Duality between Feature Selection and Info-Clustering

Chung Chan, Ali Al-Bashabsheh, Qiaoqiao Zhou and Tie Liu

Abstract

The feature-selection problem is formulated from an information-theoretic perspective. We show that the problem can be efficiently solved by an extension of the recently proposed info-clustering paradigm. This reveals the fundamental duality between feature selection and data clustering, which is a consequence of the more general duality between the principal partition and the principal lattice of partitions in combinatorial optimization.

I. Introduction

Many problems in machine learning are, in essence, the devising of a parametrized model that provides a good approximation to the functional dependency between a set of input variables (features) and an output (dependent) variable.[1] The model parameters are often determined/estimated using a training set of points, where each point is a pair consisting of a sample (i.e., a configuration) of the input variables and the corresponding output value. The set of features often contains irrelevant features to the dependent variable, which results in a high processing complexity and overfitting (due to the limited size of the training set). The feature selection problem is an attempt to resolve the above issues by selecting the features that are most relevant to the dependent variable. This of course raises the two questions of what is meant by "relevant" and how can one determine such relevant features. Shannon's

Preliminary work published in [1].

C. Chan (email: cchan@inc.cuhk.edu.hk, chungc@alum.mit.edu), A. Al-Bashabsheh and Q. Zhou are with the Institute of Network Coding at the Chinese University of Hong Kong, the Shenzhen Key Laboratory of Network Coding Key Technology and Application, China, and the Shenzhen Research Institute of the Chinese University of Hong Kong.

T. Liu is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (email: tieliu@tamu.edu).

The work is supported in part by a grant from University Grants Committee of the Hong Kong Special Administrative Region, China (Project No. AoE/E-02/08), Shenzhen Research Fund (KQCX20130628164008004) and Shenzhen Key Laboratory of Network Coding Key Technology and Application, Shenzhen, China (ZSDY20120619151314964).

The work of T. Liu was supported in part by the National Science Foundation under Grant CCF-13-20237. Part of the work was done while T. Liu was visiting the Institute of Network Coding at the Chinese University of Hong Kong.

The work of C. Chan was supported in part by the University Grants Committee of the Hong Kong Special Administrative Region, China (Project No. 14200714).

[1]Depending on the context, the input variables are some times referred to as features in the machine learning literature and independent variables or regressors in regression analysis. In this work, we will refrain from the use of the term "independent variables" in the context of regression and reserve the term to refer to statistical independence between a set of random variables.

mutual information [2] was considered in [3] for the feature selection problem. It was also recognized that such a natural formulation [3, FRn–k] is impractical without further relaxation, owing to the high computational and sample complexity in estimating the mutual information for a large set of features from data. Hence, subsequent information-theoretic approaches such as [4] have been focusing on finding good heuristics to solve the problem approximately.

Another prominent problem in machine learning is the clustering problem. In a broad sense, this is the problem of dividing a set of objects into groups such that elements in the same group are similar/relevant to each other and elements from different groups are dissimilar/irrelevant to each other. Given a mathematically justifiable notion of similarity/relevance for clustering (see [5] for details), then one may, at least intuitively, provide a satisfying answer to the two questions above. Namely, one can treat the features and the dependent variable as the objects in hand, identify the cluster that contains the dependent variable, and declare the remaining elements in the same cluster as the most relevant features to the dependent variable. While in general this remains an intuition that may lack mathematical rigor, in the special case when the features are statistically independent, we prove a duality theorem between the feature selection and data clustering problems that will provide a precise mathematical explanation of the intuition above.

The underlying pinnings to the feature selection and data clustering duality in this work are two mathematical structures called the principal partition (PP) (see, e.g., [6] for an overview of related works to the PP) and the principal lattice of partitions (PLP) [7] of a submodular function. Both the PP and PLP are polynomial-time computable, as will be pointed out in place. The recognition of a link between the PLP (more precisely, a subset of the PLP) and the clustering problem was made in [8], which led to an efficient algorithm that provides a partial solution to the hard k-clustering problem. The detailed connection was discussed in [5]. In [9] the PP (more precisely, a subset of the PP) was linked to the size-constrained submodular function minimization problem, which led to an efficient algorithm that provides a partial solution to the problem.

In this work, we connect the (entire) PP to the feature selection problem (by showing that an element of the PP is a solution to the feature selection problem) and connect the (entire) PLP to the data clustering problem (by showing that an element of the PLP is a solution to the clustering problem). When the features are independent, we prove a one-to-one correspondence between the PP and PLP, thereby a duality between the feature selection and the clustering problems. (More precisely, the duality is between the solutions of the two problems that are captured by the PP and PLP.)

We remark that the duality result can be extended to more general submodular function. The current duality can be viewed as the special case when the entropy function is taken to be the submodular function and the modularity is the statistical independence among the features. The only other duality we are aware of is in [10], which gave the fastest algorithm at the time for the computation of the PP of a graph. (By first computing the PLP of the graph and then constructing the PP via this duality.) However, that result cannot be put in the same category as the current result because it considers the PLP and PP for different submodular functions of a graph, namely, the graph cut function evaluated over subsets of vertices and the rank function of the cycle matroid of the graph evaluated over

subsets of edges instead of the vertices. The duality result appears to exploit the graphical structure; there seems to be no natural extension of such result beyond graphs.

II. Motivation

As a motivation for the duality result, we will consider a simple example involving two independent random variables X_1 and X_2, and a third random variable Y. For the feature selection problem, let X_1 and X_2 be the features and Y be the dependent variable. One is interested in selecting subsets of the features that are highly correlated with the dependent variable. More precisely, feature $i \in \{1, 2\}$ is the best feature if it maximizes Shannon's mutual information [11]:

$$\max_{i\in\{1,2\}} I(\mathsf{Y} \wedge \mathsf{X}_i).$$

As an illustration, assume the random variables are such that

$$\mathsf{Y} = (\mathsf{X}_1, \mathsf{X}_2) \quad \text{with } I(\mathsf{X}_1 \wedge \mathsf{X}_2) = 0 \text{ and} \tag{2.1a}$$

$$H(\mathsf{X}_1) = 2 > H(\mathsf{X}_2) = 1. \tag{2.1b}$$

Then $I(\mathsf{Y} \wedge \mathsf{X}_i) = H(\mathsf{X}_i)$, and so the first variable X_1 is a better feature than X_2 since it shares more mutual information with the dependent variable Y.

For the data clustering problem, we consider the info-clustering paradigm in [5] which clusters a set of random variables according to their multivariate mutual information. As an example, let Z_0, Z_1 and Z_2 be three random variables that we want to cluster. Given a threshold $\gamma \in \mathbb{R}$, a cluster is a subset

$$B \subseteq \{0, 1, 2\} : |B| > 1, I(\mathsf{Z}_B) > \gamma, \forall B' \supsetneq B, I(\mathsf{Z}_{B'}) \leq \gamma,$$

where $I(\mathsf{Z}_B)$ is the multivariate mutual information (MMI) defined in [12] (to be introduced in (3.2)). In other words, a cluster is an inclusion-wise maximal subset consisting of at least two random variables with strictly more than γ amount of mutual information. In the above, the MMI measures the mutual information among multiple random variables and may be viewed as an extension of Shannon's mutual information from the bivariate to the multivariate case.

