Journal of Machine Learning Research 25 (2024) 1-39 Submitted 4/23; Revised 1/24; Published 1/24

Efficient Modality Selection in Multimodal Learning

Yifei He* YIFEIHE3@ILLINOIS.EDU
University of Illinois Urbana-Champaign

Runxiang Cheng* RCHENG12@ILLINOIS.EDU
University of Illinois Urbana-Champaign

Gargi Balasubramaniam* GARGIB2@ILLINOIS.EDU
University of Illinois Urbana-Champaign

Yao-Hung Hubert Tsai YAOHUNGT@CS.CMU.EDU
Carnegie Mellon University

Han Zhao HANZHAO@ILLINOIS.EDU
University of Illinois Urbana-Champaign

Editor: Francis Bach

Abstract

Multimodal learning aims to learn from data of different modalities by fusing information from heterogeneous sources. Although it is beneficial to learn from more modalities, it is often infeasible to use all available modalities under limited computational resources. Modeling with all available modalities can also be inefficient and unnecessary when information across input modalities overlaps. In this paper, we study the modality selection problem, which aims to select the most useful subset of modalities for learning under a cardinality constraint. To that end, we propose a unified theoretical framework to quantify the learning utility of modalities, and we identify dependence assumptions to flexibly model the heterogeneous nature of multimodal data, which also allows efficient algorithm design. Accordingly, we derive a greedy modality selection algorithm via submodular maximization, which selects the most useful modalities with an optimality guarantee on learning performance. We also connect marginal-contribution-based feature importance scores, such as Shapley value, from the feature selection domain to the context of modality selection, to efficiently compute the importance of individual modality. We demonstrate the efficacy of our theoretical results and modality selection algorithms on 2 synthetic and 4 real-world data sets on a diverse range of multimodal data.

Keywords: Multimodal Learning, Modality Selection, Submodular Optimization, Feature Importance

1. Introduction

Multimodal learning aims to learn from data of different modalities[1] (e.g., images, texts, speech, sensors, etc) by fusing complementary information from different sources, to improve the generalizability and robustness of the underlying model. Compared with unimodal learning, multimodal learning models have shown superior performance in many real-world

∗. Equal contribution.

1. In this paper, we use the terms "modality" and "view" interchangeably.

applications (Bapna et al., 2022; Wu et al., 2021; Huang et al., 2023). The advantages of multimodal learning have also been studied from a theoretical standpoint. For example, prior work shows that learning with more modalities achieves a smaller population risk (Huang et al., 2021). Utilizing cross-modal information can also provably improve prediction in multiview learning (Zhang et al., 2019) and semi-supervised learning (Sun et al., 2020).

However, under the advent of large-scale multimodal deep learning (Huang et al., 2023; OpenAI, 2023; Radford et al., 2021), a major challenge lies in efficient learning on multimodal data. From the modeling perspective, one might be tempted to use all the modalities available, but this is often inefficient or infeasible under limited computational resources. Multimodal data is dense and high-dimensional, and modeling complexity (e.g., model size) can scale linearly or exponentially by the number of input modalities (Zadeh et al., 2017; Liu et al., 2018), which could incur significant consumption in computational and energy resources (e.g., GPUs, electricity). The marginal benefit from new modalities may also decrease as more modalities have been included. Oftentimes, fewer modalities may be sufficient to achieve the desired learning performance. Furthermore, being able to proactively select the most useful modalities can help improve computational efficiency, and reduce the cost of collecting and maintaining inferior modalities. For instance, in sensor placement problem, finding the optimal subset of sensors (modalities) for a learning objective (e.g. temperature or traffic prediction) eliminates the cost of maintaining others (Krause et al., 2011).

To this end, we study the problem of *modality selection* in multimodal learning: *Given a set of input modalities and a size constraint on the selected set, how to select a subset of modalities that yields the optimal learning performance?* There are two main challenges to this problem. First, there is no intuitive and generic way to understand the learning utility of an arbitrary set of modalities. Second, this subset selection problem is $\mathcal{NP}$-hard in general, which is computationally intractable.

To address these challenges, we propose a unified theoretical framework that generically quantifies the learning utility of any set of modalities, and identify proper assumptions that can not only model the dependence relations between input modalities but also allow efficient algorithm designs for modality selection. Under this framework, we can derive simple and efficient modality selection algorithms based on submodular optimization, whose selected subset theoretically possesses an optimality guarantee on the learning performance. Our framework also conveniently connects to feature importance scores from the feature selection domain. We specifically examine the Shapley value and Marginal Contribution Feature Importance (MCI). We show that these feature importance scores, although originally intractable to compute, can be adapted to the context of modality selection as efficient solutions for measuring the importance of individual modality.

We focus on two typical learning settings (classification and regression) to demonstrate the expressiveness and generalizability of our framework. First, we propose a generic utility function for multimodal learning. In classification with cross-entropy loss, a modality's utility is quantified as the Shannon mutual information between the modality and the target. In regression with quadratic loss, a modality's utility is quantified as the variance of the conditional expectation of the target given the modality. We identify approximate conditional and marginal dependence assumptions on the input modalities that can flexibly model the heterogeneous nature of multimodal data. These assumptions parameterize the dependence relations in multimodal data, and enable us to efficiently select a subset of modalities whose

utility has an optimality guarantee via a greedy submodular function maximization algorithm. Specifically, the learning utility of a selected k-size subset will be at least $1 - \frac{1}{e}$ of the true optimal utility minus $k\epsilon$, where ϵ is a constant that parameterizes the conditional dependence between input modalities. We also tackle modality selection as a modality importance ranking problem, by bridging feature importance scores from the feature selection domain to the context of modality selection. In particular, we adopt the Shapley value and MCI scores by applying our utility function as their evaluation function. Under our newly identified conditional and marginal dependence assumptions, the exact solutions to these scores, which are originally intractable to compute for measuring feature importance, can be computed efficiently for measuring the importance of individual modality.

Lastly, we evaluate our theoretical results on one synthetic and two real-world multimodal data sets in each learning setting (6 data sets in total spanning diverse types of modalities). Our evaluation outcome confirms the effectiveness of our framework and algorithms for modality selection.

2. Preliminaries

In this section, we describe the notations, problem setup, and background on submodular optimization and feature importance.

Notation and problem setup. We use X and Y to denote the random variables that take values in input space $\mathcal{X}$ and output space $\mathcal{Y}$, respectively. The instantiation of X and Y is denoted by x and y. We use $\mathcal{H}$ to denote the hypothesis class of predictors from input to output space, and $\hat{Y}$ to denote the predicted outcome. In our settings, $\mathcal{X}$ is multimodal, i.e., $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_k$, where each $\mathcal{X}_i$ is the input from the i-th modality. We use X_i to denote the random variable that takes value in $\mathcal{X}_i$, and V to denote the set of all input modalities, i.e., $V = \{X_1, ..., X_k\}$. For ease of presentation, we often use S and S' to denote two subsets of V.

We study the modality selection problem in both classification and regression settings. In both cases, under a fixed loss function, the predictor/model aims to minimize the loss between the target output and the predicted output. Under a modality cardinality constraint, the goal of modality selection is then to select a subset of input modalities such that the loss is minimized among the given class of predictors.

2.1 Submodular Optimization

Submodularity is a property of set functions with many applications in computer science (Krause et al., 2008; Gomez-Rodriguez et al., 2012; Leskovec et al., 2007). A definition of submodularity is as follows, where 2^V denotes the power set of V, and a set function f assigns each subset $S \subseteq V$ to a value $f(S) \in \mathbb{R}$.

Definition 2.1 (Nemhauser et al. (1978))**.** *Given a finite set V, a set function $f : 2^V \to \mathbb{R}$ is submodular if for any $A \subseteq B \subseteq V$, and $e \in V \setminus B$, we have $f(A \cup \{e\}) - f(A) \geq f(B \cup \{e\}) - f(B)$.*

Namely, adding a new element(s) to a larger set does not yield a larger marginal benefit compared to adding new elements to its subset. One common scenario of optimization on

Algorithm 1: Greedy Maximization of a Submodular Function f

Data: Full set $V = \{X_1, ..., X_k\}$, constraint $q \in \mathbb{Z}^+$.
Input: $f : 2^V \to \mathbb{R}$, and $p \in \mathbb{Z}^+$, where $p \leq q \leq |V|$
Output: Subset S_p
$S_0 = \emptyset$
for $i = 0, 1, ..., p-1$ **do**
  $X^i = \arg\max_{X_j \in V \setminus S_i} (f(S_i \cup \{X_j\}) - f(S_i))$
  $S_{i+1} = S_i \cup \{X^i\}$

the submodular function is submodular function maximization under cardinality constraint. It asks to find a subset $S \subseteq V$ that maximizes $f(S)$ subject to $|S| \leq q$, for some fixed budget q that specifies the largest size of S. Finding the optimal solution to this problem is $\mathcal{NP}$-hard in general. However, Nemhauser et al. (1978) proved that a greedy maximization algorithm (Algorithm 1) can output a subset whose value has an optimality guarantee in polynomial time.

In Algorithm 1, p is the number of iterations the algorithm will run, and q is the cardinality constraint. Algorithm 1 starts with an empty set S_0, and subsequently adds to the current set S_i the new element X^i that maximizes the marginal gain $f(S_i \cup \{X_j\}) - f(S_i)$ at each iteration i. Note that Algorithm 1 runs in time $\mathcal{O}(p|V|)$. The following theorem shows the optimality guarantee of its output under Algorithm 1.

Theorem 2.1 (Nemhauser et al. (1978)). *Let $q \in \mathbb{Z}^+$, S_p be the selected subset from Algorithm 1 at iteration p, and e is the Euler's number, we have:*

$$f(S_p) \geq (1 - e^{-\frac{p}{q}}) \max_{S:|S|\leq q} f(S) \tag{1}$$

Namely, $\max_{S:|S|\leq q} f(S)$ is the optimal value from the optimal subset whose cardinality is at most q. Note that if f is monotone, $\arg\max_{S:|S|\leq q} f(S)$ has cardinality exactly q. Then, the greedily-obtained value of the greedily-obtained set yielded from Algorithm 1 after q iterations is at least $1 - \frac{1}{e}$ ($\sim$0.63) of the optimal value.

2.2 Feature Importance Score

Prior work on feature importance study scoring methods that measure how much each feature contributes to learning. Each feature is treated as a participant in a coalitional game, in which all of them contribute to the total gain. A feature scoring method assigns each feature an importance score by measuring its individual contribution. Many notable feature importance scores are adapted from the Shapley value, which is defined as follows:

Definition 2.2 (Shapley (1952)). *Given a set of all players F in a coalitional game, where $v : 2^F \to \mathbb{R}$ is a set function that evaluates the utility of a set of players, the Shapley value of player i under v is:*

$$\phi_{v,i} = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} (v(S \cup \{i\}) - v(S)) \tag{2}$$

We can interpret the Shapley value of player i as the average marginal contribution of i over all possible subsets to which i can be included. Computing the exact Shapley value of a player is $\sharp\mathcal{P}$-hard as it requires enumerating over all possible subsets of players, which is exponential in the number of players, i.e., $\mathcal{O}(2^{|F|})$ unique subsets. (Roth, 1988; Winter, 2002). Although in certain settings, there are efficient approximations to Shapley value, e.g., Monte Carlo simulation (Faigle and Kern, 1992; Fatima et al., 2008; Michalak et al., 2013). In machine learning, the Shapley value has been used to model feature importance by treating each input feature as a player. The Shapley value can emit different properties under different evaluation functions.

