Partial Information Decomposition via Deficiency for Multivariate Gaussians

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

Praveen Venkatesh*
vpraveen@cmu.edu
Dept. of Electrical and Computer Engineering
Carnegie Mellon University

***Abstract*—We consider the problem of decomposing the information content of three jointly Gaussian random vectors using the partial information decomposition (PID) framework. Barrett [1] previously characterized the Gaussian PID for a scalar "source" or "message" in closed form—we extend this to the case where the message is a vector. Specifically, we revisit a connection between the notions of Blackwell sufficiency of statistical experiments and stochastic degradedness of broadcast channels, to provide a necessary and sufficient condition for the existence of unique information in the fully multivariate Gaussian PID. The condition we identify indicates that the closed form PID for the scalar case rarely extends to the vector case. We also provide a convex optimization approach for approximating a PID in the vector case, analyze its properties, and evaluate it empirically on randomly generated Gaussian systems.**

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 the information about a "message" variable, M, is decomposed between *two* others, X and Y, and is hence commonly referred to as the *bivariate* decomposition. More precisely, a bivariate PID decomposes the mutual information between M and the pair (X, Y) into four 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). Recently, several competing definitions have emerged for unique, redundant and synergistic information [2–7] (see [8] for a review). Some of these definitions, which we believe are operationally better motivated than others, have an important feature in common: they satisfy a property we call the *Blackwell criterion*. This property states that Y has *no* unique information about M if and only if X is *Blackwell sufficient* for Y [9]. Informally, X is said to be Blackwell sufficient for Y if, on average, one can make better decisions about M (w.r.t. any loss function) by observing X rather than Y.

This paper focuses on the bivariate PID for jointly Gaussian random vectors M, X and Y. The Gaussian case was previously examined by Barrett [1] for a few PID definitions that satisfy the Blackwell criterion. Barrett showed that when M is restricted to be scalar (while X and Y may be vectors), either X or Y is *always* Blackwell sufficient for the other, and hence only *one* of them can have unique information about M. Consequently, Barrett's result provided a straightforward way to *compute* the PID in closed form for scalar M. However, the case of *vector* M has remained unresolved in the PID literature: (1) *Under what conditions does Blackwell sufficiency hold?* and (2) *How do we compute the PID when it does not?*

Our work extends Barrett's result to the case of fully multivariate[1] Gaussian random variables, by characterizing when Blackwell sufficiency holds and providing a convex optimization framework for approximately computing the Gaussian PID when it does not. Our first result emerges from the observation that Blackwell sufficiency is equivalent to stochastic degradedness of broadcast channels (an observation also made by Raginsky [10]). We then draw on the results of Gerdes et al. [11], who (building on prior work by Shang and Poor [12]) provide a necessary and sufficient condition for the stochastic degradedness of linear additive Gaussian-noise MIMO broadcast channels. By extension, this is also a necessary and sufficient condition for Blackwell sufficiency in fully multivariate Gaussians. It should be noted that this result was also shown much earlier in the statistical decision theory literature using a different proof technique [13, Thm. 8.2.13].

Our central contribution in extending Barrett's result lies, therefore, in tying together disparate works from different literatures and arriving at an implication for the Gaussian PID that was hitherto unknown. Indeed, our work shows that Barrett's result is simply a partial extension of the well-known fact that scalar Gaussian broadcast channels are always stochastically degraded [14, Example 15.6.6].

Building on our characterization of Blackwell sufficiency, we also propose a convex optimization framework for *computing* a deficiency-based PID definition [5] for fully multivariate Gaussians. The notion of *deficiency*, first introduced by Le Cam [15, 13], quantifies how far X is from being Blackwell sufficient for Y, and so is a natural measure of unique information in Y. We show that the deficiency computed using our optimization framework satisfies some basic desirable properties, a few of which we examine empirically.

We begin with a brief review of the basics of the PID, the Blackwell criterion and broadcast channels (Section II). We then state the equivalence of Blackwell sufficiency and

*Equal contribution; author order was determined by a coin flip. GS was supported by the Picower Postdoctoral Fellowship.

[1]We use the term "fully *multivariate* Gaussian PID" 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.

stochastic degradedness and present our extension of Barrett's result characterizing the existence of unique information in fully multivariate Gaussians (Section III). In Section IV, we provide a convex optimization framework to approximate the deficiency-based PID for multivariate Gaussians, and present empirical results in Section V. We conclude with a discussion of open questions in Section VI.

II. Background

A. Partial Information Decomposition

Let M, X and Y be three random variables with sample spaces M, X and Y respectively, and joint density P_{MXY}. Williams and Beer [2] proposed the following general decomposition for the bivariate case:

$$I\big(M;(X,Y)\big) = UI(M:X\setminus Y) + UI(M:Y\setminus X) + RI(M:X;Y) + SI(M:X;Y), \tag{1}$$

where $UI(M:X\setminus Y) \geq 0$ is the information about M uniquely present in X and not in Y, while $RI(M:X;Y)\geq 0$ and $SI(M:X;Y)\geq 0$ are respectively the redundant and synergistic information about M contained between X and Y. For ease of notation, we will interchangeably represent these PID quantities as UI_X, UI_Y, RI, and SI respectively. Such a decomposition should also satisfy

$$I\big(M;X\big) = UI(M:X\setminus Y) + RI(M:X;Y), \tag{2}$$
$$I\big(M;Y\big) = UI(M:Y\setminus X) + RI(M:X;Y). \tag{3}$$

Based on these three equations, defining any one of the four quantities in the RHS of (1) suffices to determine the other three. For instance, Williams and Beer [2] proposed the following definition for redundancy.

Definition 1 (MMI-PID [2])**.** *Williams and Beer proposed that the redundant information about M present in both X and Y be given be the **M**inimum of their respective **M**utual **I**nformations with M (hence "MMI"):*

$$RI_{\text{MMI}}(M:X;Y) \triangleq \min\{I(M;X), I(M;Y)\}. \tag{4}$$

The above definition, with equations (1)–(3), fully determines the MMI-PID, i.e., $UI_{\text{MMI}}(M:X\setminus Y)$, $UI_{\text{MMI}}(M:Y\setminus X)$, and $SI_{\text{MMI}}(M:X;Y)$ are now well-defined.