For simplicity, consider the example

$$\mathsf{Z}_0 = \mathsf{Y}, \mathsf{Z}_1 = \mathsf{X}_1 \text{ and } \mathsf{Z}_2 = \mathsf{X}_2, \tag{2.2}$$

with Y, X_1, and X_2 satisfying (2.1). Then for $B \subseteq \{0, 1, 2\}$ with $|B| \geq 2$, the MMI can be calculated to be

$$I(\mathsf{Z}_B) = \begin{cases} 0 & B = \{1, 2\} \\ H(\mathsf{X}_1) & B = \{0, 1\} \\ H(\mathsf{X}_2) & B \in \{\{0, 1, 2\}, \{0, 2\}\}. \end{cases} \tag{2.3}$$

For instance, $I(\mathsf{Z}_{\{1,2\}}) = 0$ because $\mathsf{Z}_1 = \mathsf{X}_1$ and $\mathsf{Z}_2 = \mathsf{X}_2$ are independent, i.e., they share no information; $I(\mathsf{Z}_{\{0,2\}}) = I(\mathsf{Z}_0 \wedge \mathsf{Z}_2) = H(\mathsf{X}_2)$ because X_2 is the information shared by Z_0 and Z_1; and $I(\mathsf{Z}_{\{0,1\}}) = I(\mathsf{Z}_0 \wedge \mathsf{Z}_1) = H(\mathsf{X}_1)$ because X_1 is the information shared by Z_0 and Z_1.

The fact that $I(\mathsf{Z}_{\{0,1,2\}}) = H(\mathsf{X}_2)$ requires a more detailed understanding of the MMI. A concrete operational meaning [12] is through the secret key agreement problem, that $I(\mathsf{Z}_{\{0,1,2\}})$ is the maximum secret key rate, called the secrecy capacity, that can be agreed upon mutually among three users who observe privately the discrete memoryless sources Z_0, Z_1 and Z_2 respectively, and who are allowed to discuss publicly. While the secrecy capacity is characterized in [13] by a linear program, it has an alternative information-theoretically more appealing interpretation called the residual independence relation (RIR) discovered in [12, Theorem 5.1] based on result of [14]: The value $H(\mathsf{X}_2)$ of the MMI $I(\mathsf{Z}_{\{0,1,2\}})$ satisfies

$$\left[H(\mathsf{Z}_{\{0,1,2\}}) - H(\mathsf{X}_2)\right] = \sum_{i=0}^{2} \left[H(\mathsf{Z}_i) - H(\mathsf{X}_2)\right],$$

which is called the RIR because the total randomness on the L.H.S. after removing $H(\mathsf{X}_2)$ is equal to the sum of the individual randomness of each random variable on the R.H.S. after removing $H(\mathsf{X}_2)$. The equality can be taken to mean that there is no overlapping (mutual information) left in the residual randomness after removing $H(\mathsf{X}_2)$, and so $H(\mathsf{X}_2)$ reflects the amount of information mutual to the three random variables. This interpretation is information-theoretically appealing because it naturally extends the well-known Venn-diagram interpretation of Shannon's mutual information [15] (See [5, Section III–A]). There is also a more graphical interpretation of the RIR in [5, Section IV-A] for the special private source in (2.2) which forms a Markov chain $\mathsf{Z}_1 - \mathsf{Z}_0 - \mathsf{Z}_2$: Representing the source by the weighted graph

$$① \xrightarrow{I(\mathsf{Z}_0 \wedge \mathsf{Z}_1) = H(\mathsf{X}_1)} ⓪ \xrightarrow{I(\mathsf{Z}_0 \wedge \mathsf{Z}_2) = H(\mathsf{X}_2)} ②,$$

the MMI of $H(\mathsf{X}_2)$ is the smallest threshold such that the graph becomes disconnected after removing edges with weight no larger than the threshold.[2]

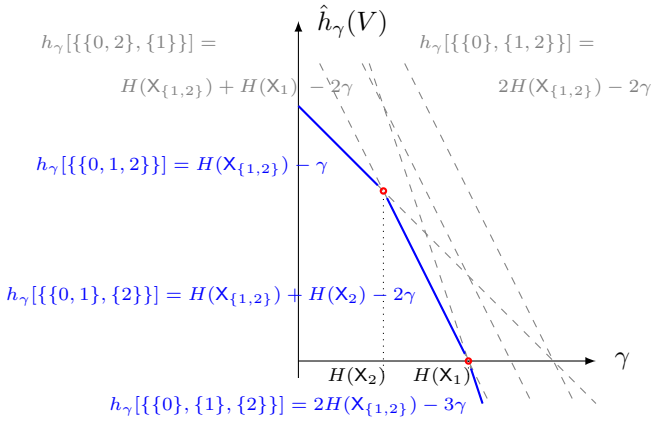
Based on (2.3), for $\gamma < H(\mathsf{X}_2)$, the entire set $\{0,1,2\}$ of random variables is a cluster because it is trivially maximal and it satisfies the required threshold constraint, i.e.,

$$I(\mathsf{Z}_{\{0,1,2\}}) = H(\mathsf{X}_2) > \gamma.$$

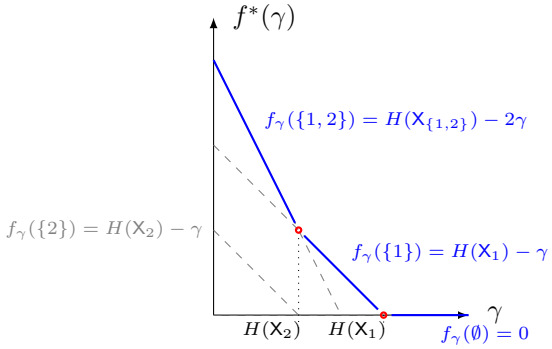
By the same reasoning, when $H(\mathsf{X}_2) \leq \gamma < H(\mathsf{X}_1)$, the set $\{0,1\}$ is a cluster. We remark that even though $\{0,2\}$ satisfies the threshold constraint for $\gamma < H(\mathsf{X}_2)$, it is not considered as a cluster because it is not maximal. More importantly, if the set $\{0,2\}$ were a cluster, then it would be inconsistent with the cluster $\{0,1\}$ which can be taken to assert that Z_0 shares more information with Z_1 (the element in the same cluster) than with Z_2 (the element outside the cluster).

The duality between feature selection and data clustering is simply that: *as the threshold γ increases, the dependent variable clusters with a smaller set of more relevant features.* In the current example, with γ large enough, i.e.,

[2]The source defined in (2.2) also belongs to a different class of graphical sources called the pairwise independent networks (PIN) [16, 17] for which the secrecy capacity was related to tree-packing. Generalizations to hypergraphical source models was considered in [14] and the result was further extended to a general source model with helpers in [18] by relating the problem secret key agreement to the problem of undirected network multicast.



(a) $\hat{h}_\gamma(V)$ vs γ in (3.6) for (2.2).



(b) $f^*(\gamma)$ vs γ in (4.3) for (2.1).

Fig. 1: Plots of (3.6) and (4.3) for the example (2.1) and (2.2) under the mapping (5.1).

exceeding $H(\mathsf{X}_2)$, the better feature X_1 is identified by the cluster $\{0,1\}$, which groups the dependent variable $\mathsf{Z}_0 = \mathsf{Y}$ with the feature $\mathsf{Z}_1 = \mathsf{X}_1$. In this work, we extend the duality result to the case allowing any number of independent random variables (as features) and any correlation between the dependent variable and features.