Feature importance scores that are based on Shapley value can underestimate the importance of correlated features, such that all correlated features are assigned lower importance values (Kumar et al., 2020). Thus, we examine an alternative—the Marginal Contribution Feature Importance (MCI) from Catav et al. (2021). MCI of a feature i is the maximum marginal contribution over all possible feature subsets. The complexity of computing the exact MCI of a feature is also exponential in the number of features.

Definition 2.3 (Catav et al. (2021))**.** *Given a set of all features F, and a non-decreasing set function $v : 2^F \to \mathbb{R}$, the MCI of feature i defined by v is:*

$$\phi_{v,i}^{mci} = \max_{S \subseteq F}(v(S \cup \{i\}) - v(S)) \tag{3}$$

3. Theoretical Framework

We propose a utility function that can generically quantify the utility of a given set of modalities in multimodal learning. We then showcase the benign properties of this utility function in the context of modality selection in both classification and regression settings. To ease the flow of reading, we defer all the proofs to the appendix.

Definition 3.1. *Let c be some constant in the output space, and $\ell(\cdot, \cdot)$ be a loss function. For a set of input modalities $S \subseteq V$, the utility of S given by the utility function $f_u : 2^V \to \mathbb{R}$ is defined to be:*

$$f_u(S) \coloneqq \inf_{c \in \mathcal{Y}} \mathbb{E}[\ell(Y, c)] - \inf_{h \in \mathcal{H}} \mathbb{E}[\ell(Y, h(S))] \tag{4}$$

Namely, the utility of a set of modalities $f_u(S)$ is defined to be the reduction of the minimum expected loss in predicting Y by observing S compared to always predicting the same constant value c, i.e., c is independent of X. The latter serves as a baseline for the bare minimum performance that can be trivially achieved. This definition intuitively corresponds to the well-known observation that multimodal input can reduce prediction loss in practice. Definition 3.1 is easily interpretable, and generalizable under different loss functions and learning settings. We have the implicit assumption that the infimum is achievable in the function class.

3.1 Classification

We mainly focus on binary classifications for ease of exposition, but our proofs, algorithms, and results directly extend to multi-class classification.[2] In this setting, a subset of input modalities $S \subseteq V$ and output $Y \in \{0, 1\}$ are observed, and the predictor aims to minimize the cross-entropy loss between target Y and prediction $\hat{Y} \in [0, 1]$. Furthermore, we identify an approximate conditional independence assumption on the input multimodal data.

Assumption 3.1 (ϵ-Approximate Conditional Independence)**.** *There exists a positive constant $\epsilon \geq 0$ such that, $\forall S, S' \subseteq V, S \cap S' = \emptyset$, we have $I(S; S' \mid Y) \leq \epsilon$.*

The strength of this conditional independence relationship is parameterized by ϵ, which is the upper bound of the conditional mutual information between disjoint modalities given the output. When $\epsilon = 0$, this assumption reduces to strict conditional independence. Although strict conditional independence is a commonly used assumption on multimodal data in prior work (White et al., 2012; Wu and Goodman, 2018; Sun et al., 2020), it is difficult to be satisfied in practice. On the other hand, our new assumption considers the heterogeneity nature of multimodal data, and is able to generically model diverse real-world scenarios.

Monotonicity. Definition 3.1 manifests as the Shannon mutual information ($I(S; Y)$) between the output Y and multimodal input S in this setting (Proposition 3.1). This result is well-proven by Grünwald and Dawid (2004); Farnia and Tse (2016). We further show that $I(S; Y)$ is monotonically non-decreasing on the set of input modalities (Proposition 3.2).

Proposition 3.1. *Given $Y \in \{0, 1\}$ and $\ell(Y, \hat{Y}) \coloneqq \mathbb{1}(Y = 1) \log \hat{Y} + \mathbb{1}(Y = 0) \log(1 - \hat{Y})$, $f_u(S) = I(S; Y)$.*

Proposition 3.2. *$\forall M \subseteq N \subseteq V$, $I(N; Y) - I(M; Y) = I(N \setminus M; Y \mid M) \geq 0$.*

Definition 3.1 intrinsically characterizes the benefit of multimodal learning. Namely, Proposition 3.1 and Proposition 3.2 show that learning from more input modalities results in equivalent or better prediction performance from an information-theoretic perspective. Under Definition 3.1, the extra benefit of using more modalities can also be quantitatively described in closed-form, e.g., $I(N \setminus M; Y \mid M)$.

Comparison to previous results. Amini et al. (2009); Huang et al. (2021) have discovered similar conclusions that more modalities will not lead to worse optimal population error in the context of multiview and multimodal learning, respectively. They obtained this observation through the analysis of excess risks of learning from multiple and single modalities, and show that the excess risk of learning from multiple modalities cannot be larger than that of a single modality. Instead, our work adopts an information-theoretic characterization, which leads to an easy-to-interpret measure of the benefits of additional modalities. In practice, estimating these measures is relatively straightforward using well-developed entropy estimators. But excess risks are hard to estimate in practice since they depend on the Bayes optimal errors, which limits their uses in many applications.

Submodularity. The utility function, which manifests as the Shannon mutual information, is approximately submodular under Assumption 3.1. Previously, Krause and Guestrin

2. We have only used the binary case to derive the conditional entropy in Proposition A.1, and to further showcase a lower bound with zero-one loss in Corollary 4.1

(2012, Corollary 4) has shown mutual information to be submodular under strict conditional independence. There are also other generalizations of submodularity such as weak submodularity (Khanna et al., 2017) or adaptive submodularity (Golovin and Krause, 2011). We provide a relaxation of the strict conditional independence in Assumption 3.1, and show the approximate submodularity under this assumption.

Proposition 3.3. *Under Assumption 3.1, $I(S;Y)$ is ϵ-approximately submodular, i.e., $\forall A \subseteq B \subseteq V$, $e \in V \setminus B$, $I(A \cup \{e\};Y) - I(A;Y) + \epsilon \geq I(B \cup \{e\};Y) - I(B;Y)$.*

If the conditional mutual information between the disjoint input modalities given output is parameterized by a threshold $\epsilon > 0$, then the mutual information between input modalities and the output emits an approximate submodularity that is also parameterized by ϵ. When conditional mutual information is zero, input modalities will be strictly conditionally independent, and the mutual information will be strictly submodular.

3.2 Regression

In this setting, a subset of input modalities $S \subseteq V$ and output $Y \in \mathbb{R}$ are observed, and the predictor aims to minimize the quadratic loss between target Y and prediction $\hat{Y} \in \mathbb{R}$. In addition, let Φ be a set of feature transformations such that $\Phi(V) = \{\Phi_i(X_i)\}_{i=1}^{|V|}$. In this setting, we assume that the Bayes optimal predictor of Y (which is the conditional expectation under squared loss) is linear in the feature representations of all modalities $\Phi(V)$ (Assumption 3.2), and $\Phi(V)$ follows a multivariate Gaussian distribution (Assumption 3.3).

The (kernelized) linear assumption is commonly used in the prior work on kernel methods (Kanagawa et al., 2018; Domingos, 2020). Despite the linear form, the kernelized representation in Assumption 3.2 is capable of encoding non-linear relationships. The Gaussian assumption can be satisfied by using representation learning methods that learn features following a multivariate Gaussian, such as the variational autoencoder (VAE) in Kingma and Welling (2013), where the latent representation is trained to be close to a standard Gaussian. VAE for representation learning is also widely studied (Higgins et al., 2017; Zhu et al., 2020; Zhang et al., 2022).

Assumption 3.2. *The conditional expectation of Y given $\Phi(V)$ is linear, i.e., $\mathbb{E}[Y \mid \Phi(V)] = \sum_{X_i \in V} \beta_i \Phi_i(X_i) + \alpha$.*

Assumption 3.3. *The marginal distribution of $\Phi(V)$ admits a multivariate Gaussian distribution with mean μ and covariance matrix Σ, i.e., $\Phi(V) \sim \mathcal{N}(\mu, \Sigma)$.*

Monotonicity. We show that Definition 3.1 manifests as the variance of the conditional expectation of the output Y given a set of input modalities S (Proposition 3.4). It is also monotonically non-decreasing (Proposition 3.5).

Proposition 3.4. *Given $S \subseteq V$ and $\ell(Y, \hat{Y}) \coloneqq (Y - \hat{Y})^2$, we have $f_u(S) = \mathrm{Var}(\mathbb{E}[Y \mid S])$.*

Proposition 3.5. *$\forall M \subseteq N \subseteq V$, $\mathrm{Var}(\mathbb{E}[Y \mid N]) - \mathrm{Var}(\mathbb{E}[Y \mid M]) = \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid N] \mid M)] \geq 0$.*

Submodularity. The modality selection problem asks to select a size-q subset of X_is from V. However, selecting from the original modalities V in the input space is difficult because having $\text{Var}(\mathbb{E}[Y \mid \cdot])$ to be submodular on V imposes strong independence among the X_is (e.g., $\forall X_i, X_j \in V, i \neq j,\ X_i \perp X_j$). Our key observation here is that instead of performing selection in the input space, we can perform modality selection in a transformed feature space, i.e., on the feature representations of the modalities V', where V' is a linear transformation of $\Phi(V)$ obtained from Singular Value Decomposition (SVD):

$$\begin{aligned} \max_{S} \quad & \text{Var}(\mathbb{E}[Y \mid S]) \\ \text{s.t.} \quad & |S| \leq q,\ S \subseteq V' \\ \text{where} \quad & V' = Q\Phi(V),\ \Sigma = Q^\top \Lambda Q \end{aligned} \tag{5}$$

Specifically, since Σ in Assumption 3.3 is symmetric positive semi-definite, there exists a unique SVD for Σ such that $\Sigma = Q^\top \Lambda Q$, where Q is an $|V| \times |V|$ orthogonal matrix. We can then linearly transform $\Phi(V)$ to V', in which $V' = Q\Phi(V)$. As a result, it readily follows that $V' \sim \mathcal{N}(Q\mu, \Lambda)$. Furthermore, $\Phi(V)$ can also be reconstructed via $\Phi(V) = Q^\top V'$. Proposition 3.4 and Proposition 3.5 still hold on V' as they do not rely on Assumption 3.2 or Assumption 3.3. The benefit of operating under V' rather than $\Phi(V)$ is that we can show that $\text{Var}(\mathbb{E}[Y \mid S])$ is submodular on $S \subseteq V'$, and use submodular optimization on V' for modality selection.

To show that $\text{Var}(\mathbb{E}[Y \mid \cdot)$ is submodular, we first show that $\mathbb{E}[Y \mid S]$ is linear for all subsets of V' (Proposition 3.6). Then we prove the diminishing gain property from Definition 2.1 for $\text{Var}(\mathbb{E}[Y \mid \cdot)$ (Proposition 3.7).

Proposition 3.6. *Under Assumption 3.2 and Assumption 3.3, $\mathbb{E}[Y \mid S]$ is linear in S for any $S \subseteq V'$.*

Proposition 3.7. *Under Assumption 3.2 and Assumption 3.3, $\text{Var}(\mathbb{E}[Y \mid S])$ is a submodular function of S for any $S \subseteq V'$.*

4. Modality Selection via Submodular Optimization

In this section, we present our theoretical results on modality selection via submodular function maximization in both the classification and regression settings.

4.1 Classification

With Proposition 3.3, we can formulate the problem of modality selection as a submodular function maximization problem under cardinality constraint, i.e., $\max_{S \subseteq V} I(S; Y)$ subject to $|S| \leq q$ where q is often smaller than $|V|$. The classic approximation guarantee in Theorem 2.1 is applicable to $I(\cdot; Y)$ only if it is strictly submodular. However, the approximation guarantee will be different in our case because the strength of submodularity of $I(\cdot; Y)$ is parameterized by the upper bound of conditional mutual information (ϵ) under the approximate conditional independence assumption (Assumption 3.1).

