})$$
$$\text{s.t.} \quad I - TT^\mathsf{T} + (H_Y - TH_X)\Sigma_M(H_Y - TH_X)^\mathsf{T} \succcurlyeq 0 \tag{123}$$

Substituting for the Σ_T within the $\log\det(\cdot)$ yields the desired expression from the statement of the proposition. $\qquad\square$

This concludes the first step of simplifying our optimization problem, which reduces the problem to an optimization over a single variable, and which is exact. Next, we proceed to approximate the terms in this simplified formulation.

First, note that the objective seeks to minimize the log-determinant of $I + (H_Y - TH_X)\Sigma_M(H_Y - TH_X)^\mathsf{T}$. This term would naturally become smaller if H_Y is made closer to TH_X. To simplify this objective, we instead look at minimizing

$$\mathbb{E}_{P_M}\left[\left\|(H_Y - TH_X)M\right\|^2_{\Sigma_T + TT^\mathsf{T}}\right], \tag{124}$$

which appears in the original optimization problem. However, to avoid having the Mahalanobis norm being taken against $\Sigma_T + TT^\mathsf{T}$, which depends on T and yields a non-trivial objective function, we instead minimize:

$$\mathbb{E}_{P_M}\left[\left\|(H_Y - TH_X)M\right\|^2_{I + H_Y\Sigma_M H_Y^\mathsf{T}}\right]. \tag{125}$$

This objective function is chosen such that (124) and (125) will be equivalent "at the extremes," i.e. when the deficiency is zero ($H_Y - TH_X = 0$) and when then deficiency is the mutual information ($T = 0$). This choice also enables the proof of Proposition 4 (see Appendix E), which guarantees that when the true deficiency equals zero or the mutual information, the approximation yields the true deficiency. We use this revised objective to estimate the optimal *argument* $\widehat{T}$. To compute the approximate Gaussian deficiency, we substitute this value of $\widehat{T}$ back into the formula for the Gaussian deficiency, given by (17).

Finally, to arrive at the estimate of $\widehat{\Sigma}_T$ we employ in our approximation, given in equation (19), we assume that $H_Y - TH_X$ is small in one part of the equation:

$$\Sigma_T = I - TT^\mathsf{T} + (H_Y - TH_X)\Sigma_M(H_Y - TH_X)^\mathsf{T} \tag{126}$$

$$= I - TT^\mathsf{T} + H_Y\Sigma_M H_Y^\mathsf{T} - H_Y\Sigma_M H_X^\mathsf{T}T^\mathsf{T} - TH_X\Sigma_M H_Y^\mathsf{T} + TH_X\Sigma_M H_X^\mathsf{T}T^\mathsf{T} \tag{127}$$

$$\approx I - TT^\mathsf{T} + H_Y\Sigma_M H_Y^\mathsf{T} - TH_X\Sigma_M H_X^\mathsf{T}T^\mathsf{T} - TH_X\Sigma_M H_X^\mathsf{T}T^\mathsf{T} + TH_X\Sigma_M H_X^\mathsf{T}T^\mathsf{T} \tag{128}$$

$$= I - TT^\mathsf{T} + H_Y\Sigma_M H_Y^\mathsf{T} - TH_X\Sigma_M H_X^\mathsf{T}T^\mathsf{T} \tag{129}$$

$$= I + H_Y\Sigma_M H_Y^\mathsf{T} - T(I + H_X\Sigma_M H_X^\mathsf{T})T^\mathsf{T} \tag{130}$$

where the approximation occurs in (128) for the two terms that are linear in T. Since $\Sigma_T \succcurlyeq 0$ becomes a constraint in our eventual optimization problem, this approximation significantly simplifies the constraint: firstly, it removes the linear terms making this a pure quadratic form; secondly, the matrix appearing in the quadratic atom, $I + H_X\Sigma_M H_X^\mathsf{T}$, is now guaranteed to be invertible. Prior to the approximation, the same matrix was $I - H_X\Sigma_M H_X^\mathsf{T}$, which did not have the same guarantee. These simplifications enable using the Schur complement in formulating our constrained optimization problem, as detailed in Appendix F.