III. Info-clustering formulation

In this section, we first introduce the general info-clustering formulation in [5] and then extend it slightly for the desired duality result. The framework considers any number of random variables with any joint distribution. More precisely, let $\mathsf{Z}_V := (\mathsf{Z}_i \mid i \in V)$ be a finite vector of random variables to be clustered, and let $\Pi(V)$ be the collection of partitions of V into non-empty disjoint sets. The set of clusters at a real-valued threshold $\gamma \in \mathbb{R}$ is defined in [5, Definition 2.1] as

$$\mathcal{C}_\gamma(\mathsf{Z}_V) := \{B \subseteq V \mid |B| > 1, I(\mathsf{Z}_B) > \gamma, \tag{3.1a}$$

$$\nexists B' \supsetneq B, I(\mathsf{Z}_{B'}) > \gamma\}, \tag{3.1b}$$

where $I(\mathsf{Z}_B)$ is the MMI defined as [12]

$$I(\mathsf{Z}_B) := \min_{\substack{\mathcal{P}\in\Pi(B):\\ |\mathcal{P}|>1}} \frac{1}{|\mathcal{P}|-1} \underbrace{D\left(P_{\mathsf{Z}_V} \middle\| \prod_{C\in\mathcal{P}} P_{\mathsf{Z}_C}\right)}_{=\sum_{C\in\mathcal{P}} H(\mathsf{Z}_C)-H(\mathsf{Z}_B)}. \tag{3.2}$$

In the bivariate case when $V = \{1,2\}$, the MMI reduces to Shannon's mutual information $I(\mathsf{Z}_1 \wedge \mathsf{Z}_2)$ with $\mathcal{P} = \{\{1\},\{2\}\}$. *The MMI naturally extends Shannon's mutual information to the multivariate case*, with concrete operational meanings in secret key agreement and undirected network coding [12, 14, 18].[3] In the above, (3.1a) *is the threshold constraint that requires the random variables in a cluster to share at least γ amount of information, while the non-existence condition in* (3.1b) *requires the cluster to be inclusion-wise maximal.*

It was shown in [5] that the clustering solution of (3.1) is given by a mathematical structure called the principal lattice of partitions (PLP) introduced by [21]. More precisely, we say that a set function $h : 2^V \to \mathbb{R}$ is submodular [22] if for all $B_1, B_2 \subseteq V$,

$$h(B_1) + h(B_2) \geq h(B_1 \cup B_2) + h(B_1 \cap B_2). \tag{3.3}$$

The function h is said to be supermodular if the inequality above is reversed, and modular if equality holds. It follows that $-h$ is supermodular iff h is submodular, while h is modular iff it is both submodular and supermodular. The entropy function

$$h(B) := H(\mathsf{Z}_B) \quad \text{for } B \subseteq V, \tag{3.4}$$

for instance, is known to be submodular [23], and so is the residual entropy function [12]

$$h_\gamma(B) := h(B) - \gamma. \tag{3.5}$$

More generally, a constant function is modular and the sum of submodular functions is submodular. For the submodular function h_γ, the Dilworth truncation [22] (evaluated at V) is

$$\hat{h}_\gamma(V) := \min_{\mathcal{P}\in\Pi(V)} \underbrace{\sum_{C\in\mathcal{P}} \overbrace{[H(\mathsf{Z}_C)-\gamma]}^{h_\gamma(C)}}_{h_\gamma[\mathcal{P}]:=}. \tag{3.6}$$

The set of optimal partitions to (3.6), i.e., the partitions attaining the minimization, for different values of $\gamma \in \mathbb{R}$ is called the *principal lattice of partitions* (PLP) [21]. As an illustration, Fig. 1a shows a plot of $\hat{h}_\gamma(V)$ against γ for the example in (2.2). For $\gamma \leq H(\mathsf{X}_2) = 1$, the trivial partition $\{\{0,1,2\}\}$ is optimal, i.e., $\hat{h}_\gamma(V) = h_\gamma[\{\{0,1,2\}\}]$. For $\gamma \in [H(\mathsf{X}_2), H(\mathsf{X}_1)] = [1,2]$, the partition $\{\{0,1\},\{2\}\}$ is optimal. For $\gamma \geq H(\mathsf{X}_1) = 2$, the partition $\{\{0\},\{1\},\{2\}\}$ into singletons is optimal.

[3]The MMI was inspired by a more general divergence bound on the secrecy capacity in [13]. In [14], the bound was shown to be loose in general with helpers and identified to be tight in the no-helper case that gives the current definition of the MMI. The earliest proof of slackness and tightness of the bound, and the attempt to interpret the MMI as a measure of mutual information can be found in [19]. The MMI was also called "shared information" in [20].

To elaborate, for any submodular function h, the set of optimal partitions to (3.6) for any γ forms a lattice called the Dilworth truncation lattice, and the sequence of Dilworth truncation lattices forms a larger lattice which is referred to as the PLP [21]. The lattice structure is respective to the partial order on partitions, denoted as $\mathcal{P} \preceq \mathcal{P}'$, meaning that

$$\forall C \in \mathcal{P}, \exists C' \in \mathcal{P}' \text{ such that } C \subseteq C'. \tag{3.7}$$

In other words, $\mathcal{P}'$ is no smaller than $\mathcal{P}$ means that $\mathcal{P}'$ is no finer than $\mathcal{P}$. We use $\prec$ to denote the strict inequality when $\mathcal{P} \neq \mathcal{P}'$. For instance, the optimal partitions in Fig. 1a form a chain, which is a special kind of lattice:

$$\{\{0,1,2\}\} \succ \{\{0,1\},\{2\}\} \succ \{\{0\},\{1\},\{2\}\}.$$

The PLP turns out to be strongly polynomial-time solvable [8, 21], and it resolves the clustering problem in hand:

Proposition 3.1 ([5, Corollary 3.1]) *For any threshold $\gamma \in \mathbb{R}$, the clusters of $\mathcal{C}_\gamma$ (3.1) are the non-singleton elements of the finest optimal partition of (3.6) with respect to the partial order (3.7).* □

This can be observed in Fig. 1a. For instance, for $\gamma \in [H(\mathsf{X}_2), H(\mathsf{X}_1)]$, the partition $\{\{0,1\},\{2\}\}$ is optimal and its non-singleton element $\{0,1\}$ is a cluster (as mentioned before).

By the above proposition, the clusters can be obtained from the optimal partitions, or more precisely, the finest optimal partitions to (3.6). In general, the finest optimal partitions form a chain called the *principal sequence of partitions* (PSP), which is a subset of the PLP [21]. In other words, the info-clustering solutions arise from the partitions in the PSP, which in general can be a proper subset of the PLP. As we will argue shortly, the additional partitions in the PLP (compared to the PSP) are meaningful and will potentially enrich the solutions of the data clustering and feature selections problems. In the following, we first extend the clustering formulation of [5] to include the entire PLP as solutions:

Definition 3.1 For a threshold $\gamma \in \mathbb{R}$, the extended set of clusters is defined as

$$\overline{\mathcal{C}}_\gamma(\mathsf{Z}_V) := \{B \subseteq V \mid |B| > 1, I(\mathsf{Z}_B) \geq \gamma, \tag{3.8a}$$

$$\nexists B' \subseteq V, \emptyset \neq \underbrace{B \cap B' \neq B'}_{\text{or equiv. } B \not\supseteq B'}, I(\mathsf{Z}_{B'}) > \gamma\}, \tag{3.8b}$$

where $I(\mathsf{Z}_B)$ is as defined in (3.2). □

The following result shows that the extended set of clusters maps to the entire PLP as desired.

Theorem 3.1 *The clusters in $\overline{\mathcal{C}}_\gamma(\mathsf{Z}_V)$ are the non-singleton elements of the optimal partition of (3.6).* □

PROOF See Appendix A. ■

The difference between the two formulations is the non-existence condition in (3.8b), which can be viewed as a relaxation of the inclusion-wise maximality constraint in (3.1b). More precisely, with Proposition 3.1, it follows that $\mathcal{C}_\gamma(\mathsf{Z}_V) \subseteq \overline{\mathcal{C}}_\gamma(\mathsf{Z}_V)$. However, the extended set of clusters may be strictly larger: The non-existence condition (3.8b)

forbids a set B' with at least one element in B and one element outside B (i.e, B bisects B') while having a mutual information strictly larger than γ; it potentially allows $C_1, C_2 \in \overline{\mathcal{C}}_\gamma(\mathsf{Z}_V)$ such that $I(\mathsf{Z}_{C_1}) = I(\mathsf{Z}_{C_2}) = \gamma$ but $C_2 \supsetneq C_1$. This allowed scenario is excluded in (3.1) even if the threshold constraint is changed to non-strict inequality.[4] For instance, consider the example (2.2) with (2.1a) and

$$H(\mathsf{X}_1) = H(\mathsf{X}_2) = 1 \tag{3.9}$$

instead of (2.1b). Then, as γ increases to 1, the set $\mathcal{C}_\gamma(\mathsf{Z}_V)$ of clusters changes from $\{\{0,1,2\}\}$ to the empty set $\emptyset$, i.e., we have $\mathcal{C}_1(\mathsf{Z}_{\{0,1,2\}}) = \emptyset$, which can be seen by noting that for $\gamma = 1$, the finest optimal partition is the partition into singletons. In contrast, one can show that as γ increases to 1, the extended set $\overline{\mathcal{C}}_1(\mathsf{Z}_{\{0,1,2\}})$ of clusters changes from $\{\{0,1,2\}\}$ to further include the sets $\{0,1\}$ and $\{0,2\}$, i.e., we have $\overline{\mathcal{C}}_1(\mathsf{Z}_{\{0,1,2\}}) = \{\{0,1,2\},\{0,1\},\{0,2\}\}$. More generally, the additional clusters in the extended set can be characterized as follows:

Corollary 3.1 *For any $\gamma \in \mathbb{R}$, we have $B \in \overline{\mathcal{C}}_\gamma(\mathsf{Z}_V) \setminus \mathcal{C}_\gamma(\mathsf{Z}_V)$ iff $|B| > 1$, $I(\mathsf{Z}_B) = \gamma$ and*

$$B \cap B' = \emptyset \quad \text{or} \quad B' \subseteq B, \tag{3.10}$$

for all $B' \in \mathcal{C}_\gamma(\mathsf{Z}_V)$ (or simply with $I(\mathsf{Z}_{B'}) > \gamma$). □

PROOF See Appendix A. ■

(3.10) means that a cluster $B \in \overline{\mathcal{C}}_\gamma(\mathsf{Z}_V) \setminus \mathcal{C}_\gamma(\mathsf{Z}_V)$ is consistent with the clusters in $\mathcal{C}_\gamma(\mathsf{Z}_V)$ in the sense that such a cluster B with γ amount of mutual information does not break apart any cluster $B' \in \mathcal{C}_\gamma(\mathsf{Z}_V)$ that has a strictly larger amount of mutual information than γ.

As an illustration, recall that the earlier example forms a Markov tree, and so the statistical dependency can be viewed as a chain ① $\overset{1}{\text{———}}$ ⓪ $\overset{1}{\text{———}}$ ② with edges of unit weight [5, Section IV]. The extended set of clusters return all the sub-chains of unit weight edges, namely ① $\overset{1}{\text{———}}$ ⓪ and ⓪ $\overset{1}{\text{———}}$ ② , as the clusters at $\gamma = 1$, in addition to the trivial cluster consisting of all the nodes. It can be seen that the extended set of clusters give more flexibility in the sense of finding a cluster of an appropriate size for the application of interest.

To summarize, if the application of interest demands a cluster of smaller size than the sizes available at a given threshold, there is no particular reason why one should not increase the threshold to identify a cluster of the desired size. (Assuming a cluster of such size is in the extended set of clusters for some threshold.) For the earlier example, even though $\{0,1\}$ and $\{0,2\}$ are not in the extended set of clusters for $\gamma < 1$, they may be considered if a cluster of size $2 < 3$ is desired. Note that $\{1,2\}$ is not a cluster because it is not consistent with (breaks apart) $\{0,1\}$, $\{0,2\}$ and therefore $\{0,1,2\}$, each of which have a strictly larger mutual information than $\{1,2\}$.

[4] A non-strict inequality for (3.1) will only shift the clustering solution very slightly, i.e., $\mathcal{C}_\gamma(\mathsf{Z}_V)$ will be changed to the one-sided limit $\lim_{\gamma' \uparrow 0} \mathcal{C}_{\gamma'}(\mathsf{Z}_V)$.

IV. Feature Selection Formulation

Let $\mathsf{X}_U := (\mathsf{X}_i \mid i \in U)$ be a finite vector of mutually independent random variables referred to as the features, and Y be a random variable that depends on X_U. The joint distribution of X_U and Y can be written as

$$P_{\mathsf{X}_U\mathsf{Y}} = P_{\mathsf{Y}|\mathsf{X}_U} \prod_{i\in U} P_{\mathsf{X}_i}. \tag{4.1}$$

For a non-negative integer k, if we are to select k features as the most relevant ones to Y, then it is natural to choose the set that maximizes the mutual information

$$\max\{I(\mathsf{Y} \wedge \mathsf{X}_B) \mid B \subseteq U, |B| = k\}. \tag{4.2}$$

This information-theoretic formulation for feature selection first appeared in [3, FRn–k], and will be referred to as the size-constrained formulation (since the size of the set of features to be selected is fixed). Note that

$$\begin{aligned} I(\mathsf{Y} \wedge \mathsf{X}_B) &= H(\mathsf{X}_B) - H(\mathsf{X}_B|\mathsf{Y}) \\ &= \sum_{i\in B} H(\mathsf{X}_i) - H(\mathsf{X}_B|\mathsf{Y}) \end{aligned}$$

by (4.1), which is supermodular in B because $H(\mathsf{X}_B|\mathsf{Y})$ is submodular and $\sum_{i\in B} H(\mathsf{X}_i)$ is modular in B.

Unfortunately, maximizing a supermodular function as in (4.2) (or minimizing a submodular function) under a cardinality constraint is NP-hard in general as it generalizes [7, Section 10.4.4][9] the dense k-subgraph problem, e.g., see [24]. Therefore, we consider a relaxation that can be solved in strongly polynomial time: Given a threshold $\gamma \in \mathbb{R}$, the preferred sets of features achieve the objective

$$f^*(\gamma) := \max_{B\subseteq U} \underbrace{\overbrace{I(\mathsf{Y} \wedge \mathsf{X}_B)}^{f(B):=} - \gamma|B|}_{f_\gamma(B):=}. \tag{4.3}$$

Intuitively, for $\gamma > 0$, the second term $-\gamma|B|$ is a penalty in favor of a smaller set of features. The closely related expression $f^*(\gamma) + \gamma k$ is the well-known Lagrangian dual of (4.2), which can serve as an upper bound of (4.2). The optimal solutions of (4.2) are related to those of the Lagrangian dual (and therefore (4.3)) as follows:

Proposition 4.1 *If B^* is optimal to (4.3) for some γ, then it is also optimal to (4.2) with $k = |B^*|$. (This holds even for dependent features, i.e., without the independence assumption (4.1).)* □

Proof Suppose to the contrary that there exists B' with $|B'| = |B^*|$ but $I(\mathsf{X}_{B'} \wedge \mathsf{Y}) > I(\mathsf{X}_{B^*} \wedge \mathsf{Y})$, then $f_\gamma(B') > f_\gamma(B^*)$, contradicting the optimality of B^*. ■

As an illustration, Fig. 1b is a plot of $f^*(\gamma)$ against γ for the example in (2.1). For $\gamma \le H(\mathsf{X}_2)$, the entire set $\{1, 2\}$ of features is the optimal solution to (4.3) achieving the maximum value of $f_\gamma(\{1, 2\})$. It is also the optimal solution to (4.2) for $k = 2$. For $\gamma \in [H(\mathsf{X}_2), H(\mathsf{X}_1)]$, the set $\{1\}$ is optimal to (4.3) and it is also the optimal solution to (4.2) for $k = 1$. For $\gamma \ge H(\mathsf{X}_1)$, the empty set $\emptyset$ is optimal to (4.3) and trivially optimal to (4.2) for $k = 0$.

The reason we regard (4.3) as a relaxation of (4.2) is because the converse of Proposition 4.1 does not hold in general, i.e., it is possible to find an example where an optimal solution to (4.2) for some integer k is not optimal to (4.3) for any γ. Such an example is given in Appendix B.