B. Blackwell Sufficiency and the Blackwell Criterion

Before stating the two other PID definitions we refer to in this paper, we introduce the notion of Blackwell sufficiency, which forms the operational basis motivating these definitions. The original definition of Blackwell sufficiency [9] was based on statistical decision theory, addressing whether or not it is possible to attain the *same risk* by making decisions based on one experiment as with another. For simplicity, we present an equivalent [9] notion of Blackwell sufficiency, which is more amenable to our setup here.

Definition 2 (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*

$$P_{Y'|X} \circ P_{X|M} \;=\; P_{Y|M}, \tag{5}$$

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

$$(P_{Y'|X}\circ P_{X|M})(y|m) \triangleq \int P_{Y'|X}(y|x)P_{X|M}(x|m)\,dx. \tag{6}$$

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. The relationship between Blackwell sufficiency and the remaining PID definitions can be summarized by the following property [3]:

Property 1 (The Blackwell criterion)**.** *For a bivariate PID of $I\big(M;(X,Y)\big)$, Y has zero unique information about M with respect to X if and only if $X \succcurlyeq_M Y$.*

We now state two PID definitions given by Bertschinger et al. [3] and Banerjee et al. [5], which have been shown to satisfy the Blackwell criterion [3, 13] but quantify departures from Blackwell sufficiency differently.[2]

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

$$\widetilde{UI}(M:X\setminus Y) \triangleq \min_{Q\in\Delta_P} I_Q(M;X\,|\,Y), \tag{7}$$

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

Definition 4 (δ-PID [5])**.** *Let the (weighted output)* deficiency[3] *of X with respect to Y about M be defined as*[4]

$$\delta(M:Y\setminus X) \triangleq \inf_{P_{Y'|X}\in\mathcal{C}(\mathsf{Y}|\mathsf{X})} \mathbb{E}_{P_M}\big[D(P_{Y|M}\,\|\,P_{Y'|X}\circ P_{X|M})\big], \tag{8}$$

Then, the deficiency-based redundant information about M present in X and Y is given by

$$RI_\delta(M:X;Y) \triangleq \min\{I(M;X)-\delta(M:X\setminus Y), I(M;Y)-\delta(M:Y\setminus X)\}. \tag{9}$$

As with the MMI-PID, equations (1)–(3) fully determine the remaining components of the $\sim$-PID and the δ-PID. We are now in a position to formally state Barrett's result for the PID of Gaussian random variables.

Theorem 1 (Barrett [1])**.** *If M, X and Y are all jointly Gaussian, and M is scalar, then we always have that either $X \succcurlyeq_M Y$ or $Y \succcurlyeq_M X$. Therefore, the $\sim$-PID (as well as the PID of Harder et al. [4]) reduce to the MMI-PID.*

We omit a proof of this theorem, since it is subsumed by the extension derived in Section III.

[2] So far, PIDs have mostly been defined for *discrete* random variables. Barrett [1] did not formally extend the PID definitions of other authors to continuous variables even though he analyzed Gaussians. In the definitions given here, we consider extensions to the continuous case, but relegate a detailed discussion of measure-theoretic issues to future work. For our purposes, we assume that all joint and conditional probability measures, as well as information measures thereof, are well defined.

[3] There are many ways to define deficiency; Raginsky [10] 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.

[4] 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.

C. Broadcast channels

To extend Barrett's result to the case of fully multivariate Gaussians, we leverage the connection between Blackwell sufficiency and the concept of stochastic degradedness from the literature on broadcast channels. Following Gerdes et al. [11], we define stochastic degradedness as follows:

Definition 5 (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.*

III. Existence of Unique Information in Multivariate Gaussians

In this section, we show that the concepts of Blackwell sufficiency and stochastic degradedness are identical (this was also mentioned in passing by Raginsky [10]). We also characterize a necessary and sufficient condition under which stochastic degradedness (and hence Blackwell sufficiency) holds for fully multivariate Gaussians. Let the sample spaces of M, X and Y respectively be $\mathsf{M} = \mathbb{R}^{d_M}$, $\mathsf{X} = \mathbb{R}^{d_X}$ and $\mathsf{Y} = \mathbb{R}^{d_Y}$. We parameterize the joint distribution P_{MXY} as:

$$M \sim P_M = \mathcal{N}(0, \Sigma_M) \tag{10}$$

$$X \mid M \sim P_{X|M} = \mathcal{N}(H_X M, \Sigma_X) \tag{11}$$

$$Y \mid M \sim P_{Y|M} = \mathcal{N}(H_Y M, \Sigma_Y) \tag{12}$$

Here, the joint covariance matrix of M, X and Y is given by Σ, and we *do not* assume that $X \perp\!\!\!\perp Y \mid M$.

Remark 1. *Without loss of generality, we assume that the noise covariance matrices Σ_X and Σ_Y are full rank. This can always be done, since removing linearly dependent elements within X and Y respectively has no effect on information-theoretic relationships.*

Lemma 2 (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 $P_{X|M}$.*

The proof is quite trivial and primarily involves juggling notation, but is provided in Appendix A for completeness. In fact, Cover and Thomas [14, Sec. 15.6.2] even *define* stochastic degradedness similar to how we have defined Blackwell sufficiency.