APPENDIX D
THE APPROXIMATE DEFICIENCY IS WELL DEFINED

Lemma 9. *For jointly Gaussian random vectors M, X and Y satisfying the assumption in Remark 3, $\widehat{T}$ and $\widehat{\Sigma}_T$ as defined by (19) and (18) yield an approximate deficiency $\widehat{\delta}_G(M : Y \setminus X)$ that is well defined, i.e. the covariance matrix $\widehat{\Sigma}_T + \widehat{T}\widehat{T}^\mathsf{T}$ for the composite channel $P_{Y'|X} \circ P_{X|M}$ is invertible.*

Proof. We know that $\widehat{\Sigma}_T \succcurlyeq 0$ by virtue of the constraint in (18) and thus $\widehat{\Sigma}_T + \widehat{T}\widehat{T}^\mathsf{T} \succcurlyeq 0$ as well. For ease of notation, let $A := H_Y\Sigma_M H_Y^\mathsf{T}$ and $B := H_X\Sigma_M H_X^\mathsf{T}$. Assume for a contradiction that $\widehat{\Sigma}_T + \widehat{T}\widehat{T}^\mathsf{T}$ is rank deficient. This implies that there exists a vector $v \neq \mathbf{0}$ such that:

$$v^\mathsf{T}(\widehat{\Sigma}_T + \widehat{T}\widehat{T}^\mathsf{T})v = v^\mathsf{T}(I + A - \widehat{T}B\widehat{T}^\mathsf{T})v = 0$$

$$\Rightarrow \quad v^\mathsf{T}(I + A)v = v^\mathsf{T}(\widehat{T}B\widehat{T}^\mathsf{T})v$$
$$\overset{(a)}{\Rightarrow} \quad (\widehat{T}^\mathsf{T}v)^\mathsf{T}B(\widehat{T}^\mathsf{T}v) > 0$$
$$\Rightarrow \quad \widehat{T}^\mathsf{T}v \neq \mathbf{0}$$

where (a) follows from $A = (H_Y\Sigma_M^{-\frac{1}{2}})(H_Y\Sigma_M^{-\frac{1}{2}})^\mathsf{T} \Rightarrow A \succcurlyeq 0$. But since $\widehat{\Sigma}_T \succcurlyeq 0$, we have:

$$v^\mathsf{T}\widehat{\Sigma}_T v \geq 0$$
$$\Rightarrow \quad v^\mathsf{T}(\widehat{\Sigma}_T + \widehat{T}\widehat{T}^\mathsf{T})v - v^\mathsf{T}(\widehat{T}\widehat{T}^\mathsf{T})v \geq 0$$
$$\Rightarrow \quad v^\mathsf{T}\widehat{T}\widehat{T}^\mathsf{T}v = \|\widehat{T}^\mathsf{T}v\|_2^2 \leq 0.$$

which is a contradiction since $\widehat{T}^\mathsf{T}v \neq \mathbf{0} \Rightarrow \|\widehat{T}^\mathsf{T}v\|_2^2 > 0$. $\qquad\square$

APPENDIX E
PROOF OF PROPOSITION 4

Proof of Proposition 4. For ease of notation, we omit the function arguments $(M : Y \setminus X)$ and refer simply to $\widehat{\delta}_G$, δ_G, and δ.

$(\delta = 0 \Rightarrow \widehat{\delta}_G = 0)$ It was shown by Torgersen [24, Theorem 8.2.13] that $\delta = 0 \Rightarrow \delta_G = 0$, thus it suffices to show that $\delta_G = 0 \Rightarrow \widehat{\delta}_G = 0$.