Theorem 4.1. *Under Assumption 3.1, let $q \in \mathbb{Z}^+$, and S_p be the selected subset from Algorithm 1 at iteration p, we have:*

$$I(S_p; Y) \geq (1 - e^{-\frac{p}{q}}) \max_{S:|S|\leq q} I(S; Y) - q\epsilon. \tag{6}$$

Since $I(\cdot; Y)$ is monotonically non-decreasing (Proposition 3.2), the value of the selected subset from Algorithm 1 after q iterations will be at least $1 - \frac{1}{e}$ of the optimal value minus $q\epsilon$ across all q-size subsets. The $q\epsilon$ term characterizes the fact that: when $I(\cdot; Y)$ is approximately submodular, selecting a larger subset via Algorithm 1 will have a larger approximation error that is caused by the conditional mutual information between modalities. When $\epsilon = 0$, Theorem 4.1 reduces to Theorem 2.1.

Based on Theorem 4.1, we can further obtain a bound on the minimum of the expected cross-entropy loss and expected zero-one loss achieved by the greedily-obtained set. Let us denote the optimal set $S^* = \arg\max_{S:|S|\leq q} I(S; Y)$, cross-entropy loss $\ell_{ce}(Y, \hat{Y}) \coloneqq \mathbb{1}(Y = 1) \log \hat{Y} + \mathbb{1}(Y = 0) \log(1 - \hat{Y})$, and zero-one loss $\ell_{01}(Y, \hat{Y}) \coloneqq \mathbb{1}(Y \neq \hat{Y})$.

Corollary 4.1. *Assume conditions in Theorem 4.1 hold, there exists optimal predictor $h^*(S_p) = \Pr(Y \mid S_p)$ such that:*

$$\mathbb{E}[\ell_{01}(Y, h^*(S_p))] \leq \mathbb{E}[\ell_{ce}(Y, h^*(S_p))] \leq H(Y) - (1 - e^{-\frac{p}{q}}) I(S^*; Y) + q\epsilon \tag{7}$$

Corollary 4.1 shows that the minimum of both losses achieved by $\Pr(Y \mid S_p)$ are no more than the uncertainty of the target output minus the lower bound of our greedily-obtained value from Theorem 4.1. We can also upper bound difference between the optimal cross-entropy loss achieved by the greedily-obtained set and the optimal set.

Corollary 4.2. *Assume conditions in Theorem 4.1 hold. There exists optimal predictors $h_1^*(S_p) = \Pr(Y \mid S_p)$, $h_2^*(S^*) = \Pr(Y \mid S^*)$ such that:*

$$\mathbb{E}[\ell_{ce}(Y, h_1^*(S_p))] - \mathbb{E}[\ell_{ce}(Y, h_2^*(S^*))] \leq e^{-\frac{p}{q}} I(S^*; Y) + q\epsilon \tag{8}$$

This result gives a guarantee on the maximum loss difference from the greedily-obtained set versus the optimal set using the optimal predictors. Both bounds from Corollary 4.1 and Corollary 4.2 are parameterized by the running iteration p and set size constraint q of Algorithm 1, as well as the approximation error induced by ϵ. As the algorithm attempts to select a larger set of modalities, both bounds become tighter.

4.2 Regression

Recall that in the regression setting, we turned to solve modality selection on a feature transformation of the original input modalities. We showed that the utility function in this setting (i.e., $\mathrm{Var}(\mathbb{E}[Y \mid S])$) is submodular for any subset $S \subseteq V'$, where V' is the transformed feature representation $Q\Phi(V)$. Selecting modalities on V' allows us to utilize the benign properties of submodularity that are otherwise difficult to satisfy on V.

Theorem 4.2. *Under Assumption 3.2 and Assumption 3.3, let $q \in \mathbb{Z}^+$, and $S_p \subseteq V'$ be the solution from Algorithm 1 at iteration p, we have:*

$$\mathrm{Var}(\mathbb{E}[Y \mid S_p]) \geq (1 - e^{-\frac{p}{q}}) \max_{S:S\subseteq V',\ |S|\leq q} \mathrm{Var}(\mathbb{E}[Y \mid S]) \tag{9}$$

Theorem 4.2 is directly extended from Theorem 2.1 because $\mathrm{Var}(\mathbb{E}[Y \mid \cdot])$ is submodular on V'. We also obtain the following upper bounds on $\inf_f \mathbb{E}[(Y - f(S_p))^2]$, similar to Corollary 4.1 and Corollary 4.2 in the classification setting. Let us denote optimal set $S^* = \max_{S:S\subseteq V',\ |S|\leq q} \mathrm{Var}(\mathbb{E}[Y \mid S])$.

Corollary 4.3. *Assume conditions in Theorem 4.2 hold. There exists an optimal predictor* $h^*(S_p) = \mathbb{E}[Y \mid S_p]$ *such that:*

$$\mathbb{E}[(Y - h^*(S_p))^2] \leq \mathrm{Var}(Y) - (1 - e^{-\frac{p}{q}})\mathrm{Var}(\mathbb{E}[Y \mid S^*]) \tag{10}$$

Corollary 4.4. *Assume conditions in Theorem 4.2 hold. There exists optimal predictors* $h_1^*(S_p) = \mathbb{E}[Y \mid S_p]$, $h_2^*(S^*) = \mathbb{E}[Y \mid S^*]$ *such that:*

$$\mathbb{E}[(Y - h_1^*(S_p))^2] - \mathbb{E}[(Y - h_2^*(S^*))^2] \leq e^{-\frac{p}{q}}\mathrm{Var}(\mathbb{E}[Y \mid S^*])$$

Remark. We show that in both classification and regression settings, we can leverage submodular function maximization to tackle the problem of modality selection in multimodal learning. By the benign property of approximate submodularity, we can obtain a selected subset of modalities from a simple greedy selection algorithm (Algorithm 1) in polynomial time, whose learning performance has an optimality guarantee to the true optimal solution. Under our theoretical formulation, we can also directly extend the results of other submodular optimization problems under different constraints and objectives into the context of modality selection (Krause and Golovin, 2014).

5. Modality Selection via Modality Importance Ranking

We now present our theoretical results of connecting feature importance scores to modality selection. In this section, we formulate modality selection as a ranking problem, where each input modality is ranked based on its "modality importance" in descending order. Then, we can select the top-k modalities (where k is the cardinality constraint) based on the ranked importance scores. In particular, we will adopt the Shapley value and MCI to measure the importance of a modality by using Definition 3.1 as the evaluation function (§2.2).

5.1 Classification

We first show that mutual information is sub-additive and super-additive under the approximate independence assumption of the input modalities. By leveraging sub- and super-additivity, we can compute the Shapley value and MCI of a modality exactly and efficiently.

Proposition 5.1. *Under Assumption 3.1,* $I(S;Y)$ *is* ϵ*-approximately sub-additive for any* $S \subseteq V$*, i.e.,* $I(S \cup S';Y) \leq I(S;Y) + I(S';Y) + \epsilon$.

Next, we show the approximate super-additivity of the mutual information under an approximate marginal independence assumption. For ease of presentation, we parameterize approximate marginal and conditional independence assumptions by the same constant ϵ.

Assumption 5.1 (ϵ-Approximate Marginal Independence)**.** *There exists a positive constant* $\epsilon > 0$ *such that,* $\forall S, S' \subseteq V, S \cap S' = \emptyset$*, we have* $I(S;S') \leq \epsilon$.

Proposition 5.2. *If Assumption 5.1 holds, $I(S;Y)$ is ϵ-approximately super-additive for any $S \subseteq V$, i.e., $\forall S, S' \subseteq V, S \cap S' = \emptyset$, $I(S \cup S';Y) \geq I(S;Y) + I(S';Y) - \epsilon$.*

In the classic definition of Shapley value, the complexity of computing the exact Shapley value of a player for general evaluation functions is exponential. However, when the evaluation function v of the Shapley value exhibits strict additivity, the exact Shapley value can be computed in polynomial time.

We provide some intuition about why additivity can make computation efficient. Specifically, we can leverage the sub-additivity to provide an upper bound of the Shapley value ϕ_{v,X_i} via a summation of $v(X_i)$ for all possible subsets. Analogously, the super-additivity should provide a lower bound of ϕ_{v,X_i} again expressed by $v(X_i)$. Putting two bounds together gives us an efficient approximation of ϕ_{v,X_i}. This solution of the exact Shapley value for a modality is parameterized by ϵ. In this setting, mutual information is the evaluation function v for the Shapley value.

Proposition 5.3. *If $I(S;Y)$ is both ϵ-approximately sub- and super-additive for any $S \subseteq V$, we have $I(X_i;Y) - \epsilon \leq \phi_{I,X_i} \leq I(X_i;Y) + \epsilon$ for any $X_i \in V$.*

When strict independence applies, i.e., $\epsilon = 0$, the Shapley value of a modality is exactly its prediction utility. In addition, the efficiency property of the Shapley value states that the sum of the Shapley values of all agents equals the value of the grand coalition. Thus, we must have $I(V;Y) = \sum_{X_i \in V} \phi_{I,X_i}$.

Next, we examine MCI (§2.2). By its definition, solving MCI of a feature requires exponential time, i.e., $\mathcal{O}(2^{|F|})$ where $|F|$ is the total number of features. But if the evaluation function of MCI is submodular, we can efficiently compute the exact MCI in polynomial time.

Proposition 5.4. *Under Assumption 3.1, we have $I(X_i;Y) \leq \phi^{mci}_{I,X_i} \leq I(X_i;Y) + \epsilon$ for any $X_i \in V$.*

If $\epsilon = 0$, the mutual information will be strictly submodular, and the MCI of a modality is exactly its prediction utility, i.e., $\phi^{mci}_{I,X_i} = I(X_i;Y)$. In addition, if Proposition 5.1 holds with $\epsilon = 0$, $I(\cdot;Y)$ will be sub-additive, and $I(V;Y) \leq \sum_{X_i \in V} I(X_i;Y) = \sum_{X_i \in V} \phi^{mci}_{I,X_i}$. If Proposition 5.2 also holds with $\epsilon = 0$, then $I(\cdot;Y)$ will be additive. Then we can obtain an efficiency property for MCI, i.e., $I(V;Y) = \sum_{X_i \in V} \phi^{mci}_{I,X_i}$.

5.2 Regression

We continue our analysis in the regression setting on the transformed version of the original modalities V (denoted as V'). Recall that Definition 3.1 manifests as the variance of conditional expectation in this setting. Next, we show that the variance of conditional expectation is additive. Hence, by using it as the evaluation function, we can compute the exact importance of a modality via Shapley value and MCI in polynomial time.

Proposition 5.5. *Under Assumption 3.2 and Assumption 3.3,*

- $\mathrm{Var}(\mathbb{E}[Y \mid S])$ *is additive for any $S \subseteq V'$.*
- $\phi_{\mathrm{Var}(\mathbb{E}[Y|\cdot]),X_i} = \phi^{mci}_{\mathrm{Var}(\mathbb{E}[Y|\cdot]),X_i} = \mathrm{Var}(\mathbb{E}[Y \mid X_i])$ *for any $X_i \in V'$.*

Since $\text{Var}(\mathbb{E}[Y \mid \cdot])$ is additive on V', the marginal contribution of $X_i \in V'$ from each unique subset $S \subseteq V'$ will be equal to the same quantity $\text{Var}(\mathbb{E}[Y \mid X_i])$ that is independent of S. Then by Definition 2.2, the Shapley value of a modality after linear transformation $\Psi(\Phi(V)) \coloneqq Q\Phi(V)$ will be equal to the utility of that modality after transformation (Proposition 5.5). Again, by the efficiency property of Shapley value, we have $\text{Var}(\mathbb{E}[Y \mid V']) = \sum_{X_i \in V'} \phi_{\text{Var}(\mathbb{E}[Y|\cdot]),X_i}$. As $\text{Var}(\mathbb{E}[Y \mid \cdot])$ is submodular, the exact MCI of a modality will also equal its prediction utility. We can also obtain an efficiency property of the MCI, i.e., $\text{Var}(\mathbb{E}[Y \mid V']) = \sum_{X_i \in V'} \phi^{mci}_{\text{Var}(\mathbb{E}[Y|\cdot]),X_i}$.