Before moving to the main result of this paper, it is worth noting that Barrett's result does *not* directly extend to the case of jointly Gaussian variables when M is a vector. Barrett appears to leave this as an open question [1, end of Sec. IV-B], but it can be answered with a simple counterexample:

Counterexample. Consider the case where $M = [M_1, M_2]$, $X = M_1$ and $Y = M_2$, with $M_1, M_2 \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. Intuitively, X and Y each have an equal amount of unique information about M, and one can show that this holds for the $\sim$- and δ-PIDs. However, under the MMI-PID, X and Y have only redundant and synergistic information and no unique information. Therefore, PIDs that satisfy the Blackwell criterion do not always reduce to the MMI-PID when M is a vector. □

Theorem 3. *For jointly Gaussian random vectors M, X and Y, $X \succcurlyeq_M Y$ if and only if $H_X^T \Sigma_X^{-1} H_X \succcurlyeq H_Y^T \Sigma_Y^{-1} H_Y$.*

The proof of this result follows directly from the result of Gerdes et al. [11], along with Lemma 2. In an effort to present a self-contained record of proofs, we provide it in Appendix B. The main result can be interpreted as follows: for jointly Gaussian vectors M, X and Y, and a PID satisfying the Blackwell criterion, X and Y *both have* unique information about M *unless* the condition in Theorem 3 holds (in one direction or the other). This is significant, as the condition is fairly restrictive: when $d_X, d_Y < d_M$, the condition will hold w.p. 0 assuming a continuous distribution over H_X and H_Y [11, Prop. 2]. Given that the MMI-PID assigns unique information to only one of X or Y, it is clear that an alternative approach to computing PIDs is needed when $d_M > 1$.

Remark 2. *Since Σ_X and Σ_Y are invertible (Remark 1), we can equivalently write the condition in Theorem 3 using a whitening transform:*

$$\widetilde{H}_X^T \widetilde{H}_X \succcurlyeq \widetilde{H}_Y^T \widetilde{H}_Y; \quad \widetilde{H}_X \triangleq \Sigma_X^{-\frac{1}{2}} H_X, \; \widetilde{H}_Y \triangleq \Sigma_Y^{-\frac{1}{2}} H_Y \tag{13}$$

Without loss of generality, we assume henceforth that H_X and H_Y have been whitened, and hence $\Sigma_X = \Sigma_Y = I$.

IV. Approximating the δ-PID

Having established a complete characterization of when unique information exists in the multivariate Gaussian setting, we now seek to compute the δ-PID. This requires computing the weighted output deficiency given by (8), which is challenging because it requires an optimization over all continuous conditional distributions $P_{Y'|X}$. Instead, we consider the restricted problem where $P_{Y'|X}$ lies in the set of linear additive Gaussian noise channels $\mathcal{C}_G(\mathsf{Y} \mid \mathsf{X}) \subset \mathcal{C}(\mathsf{Y} \mid \mathsf{X})$. 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 $\Sigma_T \in \mathbb{R}^{d_Y \times d_Y}$, $\Sigma_T \succcurlyeq 0$, so that the *Gaussian deficiency* can be defined as:

$$\begin{aligned}
&\delta_G(M : Y \setminus X) \\
&\quad \triangleq \inf_{P_{Y'|X} \in \mathcal{C}_G(\mathsf{Y}|\mathsf{X})} \mathbb{E}_{P_M}\left[D(P_{Y|M} \,\|\, P_{Y'|X} \circ P_{X|M})\right] && (14)\\
&\quad = \inf_{T, \Sigma_T \succcurlyeq 0} \frac{1}{2}\Bigg[\mathbb{E}_{P_M}\left[\left\|(TH_X - H_Y)M\right\|^2_{\Sigma_T + T\Sigma_X T^T}\right] && (15)\\
&\qquad\qquad + \operatorname{Tr}\left((\Sigma_T + T\Sigma_X T^T)^{-1}\Sigma_Y\right) \\
&\qquad\qquad + \log \frac{|\Sigma_T + T\Sigma_X T^T|}{|\Sigma_Y|} - d_Y \Bigg] \\
&\quad = \inf_{T, \Sigma_T \succcurlyeq 0} \frac{1}{2}\Bigg[\mathbb{E}_{P_M}\left[\left\|(TH_X - H_Y)M\right\|^2_{\Sigma_T + TT^T}\right] && (16)\\
&\qquad\qquad + \operatorname{Tr}\left((\Sigma_T + TT^T)^{-1}\right) \\
&\qquad\qquad + \log|\Sigma_T + TT^T| - d_Y \Bigg],
\end{aligned}$$

where $\|a\|_B^2 \triangleq aB^{-1}a^T$ is the squared Mahalanobis distance, and (16) follows from Remark 2. Unfortunately, this problem may not be convex since $\operatorname{Tr}\left((\Sigma_T + TT^T)^{-1}\right)$ is the composition of a convex function ($\operatorname{Tr}\left((\cdot)^{-1}\right)$) with a non-monotonic function of T. Therefore, we propose a convex relaxation of (14) to find an approximate minimizer $\widehat{P}_{Y'|X} = \mathcal{N}(\widehat{T}X, \widehat{\Sigma}_T)$.

To motivate the proposed relaxation, we note the deficiency is bounded as $0 \leq \delta_G(M : Y \setminus X) \leq I(M;Y)$ and target our approach to recover the true deficiency when it takes these extremal values. Letting $P^*_{Y'|X} = \mathcal{N}(T^*X, \Sigma^*_T)$ be the minimizer of (14), we note that

$$T^*H_X = H_Y,\ \Sigma^*_T = I - T^*T^{*T} \Rightarrow \delta_G(M:Y\setminus X) = 0;$$
$$T^* = \mathbf{0},\ \Sigma^*_T = I + H_Y\Sigma_M H_Y^T \Rightarrow \delta_G(M:Y\setminus X) = I(M;Y).$$

Intuitively, as the ability to construct a channel $P_{Y'|X} \circ P_{X|M}$ that replicates $P_{Y|M}$ diminishes (i.e. the deficiency approaches the mutual information), we need to begin incorporating the noise associated with the marginal distribution P_Y in $\widehat{\Sigma}_T$. Thus, given an estimated $\widehat{T}$, we construct $\widehat{\Sigma}_T$ as:

$$\widehat{\Sigma}_T = I + H_Y\Sigma_M H_Y^T - \widehat{T}(I + H_X\Sigma_M H_X^T)\widehat{T}^T \tag{17}$$

The remaining problem is identifying a reasonable gain matrix $\widehat{T}$ for $\widehat{P}_{Y'|X}$. Since we cannot directly minimize the first term in (16), we formulate a convex relaxation to approximate T^*:

$$\begin{aligned} \widehat{T} = \ & \underset{T}{\operatorname{argmin}} \quad \mathbb{E}_{P_M}\left[\left\|(TH_X - H_Y)M\right\|^2_{I+H_Y\Sigma_M H_Y^T}\right] \\ & \text{s.t.} \quad \Sigma_T \succcurlyeq 0 \end{aligned} \tag{18}$$

with Σ_T as defined in (17) with T in place of $\widehat{T}$. The choice to minimize the Mahalanobis distance w.r.t. $I + H_Y\Sigma_M H_Y^T$ is motivated in the proof of Proposition 4 and is partially justified by the empirical results in Section V, with further justifications and/or demonstrations of optimality left for future work.