When δ_G is well-defined, we know that $\Sigma_T + TT^\mathsf{T} \succcurlyeq 0$ is invertible and thus $\Sigma_T + TT^\mathsf{T}, (\Sigma_T + TT^\mathsf{T})^{-1} \succ 0$. This implies that the first term of (17) is greater than zero unless $(TH_X - H_Y)M = 0$. Thus $\delta_G = 0$ implies that $(TH_X - H_Y)M = 0$ almost everywhere in M, i.e. that there exists a T^* such that $T^*H_X = H_Y$. Thus, we also know that the objective function in (18) is minimized by $T = T^*$. To show that T^* is in the feasible set, we note that $\delta_G = 0$ implies that there exists a $\Sigma_T^* \succcurlyeq 0$ such that $\Sigma_T^* + T^*T^{*\mathsf{T}} = I \Rightarrow I - T^*T^{*\mathsf{T}} \succcurlyeq 0$. As such, when $T = T^*$ we have that $\Sigma_T = I + H_Y\Sigma_M H_Y^\mathsf{T} - T^*(I + H_X\Sigma_M H_X^\mathsf{T})T^{*\mathsf{T}} = I - T^*T^{*\mathsf{T}} = \Sigma_T^*$.

$(\delta = 0 \Leftarrow \widehat{\delta}_G = 0)$ This follows from $\widehat{\delta}_G \geq \delta \geq 0$.

$(\delta = I(M;Y) \Rightarrow \widehat{\delta}_G = I(M;Y))$ Note that if $\widehat{T} = \mathbf{0}$, we have

$$\widehat{\Sigma}_T = I + H_Y\Sigma_M H_Y^\mathsf{T}$$
$$\Rightarrow \quad \widehat{P}_{Y'|X} \circ P_{X|M} = P_Y \tag{131}$$
$$\Rightarrow \quad \widehat{\delta}_G = I(M;Y).$$

As such, it suffices to show that $\delta = I(M;Y)$ implies that the objective function in (18) is minimized at $T = \mathbf{0}$. Suppose for a contradiction that there exists a $\widetilde{T}$ such that:

$$\mathbb{E}_{P_M}\left[\|(\widetilde{T}H_X - H_Y)M\|_A^2\right] < \mathbb{E}_{P_M}\left[\|H_Y M\|_A^2\right] \tag{132}$$

with $A := I + H_Y\Sigma_M H_Y^\mathsf{T}$ giving the marginal covariance matrix for Y (i.e. $P_Y = \mathcal{N}(\mathbf{0}, A)$). This implies that

$$\mathbb{E}_{P_M}\left[\|(\lambda\widetilde{T}H_X - H_Y)M\|_A^2\right] < \mathbb{E}_{P_M}\left[\|H_Y M\|_A^2\right] \tag{133}$$

for any $\lambda \in (0,1)$. The implication follows from the convexity of the Mahalanobis distance, noting that $\lambda\widetilde{T} = \lambda\widetilde{T} + (1 - \lambda)\mathbf{0}$. In other words, we can keep making $\widetilde{T}$ "smaller" (by multiplying it by smaller values of λ), until $\widetilde{\Sigma}_T := A - \widetilde{T}\widetilde{T}^\mathsf{T}$ is PSD, while still making sure that (132) holds. Defining $\widetilde{P}_{Y'|X} := \mathcal{N}(\widetilde{T}X, \widetilde{\Sigma}_T)$, we note that $\widetilde{P}_{Y'|X} \circ P_{X|M} = \mathcal{N}(\widetilde{T}H_X M, A)$. As such:

$$\mathbb{E}_{P_M}\left[D(P_{Y|M}\|\widetilde{P}_{Y'|X} \circ P_{X|M})\right] - I(M;Y)$$
$$= \mathbb{E}_{P_M}\left[D(P_{Y|M}\|\widetilde{P}_{Y'|X} \circ P_{X|M}) - D(P_{Y|M}\|P_Y)\right]$$
$$= \mathbb{E}_{P_M}\left[\|(\widetilde{T}H_X - H_Y)M\|_A^2\right] - \mathbb{E}_{P_M}\left[\|H_Y M\|_A^2\right] < 0.$$