For a general supermodular function f, the set of optimal solutions to (4.3) for different values of γ forms a finite distributive lattice with respect to set inclusion [25]. By Birkhoff's representation theorem, the lattice can be characterized using a partial order over the elements of a partition of V. This structure was shown to be polynomial-time computable and is called the *principal partition* (PP). For the detailed definition and historical development of the concept, we refer the readers to [6, 25, 26].[5] In particular, the optimal solutions in Fig. 1b form a chain, which is a special kind of lattice:

$$\{0,1,2\} \supseteq \{0,1\} \supseteq \emptyset.$$

There is a closely related relaxation in [9, (2)] of the general size-constrained submodular function minimization problem.[6] Our relaxation (4.3) is simpler. It appeared as an intermediate step [9, (3)] that contains all the solutions of [9, (2)] (with the non-negative submodular function therein chosen to be $B \mapsto H(\mathsf{X}_B|\mathsf{Y})$). Another difference is that we consider the entire PP as solutions to the feature selection problem while [9] restricts only to the inclusion-wise maximal and minimal subsets to the general size-constrained optimization. As a result, our formulation can give more optimal solutions to (4.2) that are also meaningful. For example, consider (2.1) but with (2.1b) replaced by (3.9) $H(\mathsf{X}_1) = H(\mathsf{X}_2) = 1$. In this case, the features X_1 and X_2 are equally good as each of them contains the same amount (1 bit) of mutual information with Y. It can be shown that, for $\gamma = 1$, both $\{1\}$ and $\{2\}$ are optimal solutions to (4.3) (in addition to the optimal solutions $\emptyset$ and $\{1,2\}$). Thus, for $k = 1$, both $\{1\}$ and $\{2\}$ are solutions to (4.2) as desired. However, the relaxation in [9] considers only the minimal solution $\emptyset$ and maximal solution $\{1,2\}$ to (4.3), which therefore fails to give any solution to (4.2) for $k = 1$.

V. The Duality

The solutions to the data clustering and feature selection problems can be related by the following mapping:

$$V = \{0\} \cup U \quad \text{and} \quad \mathsf{Z}_i = \begin{cases} \mathsf{Y} & i = 0 \\ \mathsf{X}_i & i \in U, \end{cases} \tag{5.1}$$

where X_U satisfies (4.1), and we assume $0 \notin U$ without loss of generality.

Theorem 5.1 *Under the mapping* (5.1)*, we have for all* $\gamma \in \mathbb{R}$ *and* $B \subseteq U$ *that* B *is an optimal solution to* (4.3) *iff* $\{0\} \cup B$ *is an element of an optimal partition to* (3.6). □

[5] In the literature, the term PP is used to refer to both the distributive lattice and the induced (equivalent) structure consisting of a partial order defined over a partition of the ground set (hence the term PP). In this work, we follow this convention to use the term PP to refer to the distributive lattice.

[6] The idea of the relaxation has appeared in [7, Section 10.4.4], but instead of the size-constrained optimization problem (4.2), a closely-related density problem was considered.

In other words, *the dependent variable* $\mathsf{Z}_0 = \mathsf{Y}$ *is clustered with the set* B *of selected features* $\mathsf{Z}_B = \mathsf{X}_B$. The duality can be observed from Fig. 1 for the example in (2.1) using the mapping (2.2) (which agrees with (5.1)). For $\gamma \leq H(\mathsf{X}_2)$, the set $\{1,2\}$ is optimal in Fig. 1b, and its union $\{0\} \cup \{1,2\}$ with $\{0\}$ is contained by the optimal partition $\{\{0,1,2\}\}$ in Fig. 1a. For $\gamma \in [H(\mathsf{X}_2), H(\mathsf{X}_1)]$, the optimal subset $\{1\}$ in Fig. 1b union $\{0\}$ is contained by the optimal partition $\{\{0,1\},\{2\}\}$ in Fig. 1a. Finally, for $\gamma \geq H(\mathsf{X}_1)$, the optimal partition $\{\{0\},\{1\},\{2\}\}$ in Fig. 1a contains $\{0\} \cup \emptyset$, which is trivially the union of $\{0\}$ and the optimal subset $\emptyset$ in Fig. 1b.

As another example, consider (2.2) again but with (2.1a) and (3.9), i.e., the case when both features X_1 and X_2 are equally good. For $\gamma = 1$, every subset of $\{1,2\}$ is optimal to (4.3). In particular, the solutions $\{1\}$ and $\{2\}$ correspond to the partitions $\{\{0,1\},\{2\}\}$ and $\{\{1\},\{0,2\}\}$, which are optimal to (3.6). This is in alignment with Theorem 5.1. Note that neither of these optimal partitions is the finest optimal partition, i.e., the partition $\{\{0\},\{1\},\{2\}\}$, and so Proposition 3.1 dictates that neither $\{0,1\}$ nor $\{0,2\}$ is a cluster according to (3.1). Nevertheless, the duality result here is more general and the discrepancy is resolved via the extended clustering formulation in (3.8), where as mentioned earlier, the sets $\{0,1\}$ and $\{0,2\}$ are indeed in the collection of extended clustering solutions.

Before proving the theorem, we first specialize the clustering solution under the current mapping (5.1) by exploiting the independence among the features (4.1). For $C \subseteq V$, define the C-block partition of V as

$$\mathcal{P}_C := \{C\} \cup \{\{i\} \mid i \in V \setminus C\}. \tag{5.2}$$

Proposition 5.1 *For* $\gamma > 0$*, any optimal* $\mathcal{P}$ *to* (3.6) *under* (5.1) *must satisfy* $\mathcal{P} = \mathcal{P}_{\{0\}\cup B}$ (5.2) *for some* $B \subseteq U$.□

PROOF Suppose to the contrary that an optimal $\mathcal{P}$ to (3.6) contains

$$C' \in \mathcal{P} : 0 \notin C', |C'| > 1.$$

Define another partition of V as

$$\mathcal{P}' = (\mathcal{P} \setminus C') \cup \{\{i\} \mid i \in C'\}.$$

Then, the difference $h_\gamma[\mathcal{P}'] - h_\gamma[\mathcal{P}]$ is

$$\underbrace{\sum_{i \in C'} H(\mathsf{Z}_i) - H(\mathsf{Z}_{C'})}_{=0 \text{ by (5.1) and (4.1).}} - \underbrace{(|C'| - 1)}_{>0} \underbrace{\gamma}_{>0} < 0,$$

which contradicts the optimality of $\mathcal{P}$. ■

PROOF (THEOREM 5.1) we will break down the proof into three cases:

1) $\gamma > 0$: In this case, we relate (4.3) and (3.6) directly by rewriting the terms in (4.3) using $\mathcal{P}_{\{0\}\cup B}$ (5.2) for

$B \subseteq U$.

$$\begin{aligned} |B| &= |U| - |U \setminus B| \\ &= |U| - |\mathcal{P}_{\{0\}\cup B}| + 1 \\ I(Y \wedge \mathsf{X}_B) &= H(\mathsf{Y}) + \underbrace{H(\mathsf{X}_B)}_{\sum_{i\in B} H(\mathsf{X}_i) \text{ by (4.1).}} - \underbrace{H(\mathsf{Y}, \mathsf{X}_B)}_{\mathsf{Z}_{\{0\}\cup B} \text{ by (5.1).}} \\ &= H(\mathsf{Y}) + \sum_{i\in U} H(\mathsf{X}_i) - \sum_{C\in\mathcal{P}_{\{0\}\cup B}} h(C). \end{aligned}$$

Altogether, we have

$$I(\mathsf{Y} \wedge \mathsf{X}_B) - \gamma|B| = t - h_\gamma[\mathcal{P}_{\{0\}\cup B}]$$

where $t := H(\mathsf{Y}) + \sum_{i\in U} H(\mathsf{X}_i) - (|U|+1)\gamma$. Since t is independent of B,

$$\max_{B\subseteq U} I(\mathsf{Y} \wedge \mathsf{X}_B) - \gamma|B| = t - \min_{B\subseteq U} h_\gamma[\mathcal{P}_{\{0\}\cup B}].$$

Since $\gamma > 0$, by Proposition 5.1 the minimization on the R.H.S. above is the same as (3.6), which completes the proof of this case.