Remark. To provide guidance to practitioners on choosing the best modality selection algorithm that fits their needs, we compare our proposed algorithms mainly by their selected modality set, and time complexity. A key advantage of the greedy algorithm is that it tends to select more complimentary modality sets with performance guarantee—while MCI only considers the marginal utility of each modality, the greedy algorithm further accounts for utility of the entire selected set. The claim is empirically validated in Section 6.1.2, in which we also provided an illustrative example (Fig. 2). Conceptually, MCI is designed for proper credit assignments to individual modalities, while the greedy algorithm focuses more on the utility of a modality set. For instance, when there exist two identical modalities with very high utilities, MCI ranking will likely select both whereas the greedy algorithm will not.

On the other hand, one advantage of MCI ranking is its time complexity. The time complexity of ranking is $\mathcal{O}(|V| \log |V|)$, while greedy submodular maximization is $\mathcal{O}(q|V|)$, where $|V|$ is the total number of input modalities, and q is the number of selected modalities. Nevertheless, there are practices to speed up the greedy maximization algorithm with slight relaxation on its theoretical guarantee. For example, Mirzasoleiman et al. (2013) proposes Distributed Greedy, which distributes the set selection process on l machines, resulting in a complexity of $\mathcal{O}(q|V|/l)$; Mirzasoleiman et al. (2015) proposes Stochastic-Greedy Algorithm, which restricts the search space for the next elements in each iteration and achieves a $(1 - 1/e - \epsilon)$ guarantee in linear time $\mathcal{O}(|V|)$.

6. Experiments

In this section, we present our empirical evaluation of modality selection via greedy submodular maximization (Algorithm 1), and modality importance ranking via MCI.

Algorithm implementation. For Algorithm 1, in each iteration i we execute the following: (1) for each candidate modality X_j, we train two models on S_i and $S_i \cup \{X_j\}$ respectively until training losses converge, and take the difference between two losses; (2) record test loss and accuracy/R-squared from the model trained on $S_i \cup \{X_i\}$ before the model over-fits; (3) add the selected modality to S_i to construct S_{i+1} and go to next iteration. We use model parameters before over-fitting for prediction and parameters after over-fitting for utility estimation. We make such design choices because, for prediction, we want to generalize well for better test performance, while for utility estimation, we want to record the fully converged loss. For importance ranking, we compute the MCI of each input modality and then select the top-ranked modalities with the largest MCI values.

6.1 Classification

We evaluate our results on one semi-synthetic data set and two real-world data sets.

Patch-MNIST. This is a semi-synthetic data set built upon MNIST (LeCun and Cortes, 1998). Specifically, we divide each image in the original MNIST into non-overlapping square patches. Each patch location represents a single modality. We construct and experiment on two Patch-MNIST variants, where one variant has 49 patches and each patch is of size 4×4 square pixel, and another has 9 patches and each patch has a side length of 9 or 10 pixels. Patch-MNIST has ten output classes, 50,000 training images, and 10,000 testing images.

PEMS-SF. This is a real-world time-series data set from the UCI machine learning repository (Dua and Graff, 2017). It represents the traffic occupancy rate of different freeways in the San Francisco Bay Area. The task is to classify the day of the week. Data is obtained from 963 sensors placed across the bay area. Each sensor represents modality, and has a time series with 144 time steps. We down-sample 144 time steps to 36 via taking the regional means of size-4 windows. Running Algorithm 1 requires $\mathcal{O}(q|V|)$ with $|V| = 963$, and each step requires training a new model. To mitigate the extensive run-time, we experiment on 45 out of 963 sensors by filtering sensors in line for the same freeway. There are a total of 440 instances (days), with the train-val-test split being 200, 67, and 173 samples.

CMU-MOSI. This is a popular real-world benchmark in affective computing and multimodal learning (Zadeh et al., 2016). The task is 3-class sentiment classification (positive, neutral, negative) from 20 visual and 5 acoustic modalities with temporal features. Specifically, CMU-MOSI collects time-series facial action units and phonetic units from short video clips (10-second clip sampled at 5Hz rate). Each unit is a modality, and consists of a 50-dimensional feature vector. Training and testing sample sizes are 1284 and 686 respectively.

We validate the independence conditions on all data sets by comparing the mean conditional Mutual Information (MI) and the mean marginal MI of disjoint modalities (Gao et al., 2017). As shown in Table 1, the conditional MI is smaller than the marginal MI for MNIST and PEMS-SF. Both conditional and marginal MI are small for CMU-MOSI. This implies that modalities should be approximately conditionally independent in these data sets.

Table 1: Mean Marginal/Conditional Mutual Information

data set	Mean Marg. MI	Mean Cond. MI
Patch-MNIST	2.187	0.078
PEMS-SF	0.626	0.223
CMU-MOSI	0.064	0.069

6.1.1 Implementation

Utility estimation. From Proposition 3.1, utility $f_u(S)$ equals $I(S;Y)$, and $I(S;Y) = H(Y) - H(Y \mid S)$. Based on the variational formulation of the conditional entropy as the minimum cross-entropy, we approximate $H(Y \mid S)$ by using the converged training loss on S to predict Y (Farnia and Tse, 2016). Accordingly, to estimate the marginal gain

$I(X_j; Y \mid S_i)$ from Algorithm 1 over high dimensional data, we compute the difference $H(Y \mid S) - H(Y \mid S \cup \{X_j\})$ (McAllester and Stratos, 2020). To compute the MCI of each modality X_j, we just need to compute $I(X_j; Y)$, according to Proposition 5.4.

Modeling. We now describe models for prediction and utility estimation. For Patch-MNIST, we use a convolutional neural network with one convolutional layer, one max pooling layer, and two fully-connected layers with ReLU for both estimation and prediction. The network is trained with Adam optimizer on a learning rate of $1e-3$. For PEMS-SF, we use a 3-layer neural network with ReLU activation and batch normalization for estimation. This is trained with Adam optimizer on a learning rate of $5e-4$. For prediction, we use a recent time-series classification pipeline (Dempster et al., 2020) for time-series data processing, followed by a linear Ridge Classifier (Löning et al., 2019). For CMU-MOSI, we experiment with two prediction model types: a linear classifier with Rocket Transformation for time-series (same as the one for PEMS-SF); and a plain 3-layer fully-connected neural network with ReLU activation. On each data set, the number of training epochs is the same for all evaluated approaches across different modality subset sizes.

Experimental procedures. For PEMS-SF and CMU-MOSI, we record and show the training loss before over-fitting instead of the test loss. This is because PEMS-SF and CMU-MOSI have a much smaller sample size than Patch-MNIST with potentially noisier features, the model likely will not generalize stably. We first examine Theorem 4.1 and MCI ranking on a larger sample set which better represents the population and is not influenced by the generalization gap. Then we analyze the test accuracy to account for the generalization.

For Patch-MNIST with 49 modalities, PEMS-SF and CMU-MOSI, we evaluate Algorithm 1 and MCI ranking against a randomized baseline at each set size. The randomized baseline randomly selects a modality iteratively. For Patch-MNIST with 9 modalities, we further include optimal and average baselines. At each set size q, the optimal baseline is the optimal value from all possible subsets of size q, and the average baseline is the average. We only implement the optimal baseline for the 9 modalities case because evaluating all possible subsets for a larger set is expensive.

Training cost. At each iteration of Algorithm 1, the marginal utility gain for each candidate modality is evaluated. Since we estimate the conditional mutual information by training a neural network, and we need to evaluate each modality subset at different set sizes, each iteration involves model training. These experiments can be costly for large data sets and models. The training cost at each iteration of Algorithm 1 depends on different utility variants, or mutual information estimation methods in this setting.

6.1.2 Results and Empirical Analysis

Patch-MNIST. Fig. 1 shows the Patch-MNIST experiment results. In this figure, "Modality subset size" refers to the size of the selected modality set. "Utility" refers to the utility of the selected set. The "Test CE Loss" and "Test Accuracy" refer to the cross-entropy loss and prediction accuracy on test data from the model that is trained on the selected set.

Learning utility An immediate observation is the high correlation among the utility, test cross-entropy loss and accuracy in both rows. The trend of test accuracy seems identical to the utility, although they mildly differ when the set size exceeds 30. In addition, the utility and test loss are negatively correlated, matching to Definition 3.1. The utility has a larger

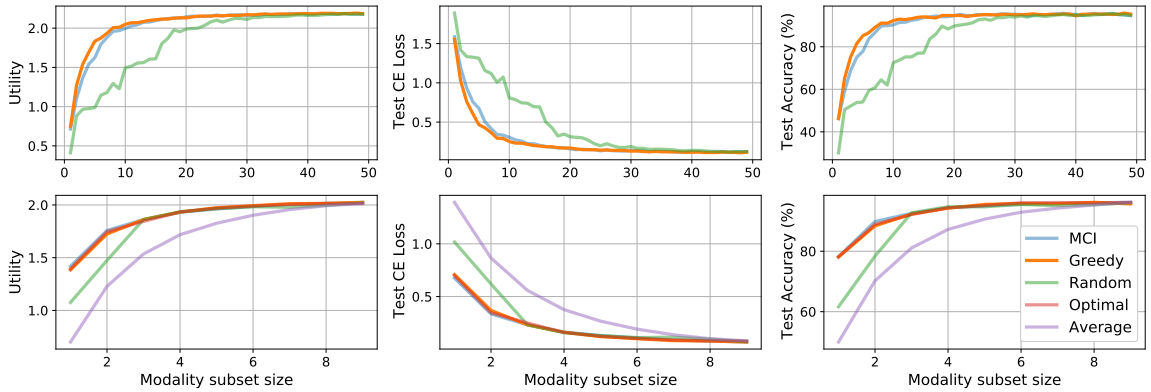

Figure 1: Results on Patch-MNIST with 49 (first row) and with 9 modalities (second row).

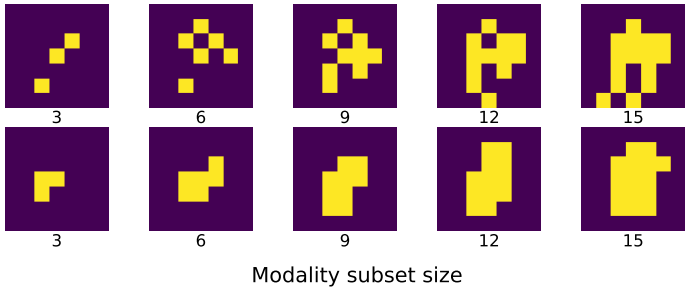

Figure 2: Modalities selected by Algorithm 1 (first row) and MCI ranking (second row) on Patch-MNIST.

upper bound than test loss, potentially because the utility is estimated by converging training loss, which is often reduced in greater magnitude than test loss. The utility has a trend of non-decreasing and diminishing gain, which matches the monotonicity and (approximate) submodularity shown in this setting. Adding more modalities is unnecessary if the subset is already large: in the 49-modalities case, accuracy barely improves after 20 modalities are selected; but in the 9-modalities case, this pattern is less obvious.

Greedy maximization Algorithm 1 beats random selection in both cases. In Fig. 1 (second row), it beats the average by selecting the modality with maximum utility from the start, and overlaps its trajectory with the optimal. In Fig. 1 (first row), Algorithm 1 achieves near-maximum utility with only 7 modalities. These results validate the approximate guarantee from Theorem 4.1. In fact, the guarantee of utility is empirically much better than theoretically proven.