But since $\widetilde{P}_{Y'|X} \in \mathcal{C}_G(\mathsf{Y} \mid \mathsf{X})$, we have:

$$\delta_G \leq \mathbb{E}_{P_M}\left[D(P_{Y|M}\|\widetilde{P}_{Y'|X} \circ P_{X|M})\right] < I(M;Y) \leq \delta, \tag{134}$$

which is a contradiction. $\qquad\square$

We briefly discuss how to reformulate the optimization problem in (18) such that it satisfies the disciplined convex programming (DCP) rules and can be solved using the CVXPY software package [35], [36]. First we note that the objective function can be rewritten using the trace trick. Letting $A = I + H_Y \Sigma_M H_Y^\mathsf{T}$ for ease of notation:

$$
\begin{aligned}
\mathbb{E}_{P_M} &\left[\| (TH_X - H_Y)M \|_A^2 \right] \\
&= \mathbb{E}_{P_M} \left[M^\mathsf{T} (TH_X - H_Y)^\mathsf{T} A^{-1} (TH_X - H_Y) M \right] \\
&= \mathbb{E}_{P_M} \left[\mathrm{Tr} \left(M^\mathsf{T} (TH_X - H_Y)^\mathsf{T} A^{-1} (TH_X - H_Y) M \right) \right] \\
&= \mathbb{E}_{P_M} \left[\mathrm{Tr} \left(MM^\mathsf{T} (TH_X - H_Y)^\mathsf{T} A^{-1} (TH_X - H_Y) \right) \right] \\
&= \mathrm{Tr} \left(\Sigma_M (TH_X - H_Y)^\mathsf{T} A^{-1} (TH_X - H_Y) \right) \\
&= \| A^{-\frac{1}{2}} TH_X \Sigma_M^{\frac{1}{2}} - A^{-\frac{1}{2}} H_Y \Sigma_M^{\frac{1}{2}} \|_F^2
\end{aligned}
$$

where $\| \cdot \|_F^2$ is the squared Frobenius norm. Next, the constraint can be rewritten using the Schur complement [37]:

$$
I + H_Y \Sigma_M H_Y^\mathsf{T} - T(I + H_X \Sigma_M H_X^\mathsf{T}) T^\mathsf{T} \succcurlyeq 0
$$

$$
\Updownarrow
$$

$$
\begin{bmatrix} I + H_Y \Sigma_M H_Y^\mathsf{T} & T \\ T^\mathsf{T} & (I + H_X \Sigma_M H_X^\mathsf{T})^{-1} \end{bmatrix} \succcurlyeq 0
$$

Combining the two, we obtain the problem in a form that can solved directly by CVX:

$$
\begin{aligned}
\widehat{T} = \ &\underset{T}{\mathrm{argmin}} \quad \| A^{-\frac{1}{2}} TH_X \Sigma_M^{\frac{1}{2}} - A^{-\frac{1}{2}} H_Y \Sigma_M^{\frac{1}{2}} \|_F^2 \\
&\text{s.t.} \quad \begin{bmatrix} I + H_Y \Sigma_M H_Y^\mathsf{T} & T \\ T^\mathsf{T} & (I + H_X \Sigma_M H_X^\mathsf{T})^{-1} \end{bmatrix} \succcurlyeq 0
\end{aligned}
\tag{135}
$$

We solved the problem used the splitting conic solver (SCS) [38], [39] with relaxation parameter 1, maximum iteration number 5000, and convergence tolerance of 1e-10. The maximum iteration number and convergence tolerance were chosen to be conservatively large and small, respectively, as we found lower and higher values of these parameters to yield inferior results (i.e. larger estimated deficiencies) and computational time was not an issue. Due to issues of numerical imprecision, the estimated $\widehat{\Sigma}_T$ was occasionally found to have very small negative eigenvalues (i.e. greater than -1e-6). To ensure that the estimated $\widehat{\Sigma}_T$ was positive semidefinite, we replaced all negative eigenvalues with zero.