2) $\gamma < 0$: By the submodularity of entropy (3.3), we have for any disjoint $C_1, C_2 \subseteq V$ that

$$\begin{aligned} h_\gamma(C_1 \cup C_2) &\le h_\gamma(C_1) + h_\gamma(C_2) - \underbrace{h_\gamma(\emptyset)}_{=-\gamma} \\ &\le h_\gamma(C_1) + h_\gamma(C_2) \end{aligned}$$

for $\gamma \le 0$. This implies that the trivial partition $\mathcal{P} = \{V\}$ is an optimal partition to (3.6) for $\gamma \le 0$ because further partitioning V will not decrease the sum in (3.6). In the current case $\gamma < 0$, the above inequality is strict, and so further partitioning V will increase the sum, and so the trivial partition is indeed the unique optimal solution.

Now, $B = U$ is an optimal solution to (4.3) for $\gamma \le 0$ because $I(\mathsf{Y} \wedge \mathsf{X}_B)$ is non-decreasing in B. In the current case $\gamma < 0$ with strict inequality, the solution is also unique because $|B|$ is strictly increasing in B. Hence, under the mapping (5.1), we have the desired conclusion for the current case that $V = \{0\} \cup U$ is contained by the unique optimal partition $\mathcal{P} = \{V\}$ of (3.6) while $B = U$ is the unique optimal solution to (4.3).

3) $\gamma = 0$: Suppose B is optimal to (4.3). Since U is also optimal, we have

$$I(\mathsf{Y} \wedge \mathsf{X}_B) = I(\mathsf{Y} \wedge \mathsf{X}_U) \tag{5.3}$$

which means that $(\mathsf{Y}, \mathsf{X}_B)$ is independent of $\mathsf{X}_{U\setminus B}$, or equivalently, by (5.1),

$$h(V) = h(\{0\} \cup B) + h(U \setminus B). \tag{5.4}$$

This implies that $\mathcal{P} = \{\{0\}\cup B, U \setminus B\}$ is also optimal for $\gamma = 0$ because $h_0 = h$ and $\mathcal{P} = \{V\}$ is an optimal solution to (3.6) as argued in the previous case.

Conversely, suppose $\{0\}\cup B$ is contained in an optimal partition $\mathcal{P}$ of (3.6) for $\gamma = 0$. Since the trivial partition $\{V\}$ is also optimal as argued in the previous case, we have

$$h(V) = \sum_{C\in\mathcal{P}} h(C) = h(\{0\}\cup B) + \sum_{C\in\mathcal{P}:0\notin C} h(C),$$

which implies (5.4) that $(\mathsf{Y}, \mathsf{X}_B)$ is independent of $\mathsf{X}_{U\setminus B}$, or equivalently (5.3). This completes the proof of the current case because $B = U$ is an optimal solution to (4.3) for $\gamma = 0$ as argued in the previous case. ■

The above proof of the duality result can be extended to a more general submodular function instead of the entropy function. Indeed, the proof of the important case $\gamma > 0$ does not even use submodularity. In comparison, the independence assumption in (4.1) appears to be rather essential in the proof. An example is given in Appendix C to show that the duality can fail without the independence assumption.

VI. Conclusion

In this work, we derived in a rigorous information-theoretic sense an intuitive duality between data clustering and feature selection. The intuition was that features that are clustered with the dependent variable are its most relevant features. We started by considering the info-clustering formulation in [5] using the MMI proposed in [12], then extended the formulation to give a more complete clustering solution that maps to the entire PLP. We also formulated the feature selection problem as a size-constrained submodular function optimization and relaxed it to a form solvable in polynomial-time by computing the PP. The general duality between the PLP and PP was derived, giving the desired duality between data clustering and feature selection.

In the feature selection formulation, the cardinality of a set of feature was considered as the model complexity of selecting that set of feature. However, it may be desirable to consider other cost functions, e.g., the entropy, which reflects the actual amount of information in the set of feature. The features may also be correlated in practice. It is an interesting, but appears non-trivial, task to extend the current result to incorporate other cost functions for the model complexity and allow statistical dependency among the features.

Acknowledgments

The authors would like to thank their colleagues at the Institute of Network Coding (INC) for their insightful comments and relevant discussions.

Appendix A
Proof of Theorem 3.1 and Corollary 3.1

To prove Theorem 3.1, we will make use of the following property of property of the PLP:

Proposition A.1 ([21]) *For $\mathcal{P}_1, \mathcal{P}_2 \in \Pi(V)$ such that $\gamma_1 < \gamma_2$ and $h_{\gamma_i}[\mathcal{P}_i] = \hat{h}_{\gamma_i}(V)$ for $i \in \{1, 2\}$, we have $\mathcal{P}_1 \succeq \mathcal{P}_2$, where "$\succeq$" is the partial order defined in (3.7).* □

This follows from the more elaborate structure of the PLP described in [5, Propositions 3.2 and 3.3], which in turn follows from [21, Theorems 3.5 and 3.7]. These results, and consequently Proposition A.1, are proved using the submodularity of entropy. The rest of the proof of Theorem 3.1 will not rely on the submodularity.

We first show that any element $B \in \overline{\mathcal{C}}_\gamma$ is a non-singleton element of some optimal partition for (3.6).

- Suppose $I(\mathsf{Z}_B) > \gamma$. It can be seen that the non-existence condition in (3.8b) implies the non-existence condition in (3.1b), and so we have $B \in \mathcal{C}_\gamma(\mathsf{Z}_V)$. By Proposition 3.1, B is contained by the finest optimal partition for (3.6) as desired.
- Suppose $I(\mathsf{Z}_B) = \gamma$ instead. Let $\mathcal{P}^*$ be the finest optimal partition to (3.6). Then, for all $C \in \mathcal{P}^* : |C| > 1$, we have $I(\mathsf{Z}_C) > \gamma$ by Proposition 3.1, and so, by the non-existence condition in (3.8b), we have

$$C \subseteq B \text{ for all } C \in \mathcal{P}^* : B \cap C \neq \emptyset.$$

(n.b., the above holds trivially for $|C| = 1$.) Let

$$\mathcal{P}'' := \{C \in \mathcal{P}^* \mid B \cap C \neq \emptyset\}$$

$$\mathcal{P} := (\mathcal{P}^* \setminus \mathcal{P}'') \cup \{B\}.$$

It follows that $\mathcal{P}'' \in \Pi(B)$ with $|\mathcal{P}''| > 1$ and $\mathcal{P} \in \Pi(V)$. By (3.2),

$$\gamma = I(\mathsf{Z}_B) \leq \frac{\sum_{C \in \mathcal{P}''} H(\mathsf{Z}_C) - H(\mathsf{Z}_B)}{|\mathcal{P}''| - 1}, \tag{A.1}$$

which implies that

$$0 \leq h_\gamma[\mathcal{P}''] - h_\gamma(B)$$

$$= h_\gamma[\mathcal{P}^*] - h_\gamma[\mathcal{P}].$$

It follows that $\mathcal{P}$ is also optimal to (3.6) since $\mathcal{P}^*$ is optimal. This completes the proof as $B \in \mathcal{P}$ by construction.

We now show that any non-singleton element B in any optimal partition $\mathcal{P}$ to (3.6) is a cluster in $\overline{\mathcal{C}}_\gamma(\mathsf{Z}_V)$.

- We first argue that $I(\mathsf{Z}_B) \geq \gamma$ as required in (3.8). Suppose to the contrary that $I(\mathsf{Z}_B) < \gamma$. Then, by (3.2), there exists $\mathcal{P}'' \in \Pi(B) : |\mathcal{P}''| > 1$ that satisfies

$$\gamma > I(\mathsf{Z}_B) = \frac{\sum_{C \in \mathcal{P}''} H(\mathsf{Z}_C) - H(\mathsf{Z}_B)}{|\mathcal{P}''| - 1}.$$

Let $\mathcal{P}^* := \mathcal{P} \setminus \{B\} \cup \mathcal{P}'' \in \Pi(V)$. The above inequality implies that

$$0 > h_\gamma[\mathcal{P}''] - h_\gamma(B)$$

$$= h_\gamma[\mathcal{P}^*] - h_\gamma[\mathcal{P}],$$

which contradicts the optimality of $\mathcal{P}$.