MCI ranking In the 9-modalities case, MCI ranking is as good as greedy maximization and the optimal baseline when the full set has fewer modalities. When more modalities are available for selection (e.g., 49 modalities), Algorithm 1 select a subset that minimizes the loss slightly further than the highest ranked modalities when set size below 15.

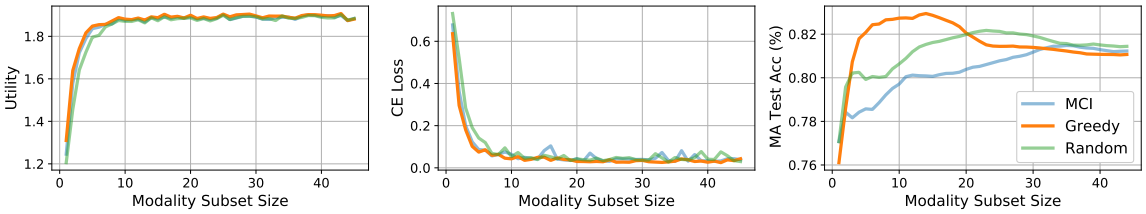

Figure 3: Experiment results for PEMS-SF.

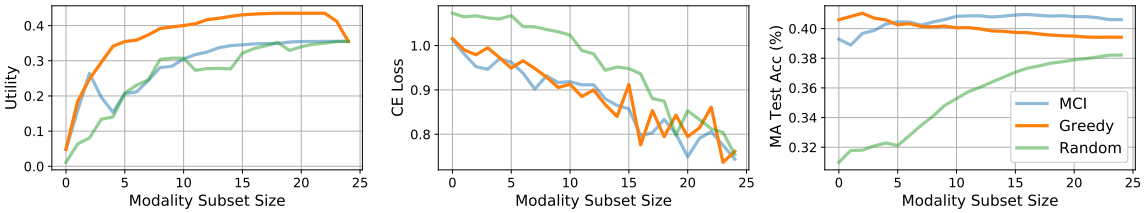

Figure 4: Experiment results for CMU-MOSI.

Modality selection path We plot the modality selection paths from Algorithm 1 and MCI ranking in Fig. 2. We can see that MCI selects the modalities that each contain the most information to output—the center regions. Whereas the modalities selected by Algorithm 1 are more diverse, covering different spatial locations of the original image, leading to an advantage in gaining more information collectively.

PEMS-SF. Fig. 3 shows our experiment results on PEMS-SF. In Fig. 3, the two leftmost plots show the utility and cross-entropy loss on the training data. The rightmost plot of Fig. 3 shows the moving average of test accuracy instead, because the model was not generalized stably under a small sample size.

Learning utility The difference in utility and loss among Algorithm 1, MCI ranking, and random baseline are small, and all of them quickly converge to the minimum possible value after selecting only a few modalities. This might be because almost every modality is sufficient to make training loss small. However, greedily selected subsets still have slightly more utility than subsets from MCI ranking and random baseline at every set size. Overall, we still observe the utility is monotone and (approximate) submodular; and Algorithm 1's achieved utility matches Theorem 4.1.

Generalization From the test accuracy plot, we can see a clear advantage from the greedily-obtained set over others when the subset size is small. Meanwhile, MCI ranking is worse than the random baseline, which could imply that MCI ranking does not have a robust performance guarantee as Algorithm 1. Other than that, the test accuracy of Algorithm 1 gradually decreases as more modalities are added. This is in line with the over-fitting artifact of greedy feature selection from Blanchet et al. (2008). However, in the regime of good generalization, greedy maximization should preserve the performance guarantee during testing.

CMU-MOSI. The results are alike for both prediction model types mentioned in Section 6.1.1 for CMU-MOSI. Thus we only use Fig. 4 to show the CMU-MOSI evaluation

results from the 3-layer fully-connected neural network. In Fig. 4, the two leftmost plots show the utility and cross-entropy loss on the training data. The rightmost plot of Fig. 4 shows the moving average of test accuracy since the model lacks the capacity to generalize well for this data set under a small sample size.

Overall, many previous observations from other data sets still hold for CMU-MOSI. For example, the utility curve is approximately submodular and monotone as the number of selected modalities increases. Modalities selected by Algorithm 1 and MCI ranking outperform randomly selected modalities by having more utility, lower training loss, and higher testing accuracy, especially when the number of modalities is still small. Meanwhile, potentially due to the simplicity of the model and noisy features, we are unable to observe an increase in testing accuracy as more modalities are included in Algorithm 1 and MCI ranking.

6.2 Regression

We evaluate our results on one semi-synthetic data set and two real-world data sets.

Synthetic. This data set contains 1000 samples with 10 modalities, each modality X_i contains 3 attributes. Following Assumption 3.3, we sample the attributes from a multivariate Gaussian with zero mean and a block-diagonal covariance matrix, i.e., the attributes within the same modality can be correlated, while the attributes in different modalities are independent. Following Assumption 3.2, the regression target is constructed by $Y = \sum_i \beta_i X_i + \alpha + \epsilon$, where the coefficients β_i and α are sampled from $[-1, 1]$ uniformly at random, and the error ϵ is sampled from a univariate Gaussian with zero mean and unit variance. Furthermore, to make the data set more realistic, we also vary the off-diagonal values to make the modalities dependent to different extents, and demonstrate how dependency affects the performance of the algorithms. Details about the data-generating process can be found in Appendix C.

Appliances. This is a real-world data set aiming at predicting the energy consumption of a household (Candanedo et al., 2017). The data set contains 10 temperature-humidity pairs from 9 rooms in a house and a nearby weather station as well as other general weather features, such as visibility, pressure and wind speed. The data is collected every 10 minutes over a 4.5-month period. We treat each temperature-humidity pair as a modality and the remaining weather features as another modality.

Communities and Crime. This is a real-world crime prediction data set from the UCI machine learning repository (Redmond and Highley, 2010). The task is to predict the number of murders per 100K population. The raw data contains 125 attributes, including demographic composition, income levels, police force, etc. After dropping the non-predictive and redundant attributes, we used 65 attributes (divided into 17 modalities) for our experiment.

6.2.1 Implementation

Utility estimation. We directly apply the definition of utility function in Definition 3.1. Specifically, for each model, we first record the converged MSE loss given a constant modality ℓ_0, then take the difference between ℓ_0 and the converged MSE loss using the modality of interest as its utility. Therefore, the utility is model-dependent.

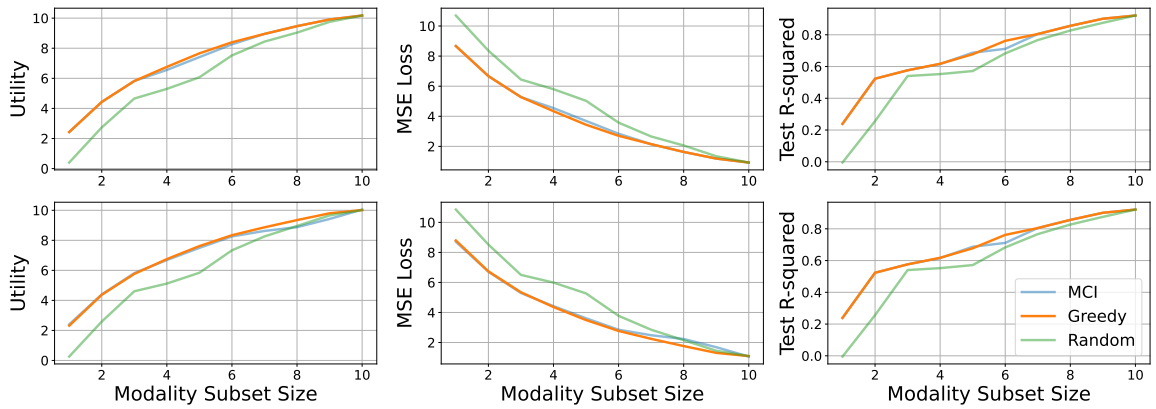

Figure 5: Experiment results on the synthetic data set with linear regressor (first row) and neural network (second row).

Modeling. For neural networks, we use the structure of multiple modality-specific feature extractors followed by a joint linear predictor. For the synthetic data set, the feature extractor is a single-layer fully connected network with 128 hidden units. For both the appliances and the crimes data set, the feature extractor is a 3-layer fully connected network with 128 hidden units. For VAE, we use two 3-layer fully connected networks with 128 hidden units as the encoder and the decoder respectively. The latent dimension is 16. The networks are trained with the Adam optimizer with a learning rate of $1e-3$.

6.2.2 Results and Empirical Analysis

To illustrate the impact of different feature transformations in Assumption 3.2, we evaluate the algorithms using linear regressors (identity feature mapping) and neural networks (both VAE encoded features and end-to-end training). To eliminate the effect of inherent variance in the data set, we report the test R-squared along with the raw MSE loss.

Synthetic. The performance of both the linear regressor and the neural network is shown in Fig. 5. The comparison under different modality dependencies is shown in Fig. 6.

Learning utility The utility has a trend of non-decrease and diminishing gain, verifying the monotonicity and submodularity. Since the ground-truth data-generating function is linear, the performance under both models is similar.

Selection algorithms The performance of Algorithm 1 and MCI ranking is alike, both outperforming the random baseline. Since we sample the attributes from a block-diagonal covariance matrix, all the modalities are independent. Under such circumstances, the algorithms have identical modality selection paths because picking the modality with the highest utility is the optimal strategy in each step. However, when there exists redundant information, i.e., the modalities are dependent, Algorithm 1 will outperform MCI ranking by selecting more complementary modalities.

Modality dependency We now study how inter-modality dependency affects performance. While we fix the block-diagonal in the covariance matrix to be between -1 and 1, we allow different amounts of covariance in the off-diagonal with a scale ranging from 0.001 to 0.7

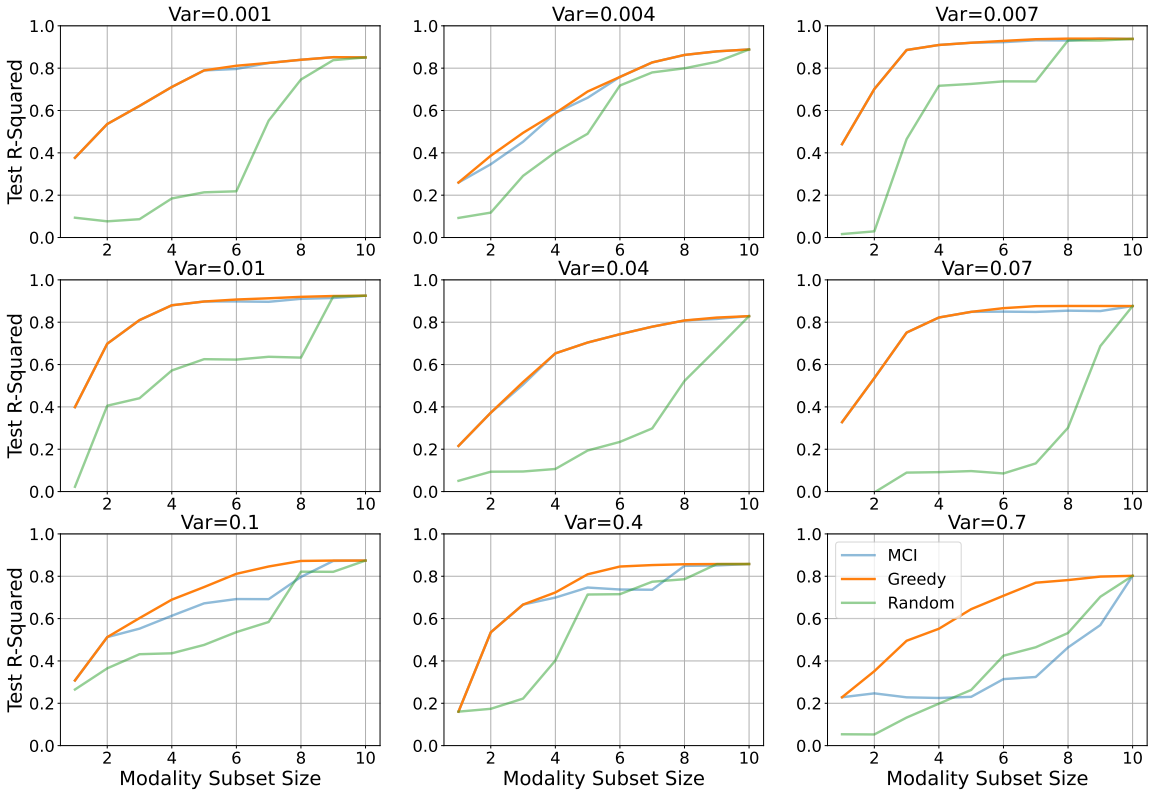

Figure 6: Experiment results for the synthetic data set with different modality dependencies using a linear regressor.