Here, we describe the procedure used to generate the plots in Figure 1.

1) Sample d_M, d_X, d_Y according one of (S1)–(S4).
2) Sample a covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ (where $d = d_M + d_X + d_Y$) from a standard Wishart distribution.
3) Compute $I(M; (X, Y))$, $I(M; X)$, and $I(M; Y)$.
4) Compute conditional mean and covariances and whiten to obtain H_X, H_Y, and $\Sigma_{X|M} = \Sigma_{Y|M} = I$.
5) Estimate the deficiencies $\widehat{\delta}_G(M : Y \backslash X)$ and $\widehat{\delta}_G(M : X \backslash Y)$, using the the Python CVX package to solve (135).
6) Compute the approximate δ-PID atoms.
7) Check if $\widehat{RI}$ and $\widehat{SI}$ are non-negative.
8) If $d_M = 1$, check if either $\widehat{UI}_X \approx 0$ or $\widehat{UI}_Y \approx 0$.

A. Simulation setup

The results shown in Figure 2 are for a *multivariate Poisson* distribution, which is defined below. We used a small number of dimensions ($d_M = 2$ and $d_X = d_Y = 1$) in order to allow for computational tractability of the discrete $\sim$-PID estimator of Banerjee et al. [23].

We took $M_i \sim$ i.i.d. Poisson($\lambda = 2$), for both components of M (i.e, $i = 1, 2$). We took $X = X_1 + X_2 + N_X$, where $X_i \sim$ Binomial(M_i, w_i^X) for some "weight" w_i^X, and $N_X \sim$ Poisson($\lambda_X = 1$). Y was defined in the same way as X, with Y_i, N_Y, X_i and N_X all being independent of each other. The results shown in Figure 2 are for $w_1^Y = w_2^Y = w_2^X = 0.5$, while w_1^X is varied between 0.0 and 1.0 in increments of 0.1 on the x-axis.

The discrete $\sim$-PID was estimated by truncating the Poisson distributions, while ensuring >95% of probability was captured. For computing the $\widehat{\delta}_G$-PID, the Poisson distribution was approximated as a Gaussian distribution by considering the sample covariance matrix, estimated from $N = 10^6$ data points. This was observed to produce a stable covariance estimate, following which channel parameters such as H_X, $\Sigma_{X|M}$, etc. were extracted from the estimated covariance matrix. The convex optimization problem in the definition of $\widehat{\delta}_G$-PID was solved using CVXPY (see Appendix F for details).

B. Results and Limitations

It should be noted that this experiment has several intrinsic limitations:

1) The Poisson distributions had to be truncated for reasons of finiteness and computational tractability, before estimating the $\sim$-PID. As mentioned above, we endeavoured to ensure that the distribution was captured well, but the truncation could nevertheless skew the discrete-PID estimates.
2) The estimation of the sample covariance matrix used to compute the $\widehat{\delta}_G$-PID could also contribute to error.
3) Lastly and most importantly, the two PIDs being considered are different, i.e., the $\sim$-PID for the Poisson distribution, and the δ-PID for the Gaussian approximation. These PIDs likely produce slightly different values: the δ-PID is closer in spirit to the PID of Harder et al. [4] (see [5] for an explanation), which was shown to have different values from the $\sim$-PID by Bertschinger et al. [3].

These limitations suggest that we have no way of truly knowing how accurate either of the two estimates are. However, their degree of agreement despite the use of completely different estimation techniques cannot be purely coincidental, and argues in favor of the practical applicability of our method. In particular, these results provide evidence suggesting that:

1) Our approximate PID is able to differentiate between cases of zero- and non-zero unique information.
2) Our proposed method matches expected trends for every one of the four PID quantities.
3) As a fraction of total mutual information, each PID value is reasonably close to its expected value, suggesting our method can be used to assess relative differences between PID quantities in different settings.
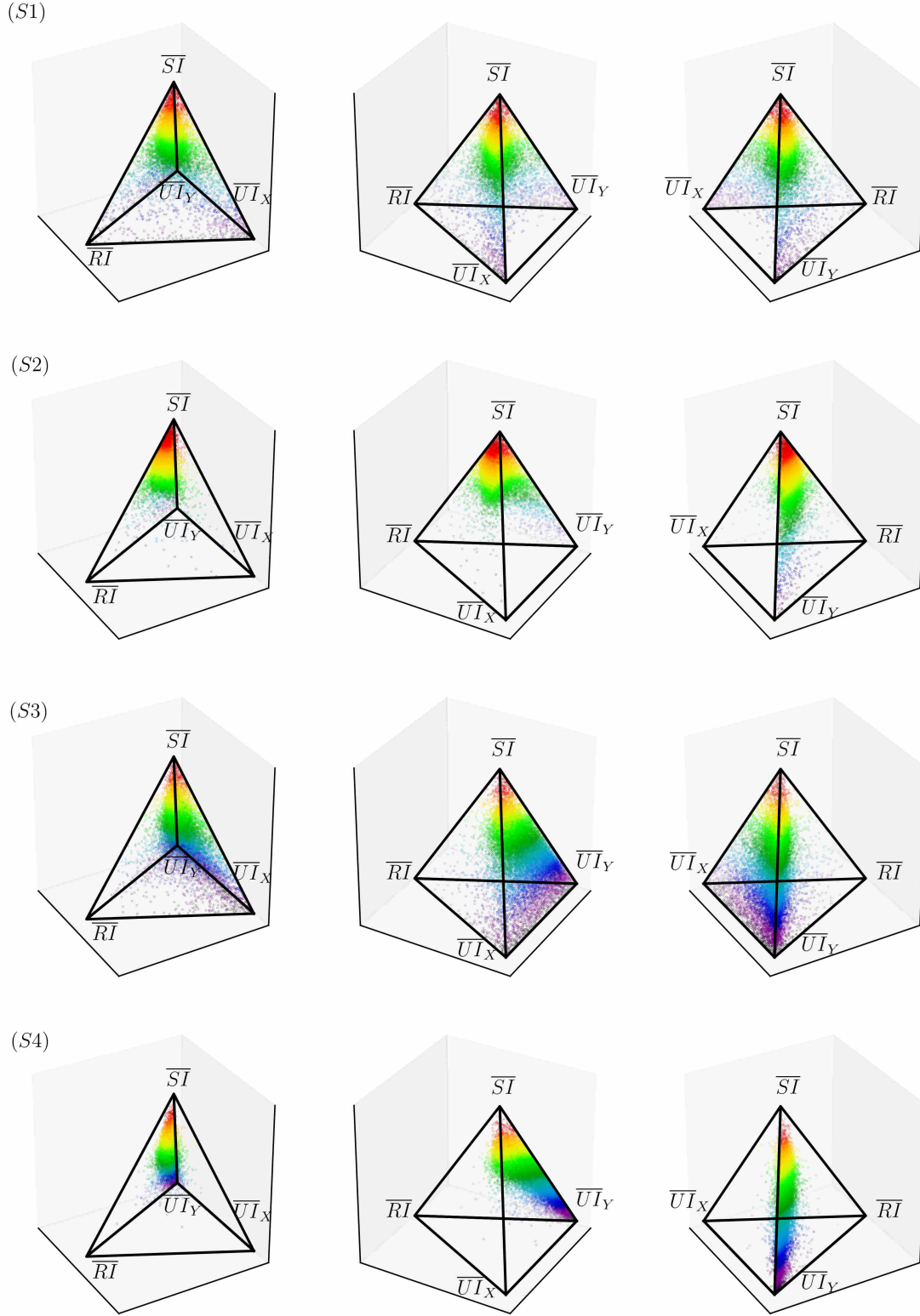
($S1$)

($S2$)

($S3$)

($S4$)

Fig. 3: Alternative views of Figure 1. Each row displays one of the four sampling schemes, and each column provides a rotated view of the simplex.