Having defined an approach for obtaining $\widehat{P}_{Y'|X}$, the resulting approximate deficiency is given by:

$$\begin{aligned} \widehat{\delta}_G(M : Y \setminus X) \triangleq \frac{1}{2}\Bigg[& \mathbb{E}_{P_M}\left[\left\|(\widehat{T}H_X - H_Y)M\right\|^2_{\widehat{\Sigma}_T + \widehat{T}\widehat{T}^T}\right] \\ & + \operatorname{Tr}((\widehat{\Sigma}_T + \widehat{T}\widehat{T}^T)^{-1}) \\ & + \log|\widehat{\Sigma}_T + \widehat{T}\widehat{T}^T| - d_Y\Bigg] \end{aligned} \tag{19}$$

It can be shown that $\widehat{\Sigma}_T + \widehat{T}\widehat{T}^T$ is guaranteed to be invertible when using the relaxation of (17) and (18). We state this formally and prove it in Appendix C. Because $\widehat{P}_{Y'|X} \in \mathcal{C}_G(\mathsf{Y}|\mathsf{X}) \subset \mathcal{C}(\mathsf{Y}|\mathsf{X})$, *any* method for obtaining $\widehat{T}$ and $\widehat{\Sigma}_T$ will yield:

$$\widehat{\delta}_G(M : Y \setminus X) \geq \delta_G(M : Y \setminus X) \geq \delta(M : Y \setminus X) \tag{20}$$

Now we provide conditions for when the estimate in (19) coincides with the true deficiency:

Proposition 4. *For jointly Gaussian random vectors M, X and Y, the approximate deficiency defined by (17)–(19) satisfies:*

$$\delta(M:Y\setminus X) = 0 \Leftrightarrow \widehat{\delta}_G(M:Y\setminus X) = 0 \tag{21}$$
$$\delta(M:Y\setminus X) = I(M;Y) \Rightarrow \widehat{\delta}_G(M:Y\setminus X) = I(M;Y) \tag{22}$$

The proof is given in Appendix D. Plugging in $\widehat{\delta}_G$ into (9), we obtain an approximation of RI_δ, and hence an approximation of the δ-PID, which we call the $\widehat{\delta}_G$-PID:

$$\begin{aligned} \widehat{RI}(M : X;Y) \triangleq \ & \min\{I(M;X) - \widehat{\delta}_G(M : X \setminus Y), \\ & I(M;Y) - \widehat{\delta}_G(M : Y \setminus X)\} \end{aligned} \tag{23}$$

For brevity, we use $\widehat{RI}$ etc. when referring to the constituent atoms of the $\widehat{\delta}_G$-PID. As with other PIDs, $\widehat{RI}$ paired with equations (1)–(3) fully determines the remaining atoms of the $\widehat{\delta}_G$-PID: $\widehat{UI}_X$, $\widehat{UI}_Y$ and $\widehat{SI}$. Conveniently, the inequalities in (20) bound the approximated $\widehat{\delta}_G$-PID by the true δ-PID:

Proposition 5. *For jointly Gaussian random vectors M, X and Y, the $\widehat{\delta}_G$-PID imposes bounds on the δ-PID:*

$$\widehat{RI}(M : X ; Y) \leq RI_\delta(M : X ; Y) \tag{24}$$
$$\widehat{UI}(M : X \setminus Y) \geq UI_\delta(M : X \setminus Y) \tag{25}$$
$$\widehat{UI}(M : Y \setminus X) \geq UI_\delta(M : Y \setminus X) \tag{26}$$
$$\widehat{SI}(M : X ; Y) \leq SI_\delta(M : X ; Y) \tag{27}$$

The proof follows directly by applying the inequality in (20) to compare (9) and (23), and from the basic PID equations. It is worth noting that this proposition holds for *any* $\widehat{T}$ and $\widehat{\Sigma}_T \succcurlyeq 0$ (not just the minimizer of (18)), and is thus of limited utility as a standalone result. For example, we could simply define $\widehat{T} = \mathbf{0}$ and $\widehat{\Sigma}_T = I + H_Y\Sigma_M H_Y^T$, and the proposition would recover trivial bounds implied by the basic PID equations ($RI_\delta \geq 0$, $UI_{\delta,X} \leq I(M;X)$, $UI_{\delta,Y} \leq I(M;Y)$, and $SI_\delta \geq I(M;(X,Y)) - I(M;X) - I(M;Y)$). However, we believe Proposition 5 is worth stating for two reasons. First, it highlights that our proposed approach is in some sense complementary to the MMI-PID, as the MMI-PID recovers trivial bounds in the opposite direction, i.e. upper bounds on redundant and synergistic information and lower bounds on unique information. Second, if the $\widehat{\delta}_G$-PID is a *valid* PID (i.e., if all atoms are non-negative and satisfy equations (1)–(3)) then each atom of the $\widehat{\delta}_G$-PID is bounded *between* the true δ-PID and the $\sim$-PID [3]. This is because the $\sim$-PID upper bounds the unique information and lower bounds the redundant and synergistic information [3, Lemma 3] of every other PID definition which satisfies a basic property[5]—one that is satisfied by the $\widehat{\delta}_G$-PID.[6] Unfortunately, proving that the $\widehat{\delta}_G$-PID is a valid PID requires proving that $\widehat{SI} \geq 0$. We have not been able to demonstrate this formally, but in the following section we test it empirically.