(larger value means higher inter-modality dependency). As shown in Fig. 6, when the covariance gets larger, Algorithm 1 significantly outperforms MCI ranking. In the case where covariance is between -0.7 to 0.7, MCI ranking plateaus at the beginning as it only selects modalities with high individual utility, but omits information overlap between them.

Appliances. The experiment results are shown in Fig. 7.

Leaning utility The utility for the neural network shows a clear diminishing return pattern, while it is less obvious in the case of linear regression. This indicates that if we directly assume that the conditional expectation of the target given the input modalities is linear, the utility function may not be submodular. However, if we relax the linearity assumption to be in the feature representations of the input modalities, the utility function will better satisfy submodularity and monotonicity. Also, as expected, the utility for each modality subset is significantly higher under both the neural network model and the regressor with VAE-encoded features due to their model complexity. As a result, both models achieve higher R-squared than linear regressor at each subset size.

Selection algorithms Algorithm 1 clearly outperforms MCI ranking and random baseline in all three models. For MCI ranking under the linear regressor model, the total utility plateaus until it selects the sixth modality—the sixth modality is highly complementary

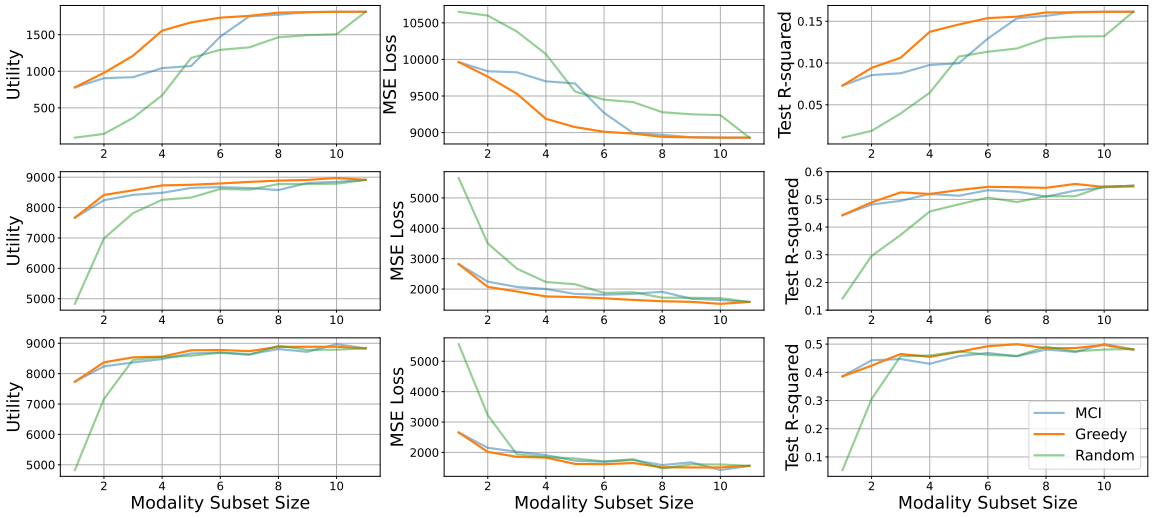

Figure 7: Results for the appliances data set with linear regressor (first row), neural network (second row), and regressor with VAE-encoded features (third row).

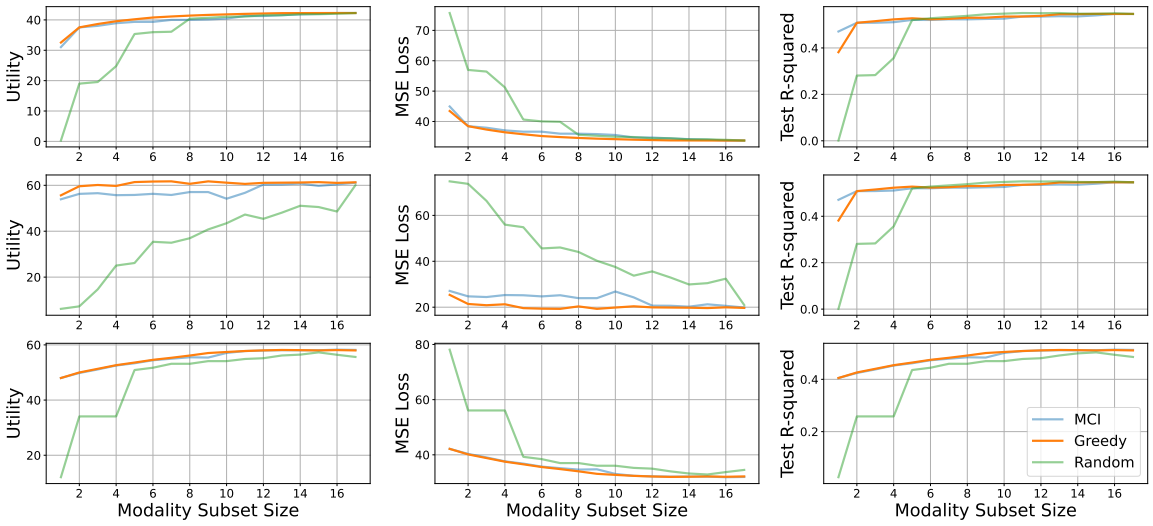

Figure 8: Results for the communities and crime data set with linear regressor (first row), neural network (second row) and regressor with VAE-encoded features (third row).

to the selected ones, but this information is not accounted by MCI ranking. Algorithm 1 accounts for this information, and selects a modality set with steadily increasing utility.

Communities and Crime. The experiment results are shown in Fig. 8.

Learning utility The utility under both models demonstrates approximate submodularity and monotonicity. However, compared to previous data sets, the trend is less clear due to the

small utility gap between using one modality and using all modalities—this often happens when one or a few modalities already have sufficient information to solve the regression task.

Selection algorithms Algorithm 1 consistently outperforms MCI ranking and the random baseline under both models. Under the neural network setting, Algorithm 1 reaches near-optimal performance using only 6 modalities, while MCI ranking uses 14. In addition, the utility of MCI ranking using the neural network shows a "plateau then increase" pattern because multiple high-utility modalities share overlapping information. This indicates that MCI ranking does not perform well at selecting complementary modalities, while Algorithm 1 does not have this shortcoming.

Feature extractors Comparing the last two rows, we can see that the regressor trained on VAE-encoded features shows a clearer diminishing return for both Algorithm 1 and MCI ranking. However, this pattern is achieved at the cost of lower performance, as shown by the lower test R-squared for the VAE-encoded features. This may be due to the fact that the features extracted by VAE are not dedicated to the regression task, as the regression targets are not used during VAE training.

7. Related Work

Multimodal Learning. Multimodal learning is a vital research area with many applications (Liu et al., 2017; Pittermann et al., 2010; Frantzidis et al., 2010). Theoretically, Huang et al. (2021) showed that learning with more modalities achieves a smaller population risk, and this marginal benefit towards prediction could be upper bounded. But the existing measure of marginal benefit (Huang et al., 2021) is hard to understand or be easily estimated, and it does not provide further insight on the emerging modality selection problem.

Submodular Optimization. Under the benign property of submodularity, many subset selection problems, which are otherwise intractable, now admit efficient approximate solutions (Fujishige, 2005; Iwata, 2008; Krause and Golovin, 2014). The first study of greedy algorithms over submodular set function dates back to Nemhauser et al. (1978). Since then, submodular optimization has been widely applied to diverse domains such as machine learning (Wei et al., 2015) , distributed computing, and social network analysis (Zhuang et al., 2013). A typical type of problem is submodular maximization, which can be subject to a variety of constraints such as cardinality, matroid, or knapsack constraints (Lee et al. (2010); Iyer and Bilmes (2013)). In our case, we extended results from Nemhauser et al. (1978) to the case of approximate submodularity of mutual information for multimodal learning.

Feature Selection. Following Chandrashekar and Sahin (2014), we categorize feature selection methods into filter method, wrapper method and embedded method.

Filter methods rank features by certain criteria and select by ordering. The ranking can be based on: correlation criteria (Weston et al., 2003) using Pearson correlation coefficient; information theoretic criteria using mutual information between the feature and target (Dumais et al., 1998); importance criteria using permutation feature importance (Breiman, 2001); game theoretic criteria using Shapley value (Shapley, 1952) or computationally tractable variants (Lundberg and Lee, 2017). However, criteria such as Pearson correlation and variance in filter methods are specifically designed for univariate features—they are not well-defined for a subset of features (i.e., a modality), which directly hinders their applications in modality selection. Note that formulation in Dumais et al. (1998) is the same

as our MCI ranking in the classification setting, except that Dumais et al. (1998) provides no theoretical guarantee on the quality of the selected features. In comparison, we first prove the submodularity of our utility function under a mild assumption of the data generative process, then provide a provable guarantee on the quality of the selected modalities.

Wrapper methods evaluate feature subsets by prediction performance. Two of the most representative algorithms are sequential feature selection (SFS) and sequential backward selection (SBS) (Pudil et al., 1994). SFS starts with an empty set, then adds one feature at a time which achieves the best predictor performance. SBS, also known as sequential backward elimination, starts with the full set and removes one feature at a time whose removal gives the lowest decrease in predictor performance. However, SBS is unfit for our context, because modality selection aims to select the most useful subset under computational resource limit (e.g., evaluating as few modalities as possible). Mutual information based feature selection (MIFS) (Battiti, 1994) uses an objective to maximize the MI between features and class output, while minimizing the MI between the selected feature and the subset of chosen features, i.e., $I(X_M;Y) - I(X_M;X_N)$. In their case, X_M is the selected feature and X_N corresponds to the subset of chosen features. In comparison, we maximize the conditional MI, i.e., $I(X_M;Y \mid X_N) = I(X_M \cup X_N;Y) - I(X_N;Y)$.

Embedded methods incorporate feature selection as part of training. LASSO (Tibshirani, 1996) and Ridge (Hoerl and Kennard, 2000) regression use regularization to enforce the model to only attend to important features. Similarly, weight-based methods (Mundra and Rajapakse, 2009) determine the feature importance by classifier weights, where higher weights indicate higher importance. Optimal brain damage (OBD) (LeCun et al., 1989) uses the second-order derivative to determine the connection weights, then prunes the unimportant features. A key distinction between this line of work and our modality selection setting is that embedded methods often do not provide an option to specify a cardinality constraint, i.e., the number of features allowed to get selected, as their selection process is performed implicitly during optimization. In addition, similar to SBS, embedded methods also require training on the whole feature set, which violates the purpose of modality selection.