- It remains to prove the non-existence condition in (3.8b). Suppose to the contrary that $B' \subseteq V$ exists with $\emptyset \neq B \cap B' \neq B'$ and $I(\mathsf{Z}_{B'}) > \gamma$. In particular, choose an inclusion-wise maximal B', and any γ' from the open interval $(\gamma, I(\mathsf{Z}_{B'}))$ (which is non-empty by assumption). We have $B' \in \mathcal{C}_{\gamma'}(\mathsf{Z}_V)$ by (3.1) and the maximality of B'. By Proposition 3.1, B' is contained by some (finest) optimal partition, say $\mathcal{P}'$, to (3.6). We will argue that $\mathcal{P} \not\succeq \mathcal{P}'$ (see (3.7)), which contradicts the property of the PLP in Proposition A.1 as desired. In particular, $B' \in \mathcal{P}'$ is not contained by B because $B \cap B' \neq B'$ by assumption. B' is not contained by any

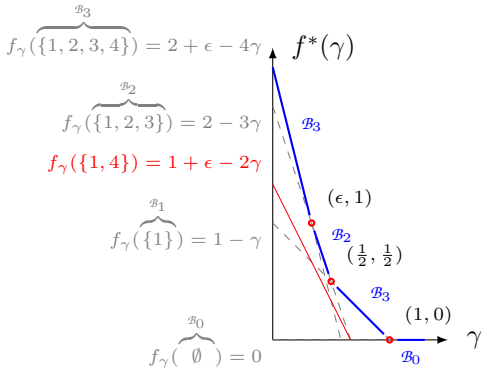


Fig. 2: The plot of $f^*(\gamma)$ vs γ for (B.1).

$C \in \mathcal{P} \setminus \{B\}$ because $B \cap B'$ is non-empty by assumption, and C does not intersect with B and therefore does not contain $B \cap B'$.

Next, we will prove Corollary 3.1. To prove the "only if" case, consider any $B \in \overline{\mathcal{C}}_\gamma(\mathsf{Z}_V) \setminus \mathcal{C}_\gamma(\mathsf{Z}_V)$. By definition (3.8), $|B| > 1$ and $I(\mathsf{Z}_B) \geq \gamma$. If in the contrary that $I(\mathsf{Z}_B) \neq \gamma$, i.e., $I(\mathsf{Z}_B) > \gamma$, (3.8) would imply (3.1), contradicting $B \notin \mathcal{C}_\gamma(\mathsf{Z}_V)$. For any $B' \in \mathcal{C}_\gamma(\mathsf{Z}_V)$, we have $I(\mathsf{Z}_{B'}) > \gamma$ and so the non-existence condition in (3.8b) implies (3.10) as desired.

To prove the "if" case, consider any B satisfying the premise and the finest optimal partition $\mathcal{P}'$ to (3.6). Let

$$\mathcal{P}'' := \{C \in \mathcal{P}' \mid B \cap C \neq \emptyset\}.$$

$\mathcal{P}'' \in \Pi(B)$ because, by Proposition 3.1, the non-singleton elements in $\mathcal{P}''$ are clusters in $\mathcal{C}_\gamma(\mathsf{Z}_V)$, and so they are subsets of B by (3.10). Thus, $\mathcal{P} := \mathcal{P}' \setminus \mathcal{P}'' \cup \{B\}$ is in $\Pi(V)$ and

$$h_\gamma[\mathcal{P}'] - h_\gamma[\mathcal{P}] = h_\gamma[\mathcal{P}''] - h_\gamma(B) \overset{(*)}{\geq} 0,$$

which will complete the proof as this implies that $\mathcal{P}$ is an optimal partition of (3.6) containing B, and so $B \in \mathcal{C}_\gamma(\mathsf{Z}_V)$ by Theorem 3.1. ($B \notin \mathcal{C}_\gamma(\mathsf{Z}_V)$ because $I(\mathsf{Z}_B) = \gamma$ as argued before.) To explain the last inequality (*), consider the non-trivial case $|\mathcal{P}''| > 1$ (because, otherwise, $\mathcal{P}'' = \{B\}$ implies equality for (*)). By assumption,

$$\gamma = I(\mathsf{Z}_B) \leq \frac{\sum_{C \in \mathcal{P}''} H(\mathsf{Z}_C) - H(\mathsf{Z}_B)}{|\mathcal{P}''| - 1},$$

where the last inequality is because $\mathcal{P}''$ is a feasible solution to (3.2). Rearranging the terms give (*) as desired.

APPENDIX B
COUNTER-EXAMPLE FOR THE CONVERSE OF PROPOSITION 4.1

Let $U := \{1, 2, 3, 4\}$ and

$$\begin{aligned} \mathsf{Y} &:= (\mathsf{W}_1, \mathsf{W}_2 \oplus \mathsf{W}_3 \oplus \mathsf{W}_4, \mathsf{W}_5) \\ \mathsf{X}_1 := (\mathsf{W}_1, \mathsf{W}_2), \mathsf{X}_2 &:= \mathsf{W}_3, \mathsf{X}_3 := \mathsf{W}_4, \mathsf{X}_4 := \mathsf{W}_5 \end{aligned} \tag{B.1}$$

where W_i's are independent random bits with $H(\mathsf{W}_i) = 1$ for $i \leq 4$ and $H(\mathsf{W}_5) = \epsilon := \frac{1}{3}$, and $\oplus$ is the XOR operator.

Fig. 2 shows the plot of $f^*(\gamma)$ against γ and also the plots of $f_\gamma(B)$ for the following subsets B:

- Among all the subsets $B \subseteq U$ of size $|B| = 1$, the choice $B = \{1\}$ maximizes the mutual information $I(\mathsf{Y} \wedge \mathsf{X}_B)$ to $H(\mathsf{W}_1) = 1$: $I(\mathsf{Y} \wedge \mathsf{X}_i)$ is 0 for $2 \leq i \leq 3$, and it is $\epsilon < 1$ for $i = 4$.
- Among all the subsets of size 2, the choice $B = \{1, 4\}$ maximizes the mutual information to $H(\mathsf{W}_1, \mathsf{W}_5) = 1 + \epsilon$: all the other mutual information are no larger than 1.
- Among all the subsets of size 3, the choice $B = \{1, 2, 3\}$ maximizes the mutual information to $H(\mathsf{W}_1, \mathsf{W}_2 + \mathsf{W}_3 + \mathsf{W}_4) = 2$: all the other mutual information are no larger than $H(\mathsf{W}_1, \mathsf{W}_5) = 1 + \epsilon < 2$.
- The only subset U of size 4 achieves a mutual information of $H(\mathsf{Y}) = 2 + \epsilon$.

It follows that $B = \{1, 4\}$ is the unique optimal solution to (4.2) for $k = 2$. In Fig. 2, it can be seen that the curve $f_\gamma(\{1, 4\})$ does not touch $f^*(\gamma)$, and so $B = \{1, 4\}$ is not an optimal solution to (4.3) for any γ as desired.