V. Empirical Results

Having established a framework for approximating the δ-PID, we seek to address three questions using simulations: (Q1) *Does $\widehat{\delta}_G$ yield a valid PID for jointly Gaussian systems?* (Q2) *Is the $\widehat{\delta}_G$-PID consistent with the results of Barrett [1] for the case of univariate M?* (Q3) *How does information decompose in multivariate Gaussian systems?*

To address these questions, we sample a random covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ (where $d = d_M + d_X + d_Y$) from a

[5] This property, known as Assumption $(*)$ in [3], requires that the redundant and unique informations depend only on the marginal distribution P_M, and the channels $P_{X|M}$ and $P_{Y|M}$. This is easily seen to be satisfied by the δ-PID, as well as the $\widehat{\delta}_G$-PID.

[6] This line of reasoning assumes that Lemma 3 from [3] holds for the $\sim$-PID defined for *continuous* random variables. This assumption was implicit in the demonstration of Barrett's central result in [1], but it has not been explicitly verified to our knowledge.

standard Wishart distribution, which fully characterizes the joint distribution P_{MXY}. To see how the PID varies with different dimensionalities d_M, d_X, and d_Y, we sample them using one of four sampling schemes:

(S1) $d_M \sim \text{Unif}\{1 \dots 10\}$ and $d_X = d_Y = d_M$
(S2) $d_M \sim \text{Unif}\{1 \dots 9\}$ and $d_X, d_Y \overset{\text{iid}}{\sim} \text{Unif}\{d_M + 1 \dots 10\}$
(S3) $d_M \sim \text{Unif}\{2 \dots 10\}$ and $d_X, d_Y \overset{\text{iid}}{\sim} \text{Unif}\{1 \dots d_M - 1\}$
(S4) $d_M \sim \text{Unif}\{2 \dots 9\}$, $d_X \sim \text{Unif}\{1 \dots d_M - 1\}$, and $d_Y \sim \text{Unif}\{d_M + 1 \dots 10\}$

Without loss of generality, we assume $d_X \leq d_Y$ (i.e. we switch their values if $d_Y < d_X$). For each of these sampling schemes, we consider 20,000 samples of d_M, d_X, d_Y and Σ, yielding 80,000 total systems and PID approximations. Further details on the implementation and experimental setup are provided in Appendices E and F and all of the associated code is provided at https://github.com/gabeschamberg/mvar-gauss-pid.

The first noteworthy result is that all of the approximated PID quantities were non-negative for every sampled system. This suggests an affirmative answer to (Q1), perhaps with the exception of a zero-measure set of systems. Our results also suggest an affirmative answer to (Q2), as every system for which $d_M = 1$ ($N = 4256$) resulted in our approximation assigning unique information to only one of X or Y.

To address (Q3), we visualize how the distribution of unique, redundant, and synergistic information changes for the different relationships between d_M, d_X, and d_Y in Figure 1. Given that the scale of the PID components varies with $I(M;(X,Y))$, we consider normalized PID quantities $\overline{UI}_X$, $\overline{UI}_Y$, $\overline{RI}$, and $\overline{SI}$ obtained by dividing the corresponding approximate PID quantities by $I(M;(X,Y))$. Thus, 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 system by its location on a 4-simplex characterizing the proportion of $I(M;(X,Y))$ that is accounted for by each PID component.

Figure 1 provides numerous insights into the relative prevalence of the PID components in multivariate Gaussian systems. First, synergy has a strong presence in all four panels. Synergy appears to be most prevalent when $d_M \leq d_X, d_Y$ (S1 and S2). This result is not entirely surprising—the prevalence of synergy in systems where $d_M = 1$ was acknowledged by Barrett [1]. Correspondingly, (S2), which gives the closest multivariate analogue to the univariate case in the sense that $d_M = 1 \Rightarrow d_M \leq d_X, d_Y$, shows the largest synergy. Redundancy is never particularly prevalent—this may be a consequence of how we have drawn a random covariance matrix, which reduces the likelihood that X and Y capture similar "directions" of M. In other words, since each dimension of X and Y represents a linear combination of the dimensions of M plus i.i.d. Gaussian noise (after whitening), the redundant information loosely captures the extent to which the measurement vectors associated with X are aligned with those of Y. Under this interpretation, we would expect there to be less alignment when there are fewer vectors (d_X, d_Y are small) of larger dimensionality (d_M is large). That may be why (S3) and (S4) have the least redundancy, since X has

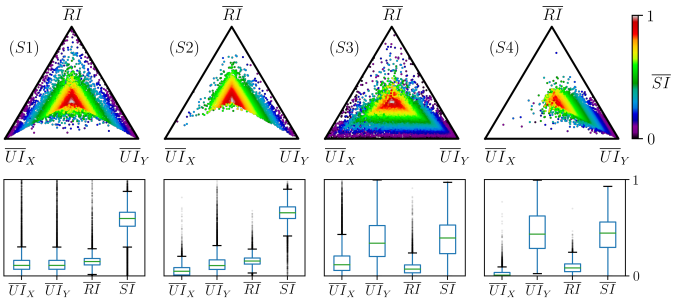

Fig. 1: Distribution of normalized unique, redundant, and synergistic information for Gaussian systems sampled from the standard Wishart distribution: (S1) $d_M = d_X = d_Y$; (S2) $d_M < d_X \leq d_Y$; (S3) $d_X \leq d_Y \leq d_M$; and (S4) $d_X \leq d_M < d_Y$. Top row: scatter plots of the computed $\overline{UI}_X$, $\overline{UI}_Y$, $\overline{RI}$ and $\overline{SI}$ on the 4-simplex (3D views in Appendix G). Bottom row: box plots showing the relative prevalence of each partial information atom.

very few dimensions to begin with in both these cases. Finally, because we have set $d_Y \geq d_X$, we see that Y tends to have more unique information than X. The only exception is that of (S1), where $d_X = d_Y$. Note how (S2) closely mimics the scalar-M case, rarely exhibiting unique information in both X and Y simultaneously. (S3) and (S4) have large amounts of unique information in Y, since d_M is large in both these cases, implying that there are more dimensions of M that X and Y might uniquely capture. This effect is particularly pronounced in (S4), as Y has enough dimensions to capture M, while X rarely captures any part of M uniquely.