One key novelty of our work beyond existing standard feature selection methods lies in our theoretical analysis based on submodularity. We identify suitable independence assumptions for the modality selection problem that are essential for the applicability of efficient algorithmic design from the submodular optimization literature. On the contrary, we do not find any feature selection approaches that directly employ submodularity on loss functions for both classification and regression problems under realistic assumptions. The reason behind this scarcity is the inherent difficulty in achieving submodularity in the feature selection context. At the feature level, submodularity is in general *unattainable* due to its reliance on strong independence assumptions, which is unrealistic and impractical in real-world scenarios.

For example, in our evaluated PEMS-SF data set, each feature represents the occupancy of traffic lanes at a specific time step, whereas each modality captures the occupancy over a time horizon. Conditioned on the label (day of the week), the features are clearly correlated since traffic occupancy cannot undergo sudden changes. Meanwhile, however, the modalities are approximately independent (Table 1), as a modality represents the whole time series within a single day. It also provides little value to quantify the utility of all features individually, as each isolated feature has negligible predictive utility for the target.

Feature Selection with Submodularity. The literature in feature selection with submodularity primarily focused on linear regression problems. However, even in this constrained context, the applicability of submodularity is largely hindered by strong independence assumptions. For instance, Das and Kempe (2008) identifies that metrics like R-squared significantly deviate from submodularity, necessitating the imposition of strong conditions, such as absence of conditional suppressors, to enforce submodularity. Das and Kempe (2011) relaxes the condition and proposes submodularity ratio to measure the closeness of a function to submodularity. They also derive a relaxed performance guarantee for the greedy algorithm depending on the deviation of the utility function from submodularity. Khanna et al. (2017) further uses the submodularity ratio to derive performance bound for variants of the greedy algorithm, including stochastic greedy and distributed greedy.

In addition, submodularity is rarely applied to classification problems. The closest work to our knowledge is Kusner et al. (2014), where they propose a tree of classifier model and apply approximate submodularity for each node of the tree. Notably, the application is specific to optimizing test-time CPU cost, and does not address the broader context of a general classification setting. The scarcity of literature in this domain further validates our claim that at the feature level, submodularity is in general *unattainable* especially in the classification setting. Note that although there exists works applying submodularity to feature selection problems (Kawahara et al., 2009; Liu et al., 2013), they study submodularity on surrogate utility functions instead of directly on loss functions.

In comparison, we provide *the first unified theoretical framework based on submodularity for tackling the modality selection problem in both the classification and regression settings.*

8. Conclusion

We formulate a theoretical framework to study modality selection in multimodal learning. Under our framework, we propose a generic function that quantifies the learning utility of a modality, and identifies proper assumption(s) suitable for modeling heterogeneous multimodal data in various scenarios. We demonstrate the expressiveness and effectiveness of our framework in two classic learning settings. In classification setting with cross-entropy loss, we show the utility function manifests as Shannon mutual information. In regression setting with quadratic loss, the utility function manifests as the variance of the conditional expectation. In both settings, the utility function emits approximate submodularity, which allows us to derive efficient modality selection algorithms with an optimality guarantee. We connect feature importance scores to the context of modality selection, in which we can compute the Shapley value and MCI of a modality under tractable complexity. We evaluate our results on 2 synthetic and 4 real-world data sets.

Acknowledgements

The work is partially supported by the Defense Advanced Research Projects Agency (DARPA) under Cooperative Agreement Number: HR00112320012 and a research grant from the Amazon-Illinois Center on AI for Interactive Conversational Experiences (AICE).

Appendix A. Preliminary

Proposition A.1. *Let $X \in \mathcal{X}$, $Y \in \{0,1\}$ be random variables, $\mathcal{H}$ be the function class of all valid binary classifiers, i.e., $\mathcal{H} = \{h : \mathcal{X} \to [0,1]\}$, and $\ell(\cdot,\cdot)$ be the cross-entropy loss. We have:*

$$\inf_{h\in\mathcal{H}} \mathbb{E}[\ell(Y, h(X))] = H(Y \mid X)$$

Proof Let $x, \hat{y}$ be the instantiation of $X, \hat{Y}$ respectively, where $\hat{Y} \coloneqq h(X)$. $\mathbb{1}(\cdot)$ denotes the indicator function, and $D_{\mathrm{KL}}(\cdot \parallel \cdot)$ denotes the Kullback–Leibler divergence.

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\ell(Y,h(X))] &= \mathbb{E}_{X,Y}[-\mathbb{1}(Y=1)\log\hat{Y} - \mathbb{1}(Y=0)\log(1-\hat{Y})] \\
&= -\mathbb{E}_X[\mathbb{E}_{Y|x}[\mathbb{1}(Y=1)\log\hat{y} + \mathbb{1}(Y=0)\log(1-\hat{y})]] \\
&= -\mathbb{E}_X[\Pr(Y=1\mid x)\log\hat{y} + \Pr(Y=0\mid x)\log(1-\hat{y})] \\
&= \mathbb{E}_X[\Pr(Y=1\mid x)\log\frac{1}{\hat{y}} + \Pr(Y=0\mid x)\log\frac{1}{1-\hat{y}}] \\
&= \mathbb{E}_X[\Pr(Y=1\mid x)\log\frac{\Pr(Y=1\mid x)}{\hat{y}} + \Pr(Y=0\mid x)\log\frac{\Pr(Y=0\mid x)}{1-\hat{y}}] \\
&\quad + \mathbb{E}_X[-\Pr(Y=1\mid x)\log\Pr(Y=1\mid x) - \Pr(Y=0\mid x)\log\Pr(Y=0\mid x)] \\
&= \mathbb{E}_X[D_{\mathrm{KL}}(\Pr(Y\mid x) \parallel h(x))] + \mathbb{E}_X[H(Y\mid x)] \\
&= D_{\mathrm{KL}}(\Pr(Y\mid X) \parallel h(X)) + H(Y\mid X)
\end{aligned}$$

Since $H(Y \mid X) \geq 0$ and is unrelated to $h(X)$, $\mathbb{E}_{\mathcal{D}}[\ell(Y,h(X))]$ is minimum when $h(X) = \Pr(Y \mid X)$. ■

Proposition A.2. *Let X, Y be random variables, $f(X)$ be any function of X, we have:*

$$\inf_f \mathbb{E}[(Y - f(X))^2] = \mathbb{E}[\mathrm{Var}(Y \mid X)]$$

Proof By the law of total expectation,

$$\begin{aligned}
\mathbb{E}[(Y-f(X))^2] &= \mathbb{E}[(Y - \mathbb{E}[Y\mid X] + \mathbb{E}[Y\mid X] - f(X))^2] \\
&= \mathbb{E}[\mathbb{E}[(Y - \mathbb{E}[Y\mid X] + \mathbb{E}[Y\mid X] - f(X))^2 \mid X]] \\
&= \mathbb{E}[\mathbb{E}[(Y-\mathbb{E}[Y\mid X])^2 \mid X]] + \mathbb{E}[\mathbb{E}[(\mathbb{E}[Y\mid X] - f(X))^2 \mid X]] \\
&\quad + 2\mathbb{E}[\mathbb{E}[(Y-\mathbb{E}[Y\mid X])(\mathbb{E}[Y\mid X] - f(X)) \mid X]]
\end{aligned}$$

Because $\mathbb{E}[g(X)Y \mid X] = g(X)\mathbb{E}[Y \mid X]$ for any function $g(X)$, we can see that:

$$\mathbb{E}[\mathbb{E}[(Y-\mathbb{E}[Y\mid X])(\mathbb{E}[Y\mid X] - f(X)) \mid X]] = 0$$

Applying the definition of conditional variance, the equation above can be simplified as:

$$\begin{aligned}
\mathbb{E}[(Y-f(X))^2] &= \mathbb{E}[\mathbb{E}[(Y-\mathbb{E}[Y\mid X])^2 \mid X]] + \mathbb{E}[\mathbb{E}[(\mathbb{E}[Y\mid X] - f(X))^2 \mid X]] \\
&= \mathbb{E}[\mathrm{Var}(Y\mid X)] + \mathbb{E}[(\mathbb{E}[Y\mid X] - f(X))^2]
\end{aligned}$$

Since $\mathbb{E}[\mathrm{Var}(Y \mid X)] \geq 0$ and unaffected by $f(X)$, $\mathbb{E}[(Y - f(X))^2]$ is minimum when $f(X) = \mathbb{E}[Y \mid X]$. ■

Appendix B. Proofs for Main Text

Proposition 3.1. *Given* $Y \in \{0,1\}$ *and* $\ell(Y,\hat{Y}) \coloneqq \mathbb{1}(Y=1)\log\hat{Y} + \mathbb{1}(Y=0)\log(1-\hat{Y})$, $f_u(S) = I(S;Y)$.

Proof By Definition 3.1, we have:

$$f_u(S) = \inf_{c\in\mathcal{Y}} \mathbb{E}[\ell(Y,c)] - \inf_{h\in\mathcal{H}} \mathbb{E}[\ell(Y,h(S))]. \tag{11}$$

By Proposition A.1, we directly have

$$\inf_{h\in\mathcal{H}} \mathbb{E}[\ell(Y,h(S))] = H(Y \mid S).$$

Also, by the definition of log loss,

$$\inf_{c\in\mathcal{Y}} \mathbb{E}[\ell(Y,c)] = \inf_{c\in\mathcal{Y}} \Pr(Y=1)\log c + \Pr(Y=0)\log(1-c).$$

It is clear that $c = \Pr(Y=1)$ is the minimizer, namely

$$\inf_{c\in\mathcal{Y}} \mathbb{E}[\ell(Y,c)] = H(Y).$$

This result is intuitive because knowing the constant c does not help reduce any uncertainty in the label Y. Plugging in the derivation back into Equation (11), we have

$$\begin{aligned} f_u(S) &= H(Y) - H(Y \mid S) \\ &= I(S;Y). \end{aligned}$$

■

Proposition 3.2. $\forall M \subseteq N \subseteq V$, $I(N;Y) - I(M;Y) = I(N \setminus M; Y \mid M) \geq 0$.

Proof Let $N \coloneqq \{X_1, ..., X_n\}$, $M \coloneqq \{X_1, ..., X_m\}$, $n \geq m$.

$$\begin{aligned} I(N;Y) - I(M;Y) &= \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, ..., X_1) - \sum_{i=1}^{m} I(X_i; Y \mid X_{i-1}, ..., X_1) \\ &= \sum_{i=m+1}^{n} I(X_i; Y \mid X_{i-1}, ..., X_1) \\ &= I(N \setminus M; Y \mid M) \\ &\geq 0 \end{aligned}$$

■

Proposition 3.3. *Under Assumption 3.1,* $I(S;Y)$ *is* ϵ*-approximately submodular, i.e.,* $\forall A \subseteq B \subseteq V$, $e \in V \setminus B$, $I(A \cup \{e\}; Y) - I(A;Y) + \epsilon \geq I(B \cup \{e\}; Y) - I(B;Y)$.