APPENDIX C
EXAMPLE WHERE DUALITY FAILS FOR DEPENDENT FEATURES

Let $U := \{1, 2, 3\}$ and

$$\begin{aligned} \mathsf{Y} &:= (\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3) \\ \mathsf{X}_1 := \mathsf{W}_1, \mathsf{X}_2 &:= (\mathsf{W}_2, \mathsf{W}_4), \mathsf{X}_3 := (\mathsf{W}_3, \mathsf{W}_4) \end{aligned} \tag{C.1}$$

where W_i's are independent random bits with $H(\mathsf{W}_1) = 1 + \epsilon > H(\mathsf{W}_i) = 1$ for $i \geq 2$ and some $\epsilon \in (0, 0.5)$. Note that the independence assumption 4.1 does not hold because $I(\mathsf{X}_2 \wedge \mathsf{X}_3) = 1$.[7]

Note that $\{1, 2\}$ and $\{1, 3\}$ are optimal solutions to (4.2) for $k = 2$ but $\{2, 3\}$ is not, because

$$I(\mathsf{Y} \wedge \mathsf{X}_{\{1,2\}}) = H(\mathsf{W}_{\{1,2\}}) = 2 + \epsilon \quad \text{and}$$

$$I(\mathsf{Y} \wedge \mathsf{X}_{\{1,3\}}) = H(\mathsf{W}_{\{1,3\}}) = 2 + \epsilon \quad \text{but}$$

$$I(\mathsf{Y} \wedge \mathsf{X}_{\{2,3\}}) = H(\mathsf{W}_{\{2,3\}}) = 2 < 2 + \epsilon.$$

By Proposition 4.1, $\{2, 3\}$ cannot be optimal to (4.3) for any value of γ either, while it can be shown that $\{1, 2\}$ and $\{1, 3\}$ are optimal solutions to (4.3) for $\gamma = 1$ (in addition to the solution $\{1\}$ and $\{1, 2, 3\}$).

Under the mapping (5.1), we have

$$\begin{aligned} I(\mathsf{Z}_{\{0,1,2\}}) &= I(\mathsf{Y}, \mathsf{X}_1 \wedge \mathsf{X}_2) = H(\mathsf{W}_2) = 1 \\ I(\mathsf{Z}_{\{0,1,3\}}) &= I(\mathsf{Y}, \mathsf{X}_1 \wedge \mathsf{X}_3) = H(\mathsf{W}_3) = 1 \\ I(\mathsf{Z}_{\{0,2,3\}}) &= \tfrac{H(\mathsf{Y}) + H(\mathsf{X}_2) + H(\mathsf{X}_3) - H(\mathsf{Y}, \mathsf{X}_2, \mathsf{X}_3)}{2} \\ &= \tfrac{H(\mathsf{W}_2) + H(\mathsf{W}_3) + H(\mathsf{W}_4)}{2} = 1.5 > 1. \end{aligned}$$

[7]The source is also a PIN model [16, 17]. It is also possible to give a Markov tree example where the duality fails, e.g., with $U = 1, 2, 3$ and $\mathsf{Y} = (\mathsf{X}_1, \mathsf{X}_2, \mathsf{X}_3)$ where $\mathsf{X}_1 = \mathsf{X}_2$ and X_3 are uniformly random and independent bits. For $\gamma = 1$, it can be shown that $\{0, 1, 2\}$ is a cluster according to (3.8), but $\{1, 2\}$ is an inferior set of feature to $\{1, 3\}$ or $\{2, 3\}$ because $I(\mathsf{Y} \wedge \mathsf{X}_{\{1,2\}}) = 1 < 2 = I(\mathsf{Y} \wedge \mathsf{X}_{\{1,3\}}) = I(\mathsf{Y} \wedge \mathsf{X}_{\{2,3\}})$.

It follows that neither $\{0\} \cup \{1,2\}$ nor $\{0\} \cup \{1,3\}$ is in $\overline{\mathcal{C}}_\gamma$ for any $\gamma \in \mathbb{R}$ because they fail to satisfy (3.8b) (with $B' = \{0,2,3\}$) for $\gamma \leq 1$ and (3.8a) for $\gamma > 1$. This shows that the "only if" statement of the duality result in Theorem 5.1 can fail when (4.1) does not hold. Furthermore, it can be shown that $\{0\} \cup \{2,3\}$ is a cluster in $\mathcal{C}_\gamma(\mathsf{Z}_V)$ (and therefore in $\overline{\mathcal{C}}_\gamma(\mathsf{Z}_V)$) for $\gamma \in [1, 1.5)$. Hence, the "if" statement of Theorem 5.1 also fails to hold for this example.

REFERENCES

[1] C. Chan, A. Al-Bashabsheh, Q. Zhou, and T. Liu, "Duality between feature selection and data clustering," in *54th Annual Allerton Conference on Communication, Control, and Computing, Allerton Retreat Center, Monticello, Illinois*, 2016.

[2] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, July 1948.

[3] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on neural networks*, vol. 5, no. 4, pp. 537–550, 1994.

[4] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520–8532, 2015.

[5] C. Chan, A. Al-Bashabsheh, Q. Zhou, T. Kaced, and T. Liu, "Info-clustering: A mathematical theory for data clustering," *arXiv preprint arXiv:1605.01233*, 2016.

[6] S. Fujishige, "Theory of principal partitions revisited," in *Research Trends in Combinatorial Optimization*. Springer, 2009, pp. 127–162.

[7] H. Narayanan, *Submodular functions and electrical networks*. Elsevier, 1997, vol. 54.

[8] K. Nagano, Y. Kawahara, and S. Iwata, "Minimum average cost clustering." in *NIPS*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 1759–1767.

[9] K. Nagano, Y. Kawahara, and K. Aihara, "Size-constrained submodular minimization through minimum norm base," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 977–984.

[10] S. Patkar and H. Narayanan, "Principal lattice of partitions of submodular functions on graphs: fast algorithms for principal partition and generic rigidity," in *International Symposium on Algorithms and Computation*. Springer, 1992, pp. 41–50.

[11] R. W. Yeung, *Information Theory and Network Coding*. Springer, 2008.

[12] C. Chan, A. Al-Bashabsheh, J. Ebrahimi, T. Kaced, and T. Liu, "Multivariate mutual information inspired by secret-key agreement," *Proc. of the IEEE*, vol. 103, no. 10, pp. 1883–1913, Oct 2015.

[13] I. Csiszár and P. Narayan, "Secrecy capacities for multiple terminals," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, Dec. 2004.

[14] C. Chan and L. Zheng, "Mutual dependence for secret key agreement," in *Proc. of 44th Annual Conf. on Inf. Sciences and Systems*, 2010.

[15] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[16] S. Nitinawarat, C. Ye, A. Barg, P. Narayan, and A. Reznik, "Secret key generation for a pairwise independent network model," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 6482–6489, Dec 2010.

[17] S. Nitinawarat and P. Narayan, "Perfect omniscience, perfect secrecy, and steiner tree packing," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 6490–6500, Dec. 2010.

[18] C. Chan, "The hidden flow of information," in *Proc. IEEE Int. Symp. on Inf. Theory*, St. Petersburg, Russia, Jul. 2011.

[19] ——, "On tightness of mutual dependence upperbound for secret-key capacity of multiple terminals," *arXiv preprint arXiv:0805.3200*, 2008.

[20] P. Narayan and H. Tyagi, "Multiterminal secrecy by public discussion," *Foundations and Trends in Communications and Information Theory*, vol. 13, no. 2-3, pp. 129–275, 2016. [Online]. Available: http://dx.doi.org/10.1561/0100000072

[21] H. Narayanan, "The principal lattice of partitions of a submodular function," *Linear Algebra and its Applications*, vol. 144, no. 0, pp. 179 – 216, 1990.

[22] A. Schrijver, *Combinatorial Optimization: Polyhedra and Efficiency*. Springer, 2002.

[23] S. Fujishige, "Polymatroidal dependence structure of a set of random variables," *Information and Control*, vol. 39, no. 1, pp. 55 – 72, 1978.

[24] U. Feige, D. Peleg, and G. Kortsarz, "The dense k-subgraph problem," *Algorithmica*, vol. 29, no. 3, pp. 410–421, 2001.

[25] S. Fujishige, *Submodular functions and optimization*, 2nd ed. Elsevier, 2005.

[26] ——, “Lexicographically optimal base of a polymatroid with respect to a weight vector,” *Mathematics of Operations Research*, vol. 5, no. 2, pp. 186–196, 1980.