VI. Discussion and Open Questions

The most important implication of Barrett's work [1] was that any PID satisfying the Blackwell criterion could be *easily computed* for Gaussian distributions with scalar M. We showed in Theorem 3 that this extends to vector M under a specific condition. We also provided a means to approximate the δ-PID when the aforementioned condition does not hold.

While we believe that our proposed approximation provides the best existing method for computing bounds on the PID in multivariate Gaussian systems, there are shortcomings with the proposed approach. Most importantly, we have not been able to prove that our approximate PID is always non-negative. This is not a huge issue in practice, as negative estimates of PID quantities can be discarded and replaced by the bounds implied by the basic PID equations. But proving that the $\widehat{\delta}_G$-PID is a valid PID could have implications for bounding it between the $\sim$-PID and δ-PID. A second shortcoming is that the $\widehat{\delta}_G$-PID is not guaranteed to satisfy the Blackwell criterion as a result of not having an upper bound on the approximate deficiency. The simulation results are favorable for both of these issues, but formal proofs are still desirable.

A number of questions regarding the computation of deficiency remain, including whether it can be computed exactly and conditions under which (8) and (14) are equivalent. Applying the $\widehat{\delta}_G$-PID in practice also requires studying its statistical properties when estimating covariances from data and providing statements of confidence. Finally, another future direction involves extending Blackwell sufficiency from Gaussian channels to general continuous channels [16].

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. 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," p. 307, 2018.

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

[10] 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.

[11] 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.

[12] 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.

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

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

[15] L. Le Cam, "Sufficiency and approximate sufficiency," *Ann. Math. Statist.*, vol. 35, no. 4, pp. 1419–1455, 12 1964. [Online]. Available: https://doi.org/10.1214/aoms/1177700372

[16] 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.

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

Appendix A
Proof of Lemma 2

Proof of Lemma 2. 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.

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

$$P_{X'|M} = P_{X|M} \text{ and } P_{YX'|M} = P_{Y|X'}P_{X'|M}. \tag{28}$$

Let $P_{Y'|X} \triangleq P_{Y|X}$. Then, (28) implies that

$$P_{Y'|X}P_{X|M} = P_{Y|X'}P_{X'|M}. \tag{29}$$

Therefore,

$$\int P_{Y'|X}P_{X|M}\,dx = \int P_{Y|X'}P_{X'|M}\,dx \tag{30}$$

$$= \int P_{YX'|M}\,dx \;=\; P_{Y|M}, \tag{31}$$

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

$$\int P_{Y'|X}P_{X|M}\,dx = P_{Y|M}. \tag{32}$$

In other words,

$$P_{Y'|M} = P_{Y|M} \text{ and hence } P_{Y'M} = P_{YM}. \tag{33}$$

Let X' be defined through a stochastic transformation of Y and M: $P_{X'|YM}(x|y,m) \triangleq P_{X|Y'M}(x|y,m)\ \forall\ x \in \mathsf{X}, y \in \mathsf{Y}, m \in \mathsf{M}$. Then, using (33), we find

$$P_{X'YM} = P_{X'|YM}P_{YM} \tag{34}$$

$$= P_{X|Y'M}P_{Y'M} = P_{XY'M}. \tag{35}$$

This in turn implies

$$P_{X'Y|M} = P_{XY'|M} \overset{(a)}{=} P_{Y'|X}P_{X|M} \tag{36}$$

$$\overset{(b)}{=} P_{Y|X'}P_{X'|M}, \tag{37}$$

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 (35). □

Appendix B
Proof of Theorem 3

Before proceeding to the proof, we introduce a lemma borrowed from Shang and Poor [12, Lemma 5] that provides an equivalent characterization of the condition in Theorem 3.

Lemma 6. *The condition $H_X^T\Sigma_X^{-1}H_X \succcurlyeq H_Y^T\Sigma_Y^{-1}H_Y$ holds if and only if*

$$\exists\ T \quad s.t. \quad H_Y = TH_X \ \text{ and } \ T\Sigma_X T^T \preccurlyeq \Sigma_Y. \tag{38}$$

Proof. Consider the whitened form of the channels:

$$\widetilde{H}_X \triangleq \Sigma_X^{-\frac{1}{2}}H_X,\ \widetilde{H}_Y \triangleq \Sigma_Y^{-\frac{1}{2}}H_Y. \tag{39}$$

Then, we need to show $\widetilde{H}_X^T\widetilde{H}_X \succcurlyeq \widetilde{H}_Y^T\widetilde{H}_Y$ if and only if

$$\exists\ T \quad \text{s.t.} \quad \widetilde{H}_Y = T\widetilde{H}_X \quad \text{and} \quad TT^T \preccurlyeq I. \tag{40}$$

The remainder of the proof follows from [12, Lemma 5]. □

Proof of Theorem 3. This proof is derived in large part from the work of Gerdes et al. [11].

($\Leftarrow$) Suppose that $H_X^T\Sigma_X^{-1}H_X \succcurlyeq H_Y^T\Sigma_Y^{-1}H_Y$. Then, by Lemma 6, $\exists\ T$ such that $H_Y = TH_X$ and $T\Sigma_X T^T \preccurlyeq \Sigma_Y$. In other words, we may write

$$Y' = TX + N, \tag{41}$$

where $N \sim \mathcal{N}(0, \Sigma_T)$, with $\Sigma_T \triangleq \Sigma_Y - T\Sigma_X T^T$ (the fact that $T\Sigma_X T^T \preccurlyeq \Sigma_Y$ 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

$$P_{Y'|M} = \mathcal{N}(TH_X M,\ T\Sigma_X T^T + \Sigma_T) \tag{42}$$

$$= \mathcal{N}(H_Y M,\ \Sigma_Y) = P_{Y|M}. \tag{43}$$

Hence, by Definition 2, $X \succcurlyeq_M Y$.

($\Rightarrow$) 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,

$$I(M;X) \overset{(a)}{\geq} I(M;Y') \overset{(b)}{=} I(M;Y), \tag{44}$$

where (a) follows from the Data Processing Inequality [14, Ch. 2], and (b) follows because $P_{Y'M} = P_{YM}$. For a Gaussian channel $P_{X|M}$, the mutual information is given by [14, Thm. 8.4.1]

$$\begin{aligned} I(M;X) &= h(X) - h(X \,|\, M) && (45)\\ &= \frac{1}{2}\log|(2\pi e)(\Sigma_X + H_X\Sigma_M H_X^T)| \\ &\qquad - \frac{1}{2}\log|2\pi e\Sigma_X| && (46)\\ &= \frac{1}{2}\log\left(\frac{|\Sigma_X + H_X\Sigma_M H_X^T|}{|\Sigma_X|}\right) && (47)\\ &= \frac{1}{2}\log\left(\frac{|\Sigma_X||I + \Sigma_X^{-1}H_X\Sigma_M H_X^T|}{|\Sigma_X|}\right) && (48)\\ &= \frac{1}{2}\log|I + \Sigma_X^{-1}H_X\Sigma_M H_X^T|. && (49) \end{aligned}$$

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

$$c^T H_X^T\Sigma_X^{-1}H_X c < c^T H_Y^T\Sigma_Y^{-1}H_Y c. \tag{50}$$

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

$$\begin{aligned} &\log|1 + c^T H_X^T\Sigma_X^{-1}H_X c| < \log|1 + c^T H_Y^T\Sigma_Y^{-1}H_Y c| && (51)\\ \Rightarrow\ &\log|I + \Sigma_X^{-1}H_X cc^T H_X^T| < \log|I + \Sigma_Y^{-1}H_Y cc^T H_Y^T| && (52) \end{aligned}$$

Now, since $cc^T \succcurlyeq 0$, it is a valid covariance matrix. So if we set $\Sigma_M \triangleq cc^T$, we get

$$\begin{aligned} I(M;X) &= \log|I + \Sigma_X^{-1}H_X\Sigma_M H_X^T| && (53)\\ &< \log|I + \Sigma_Y^{-1}H_Y\Sigma_M H_Y^T| && (54)\\ &= I(M;Y). && (55) \end{aligned}$$

However, this contradicts (44), which holds no matter what Σ_M is. Therefore, we must have that $H_X^T\Sigma_X^{-1}H_X \succcurlyeq H_Y^T\Sigma_Y^{-1}H_Y$. □

Appendix C
The Approximate Deficiency is Well Defined

Lemma 7. *For jointly Gaussian random vectors M, X and Y satisfying the assumption in Remark 1, $\widehat{T}$ and $\widehat{\Sigma}_T$ as defined by (17) and (18) yield an approximate deficiency $\widehat{\delta}_G(M : Y \setminus X)$ that is well defined, i.e. the covariance matrix $\widehat{\Sigma}_T + \widehat{T}\widehat{T}^T$ 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 $\widehat{\Sigma}_T + \widehat{T}\widehat{T}^T \succcurlyeq 0$ as well. For ease of notation, let $A \triangleq H_Y\Sigma_M H_Y^T$ and $B \triangleq H_X\Sigma_M H_X^T$. Assume for a contradiction that $\widehat{\Sigma}_T + \widehat{T}\widehat{T}^T$ is rank deficient. This implies that there exists a vector $v \neq \mathbf{0}$ such that:

$$\begin{aligned} & v^T(\widehat{\Sigma}_T + \widehat{T}\widehat{T}^T)v = v^T(I + A - \widehat{T}B\widehat{T}^T)v = 0 \\ \Rightarrow\ & v^T(I + A)v = v^T(\widehat{T}B\widehat{T}^T)v \\ \overset{(a)}{\Rightarrow}\ & (\widehat{T}^T v)^T B(\widehat{T}^T v) > 0 \\ \Rightarrow\ & \widehat{T}^T v \neq \mathbf{0} \end{aligned}$$

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

$$\begin{aligned} & v^T\widehat{\Sigma}_T v \geq 0 \\ \Rightarrow\ & v^T(\widehat{\Sigma}_T + \widehat{T}\widehat{T}^T)v - v^T(\widehat{T}\widehat{T}^T)v \geq 0 \\ \Rightarrow\ & v^T\widehat{T}\widehat{T}^T v = \|\widehat{T}^T v\|_2^2 \leq 0. \end{aligned}$$

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

Appendix D
Proof of Proposition 4

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

($\delta = 0 \Rightarrow \widehat{\delta}_G = 0$) It was shown by Torgersen [13, 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 $\Sigma_T + TT^T \succcurlyeq 0$ is invertible and thus $\Sigma_T + TT^T, (\Sigma_T + TT^T)^{-1} \succ 0$. This implies that the first term of (16) 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 $\Sigma_T^* + T^*T^{*T} = I \Rightarrow I - T^*T^{*T} \succcurlyeq 0$. As such, when $T = T^*$ we have that $\Sigma_T = I + H_Y\Sigma_M H_Y^T - T^*(I + H_X\Sigma_M H_X^T)T^{*T} = I - T^*T^{*T} = \Sigma_T^*$.

($\delta = 0 \Leftarrow \widehat{\delta}_G = 0$) 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 $\widehat{\Sigma}_T = I + H_Y\Sigma_M H_Y^T \Rightarrow \widehat{P}_{Y'|X} \circ P_{X|M} = P_Y \Rightarrow \widehat{\delta}_G = I(M;Y)$. 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:

$$\mathbb{E}_{P_M}\left[\|(\widetilde{T}H_X - H_Y)M\|_A^2\right] < \mathbb{E}_{P_M}\left[\|H_Y M\|_A^2\right] \tag{56}$$

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

$$\mathbb{E}_{P_M}\left[\|(\lambda\widetilde{T}H_X - H_Y)M\|_A^2\right] < \mathbb{E}_{P_M}\left[\|H_Y M\|_A^2\right] \tag{57}$$

for any $\lambda \in (0,1)$. The implication follows from the convexity of the Mahalanobis distance, noting that $\lambda\widetilde{T} = \lambda\widetilde{T} + (1-\lambda)\mathbf{0}$. As such, we can assume without loss of generality that $\widetilde{T}$ can be chosen such that (56) holds and $\widetilde{\Sigma}_T \triangleq A - \widetilde{T}\widetilde{T}^T$ is PSD.

Defining $\widetilde{P}_{Y'|X} \triangleq \mathcal{N}(\widetilde{T}X, \widetilde{\Sigma}_T)$, we note that $\widetilde{P}_{Y'|X} \circ P_{X|M} = \mathcal{N}(\widetilde{T}H_X M, A)$. As such:

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

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

$$\delta_G \leq \mathbb{E}_{P_M}\left[D(P_{Y|M}\|\widetilde{P}_{Y'|X} \circ P_{X|M})\right] < I(M;Y) \leq \delta, \tag{58}$$

which is a contradiction. □

Appendix E
Details on CVX Implementation

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 CVX software package. First we note that the objective function can be rewritten using the trace trick. Letting $A = I + H_Y \Sigma_M H_Y^T$ for ease of notation:

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

where $\|\cdot\|_F^2$ is the squared Frobenius norm. Next, the constraint can be rewritten using the Schur complement [17]:

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

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

$$\begin{aligned}
\widehat{T} = \ \underset{T}{\mathrm{argmin}} \quad & \|A^{-\frac{1}{2}}TH_X\Sigma_M^{\frac{1}{2}} - A^{-\frac{1}{2}}H_Y\Sigma_M^{\frac{1}{2}}\|_F^2 \\
\text{s.t.} \quad & \begin{bmatrix} I + H_Y\Sigma_M H_Y^T & T \\ T^T & (I + H_X\Sigma_M H_X^T)^{-1} \end{bmatrix} \succcurlyeq 0
\end{aligned} \tag{59}$$

Appendix F
Step-By-Step Experimental Procedure

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 $\Sigma_X = \Sigma_Y = I$.
5) Estimate the deficiencies $\widehat{\delta}_G(M : Y \setminus X)$ and $\widehat{\delta}_G(M : X \setminus Y)$, using the the Python CVX package to solve (59).
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 $\widehat{UI}_Y \approx 0$.

APPENDIX G
3-DIMENSIONAL SIMPLEX VIEWS

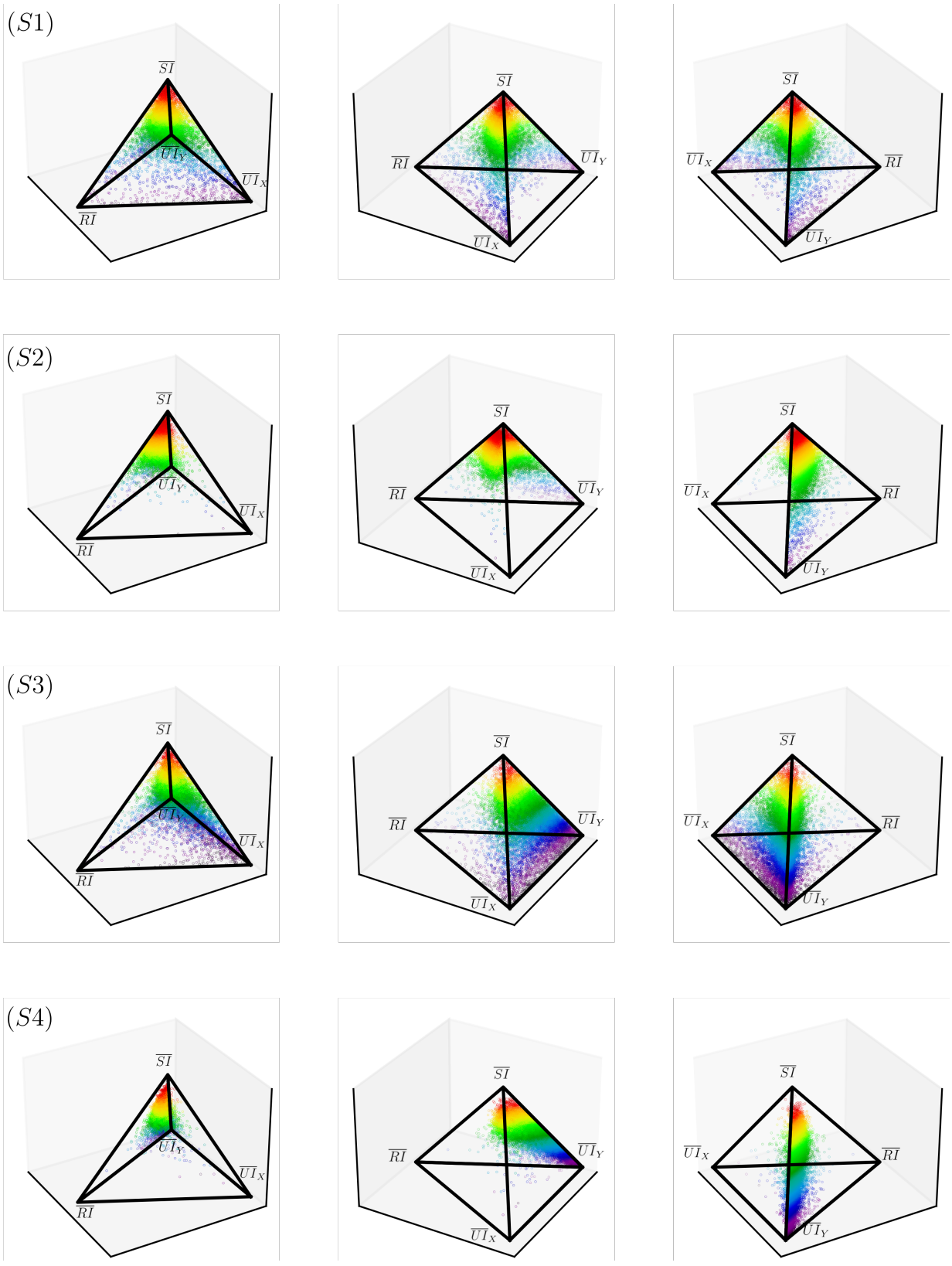

Fig. 2: Alternative views of Figure 1. Each row displays one of the four sampling schemes, and each column provides a rotated view of the simplex.