Proof For A, we have:

$$\begin{aligned} I(A \cup \{e\}; Y) - I(A; Y) &= I(\{e\}; Y \mid A) \\ &= I(\{e\}; Y, A) - I(\{e\}; A) \\ &= I(\{e\}; Y) + I(\{e\}; A \mid Y) - I(\{e\}; A) \end{aligned}$$

Similarly, $I(B \cup \{e\}; Y) - I(B; Y) = I(\{e\}; Y) + I(\{e\}; B \mid Y) - I(\{e\}; B)$. Given Assumption 3.1 holds, we denote $I(\{e\}; A \mid Y) = \epsilon_A$ and $I(\{e\}; B \mid Y) = \epsilon_B$ where $\epsilon_A, \epsilon_B \leq \epsilon$. In the worst case where $\epsilon_A = 0$, strict submodularity is still satisfied if $\epsilon_B \leq I(\{e\}; B) - I(\{e\}; A)$, i.e.,

$$\begin{aligned} I(B \cup \{e\}; Y) - I(B; Y) &= I(\{e\}; Y) + I(\{e\}; B \mid Y) - I(\{e\}; B) \\ &= I(\{e\}; Y) - I(\{e\}; B) + \epsilon_B \\ &\leq I(\{e\}; Y) - I(\{e\}; B) + I(\{e\}; B) - I(\{e\}; A) \\ &= I(A \cup \{e\}; Y) - I(A; Y) \end{aligned}$$

But if $\epsilon_B > I(\{e\}; B) - I(\{e\}; A)$, strict submodularity will not hold. However, because $\epsilon_B \leq \epsilon$, we can define approximate submodularity characterized by the constant $\epsilon \geq 0$. Specifically:

$$\begin{aligned} I(B \cup \{e\}; Y) - I(B; Y) &= I(\{e\}; Y) + I(\{e\}; B \mid Y) - I(\{e\}; B) \\ &= I(\{e\}; Y) - I(\{e\}; B) + \epsilon_B \\ &\leq I(\{e\}; Y) - I(\{e\}; B) + \epsilon \\ &\leq I(\{e\}; Y) - I(\{e\}; A) + \epsilon \\ &\leq I(\{e\}; Y) - I(\{e\}; A) + \epsilon_A + \epsilon \\ &\leq I(A \cup \{e\}; Y) - I(A; Y) + \epsilon \end{aligned}$$

■

Proposition 3.4. *Given $S \subseteq V$ and $\ell(Y, \hat{Y}) \coloneqq (Y - \hat{Y})^2$, we have $f_u(S) = \mathrm{Var}(\mathbb{E}[Y \mid S])$.*

Proof By Definition 3.1 and Proposition A.2, we have:

$$\begin{aligned} f_u(S) &= \inf_{c \in \mathcal{Y}} \mathbb{E}[(Y - c)^2] - \inf_{h \in \mathcal{H}} \mathbb{E}[(Y - h(S))^2] \\ &= \mathbb{E}[(Y - \mathbb{E}[Y])^2] - \mathbb{E}[(Y - \mathbb{E}[Y \mid S])^2] \\ &= \mathrm{Var}(Y) - \mathbb{E}[\mathrm{Var}(Y \mid S)] \\ &= \mathrm{Var}(\mathbb{E}[Y \mid S]) \end{aligned}$$

■

Proposition 3.5. $\forall M \subseteq N \subseteq V$, $\mathrm{Var}(\mathbb{E}[Y \mid N]) - \mathrm{Var}(\mathbb{E}[Y \mid M]) = \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid N] \mid M)] \geq 0$.

Proof Let $K = \mathbb{E}[Y \mid N]$. Then by the law of total variance,

$$\mathrm{Var}(\mathbb{E}[Y \mid N]) = \mathrm{Var}(\mathbb{E}[K \mid M]) + \mathbb{E}[\mathrm{Var}(K \mid M)]$$

Further, by the law of total expectation, we have:

$$\mathrm{Var}(\mathbb{E}[K \mid M]) = \mathrm{Var}(\mathbb{E}[\mathbb{E}[Y \mid N] \mid M]) = \mathrm{Var}(\mathbb{E}[Y \mid M])$$

Therefore,

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[Y \mid N]) - \mathrm{Var}(\mathbb{E}[Y \mid M]) &= \mathrm{Var}(\mathbb{E}[Y \mid M]) + \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid N] \mid M)] - \mathrm{Var}(\mathbb{E}[Y \mid M]) \\ &= \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid N] \mid M)] \\ &\geq 0\end{aligned}$$

■

Proposition 3.6. *Under Assumption 3.2 and Assumption 3.3,* $\mathbb{E}[Y \mid S]$ *is linear in* S *for any* $S \subseteq V'$.

Proof Recall that given $\Phi(V) \sim \mathcal{N}(\mu, \Sigma)$ (Assumption 3.3), we have $\Sigma = Q\Lambda Q^\top$, and $V' = Q^{-1}\Phi(V) \sim \mathcal{N}(0, \Lambda)$. Because V' is a linear transformation of $\Phi(V)$, under Assumption 3.2, $\mathbb{E}[Y \mid V'] = \mathbb{E}[Y \mid Q^{-1}\Phi(V)]$ is also linear. Let $\mathbb{E}[Y \mid V'] \coloneqq \sum_{X_i \in V'} \alpha_i X_i + \alpha$, then we can show the following for any $S \subseteq V'$.

By the law of total expectation,

$$\begin{aligned}\mathbb{E}[Y \mid S] &= \mathbb{E}[\mathbb{E}[Y \mid V'] \mid S] \\ &= \mathbb{E}[\sum_{X_i \in V'} \alpha_i X_i + \alpha \mid S] \\ &= \sum_{X_i \in S} \alpha_i \mathbb{E}[X_i \mid S] + \sum_{X_i \in V' \setminus S} \alpha_i \mathbb{E}[X_i \mid S] + \mathbb{E}[\alpha \mid S]\end{aligned}$$

Because $V' \sim \mathcal{N}(0, \Lambda)$ and Λ is diagonal, each distinct pair of X_i, X_j from V' are independent. Thus, $\forall X_i \in S$, $\mathbb{E}[X_i \mid S] = \mathbb{E}[X_i \mid X_i] = X_i$; and $\forall X_i \in V' \setminus S$, $\mathbb{E}[X_i \mid S] = \mathbb{E}[X_i]$. Thus:

$$\sum_{X_i \in S} \alpha_i \mathbb{E}[X_i \mid S] + \sum_{X_i \in V' \setminus S} \alpha_i \mathbb{E}[X_i \mid S] + \mathbb{E}[\alpha \mid S] = \sum_{X_i \in S} \alpha_i X_i + c + \alpha$$

where $c = \sum_{X_i \in V' \setminus S} \alpha_i \mathbb{E}[X_i]$ and α are constants independent of S. This shows $\mathbb{E}[Y \mid S]$ is linear for any $S \subseteq V'$. ■

Proposition 3.7. *Under Assumption 3.2 and Assumption 3.3,* $\mathrm{Var}(\mathbb{E}[Y \mid S])$ *is a submodular function of* S *for any* $S \subseteq V'$.

Proof

By Definition 2.1, it is equivalent to prove: $\forall A \subseteq B \subseteq V',\ e \in V' \setminus B$,

$$\mathrm{Var}(\mathbb{E}[Y \mid A \cup \{e\}]) - \mathrm{Var}(\mathbb{E}[Y \mid A]) \geq \mathrm{Var}(\mathbb{E}[Y \mid B \cup \{e\}]) - \mathrm{Var}(\mathbb{E}[Y \mid B])$$

By Proposition 3.5, we have $\mathrm{Var}(\mathbb{E}[Y \mid A \cup \{e\}]) - \mathrm{Var}(\mathbb{E}[Y \mid A]) = \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid A \cup \{e\}] \mid A)]$ for A. Similarly, apply Proposition 3.5 to B, Appendix B is simplified to:

$$\mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid A \cup \{e\}] \mid A)] \geq \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid B \cup \{e\}] \mid B)]$$

Then by Proposition 3.6, we can express $\mathbb{E}[Y \mid A \cup \{e\}]$ as a linear function of $A \cup \{e\}$, i.e.,

$$\mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid A \cup \{e\}] \mid A)] = \mathbb{E}[\mathrm{Var}(\sum_{X_i \in A} \alpha_i X_i + \alpha_e e + c + \alpha \mid A)]$$

where $c = \sum_{X_i \in V \setminus A \cup \{e\}} \alpha_i \mathbb{E}[X_i]$ is a constant independent of $A \cup \{e\}$.

Because each distinct pair of $X_i,\ X_j$ from V' are independent by the construction of V', we can simplify the $\mathrm{Var}(\cdot)$ term from the last equation as the following:

$$\begin{aligned}
\mathrm{Var}(\sum_{X_i \in A} \alpha_i X_i + \alpha_e e + c + \alpha \mid A) &= \mathrm{Var}(\sum_{X_i \in A} \alpha_i X_i \mid A) + \mathrm{Var}(\alpha_e e \mid A) + \mathrm{Var}(c \mid A) + \mathrm{Var}(\alpha \mid A) \\
&= \mathrm{Var}(\sum_{X_i \in A} \alpha_i X_i \mid X_i) + \mathrm{Var}(\alpha_e e \mid A) \\
&= \alpha_e^2 \mathrm{Var}(e \mid A) \\
&= \alpha_e^2 \mathrm{Var}(e)
\end{aligned}$$

Thus, because $e \notin B$,

$$\mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid A \cup \{e\}] \mid A)] = \mathbb{E}_A[\alpha_e^2 \mathrm{Var}(e)] = \alpha_e^2 \mathrm{Var}(e)$$

Similarly, $\mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid B \cup \{e\}] \mid B)] = \alpha_e^2 \mathrm{Var}(e)$. Therefore,

$$\mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid A \cup \{e\}] \mid A)] = \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid B \cup \{e\}] \mid B)]$$

■

Theorem 4.1. *Under Assumption 3.1, let $q \in \mathbb{Z}^+$, and S_p be the selected subset from Algorithm 1 at iteration p, we have:*

$$I(S_p; Y) \geq (1 - e^{-\frac{p}{q}}) \max_{S:|S| \leq q} I(S; Y) - q\epsilon. \tag{6}$$

Proof Let $S^* \coloneqq \max_{S:|S|\leq q} I(S;Y)$ be the optimal subset with cardinality at most q. By Proposition 3.2, $|S^*| = q$. We order S^* as $\{X_1^*,\ ...,\ X_q^*\}$. Then for all positive integer $i \leq p$,

$$
\begin{aligned}
I(S^*;Y) &\leq I(S^* \cup S_i;Y) && (12)\\
&= I(S_i;Y) + \sum_{j=1}^{q} I(X_j^*;Y \mid S_i \cup \{X_{j-1}^*,...,X_1^*\}) && (13)\\
&= I(S_i;Y) + \sum_{j=1}^{q} (I(\{X_j^*,...,X_1^*\} \cup S_i;Y) - I(\{X_{j-1}^*,...,X_1^*\} \cup S_i;Y)) && (14)\\
&\leq I(S_i;Y) + \sum_{j=1}^{q} (I(\{X_j^*\} \cup S_i;Y) - I(S_i;Y) + \epsilon) && (15)\\
&\leq I(S_i;Y) + \sum_{j=1}^{q} (I(S_{i+1};Y) - I(S_i;Y) + \epsilon) && (16)\\
&\leq I(S_i;Y) + q(I(S_{i+1};Y) - I(S_i;Y) + \epsilon) && (17)
\end{aligned}
$$

Eq. (12) is from Proposition 3.2, Eq. (13) and Eq. (14) are by the chain rule of mutual information, Eq. (15) is from Proposition 3.3, Eq. (16) is by the definition of Algorithm 1 that $I(S_{i+1};Y) - I(S_i;Y)$ is maximized in each iteration i. Let $\delta_i \coloneqq I(S^*;Y) - I(S_i;Y)$, we can rewrite Eq. (17) into $\delta_i \leq q(\delta_i - \delta_{i+1} + \epsilon)$, which can be rearranged into $\delta_{i+1} \leq (1 - \frac{1}{q})\delta_i + \epsilon$.

Let $\delta_0 = I(S^*;Y) - I(S_0;Y)$. Since $S_0 = \emptyset$, we have $\delta_0 = I(S^*;Y)$. By the previous results, we can upper bound the quantity $\delta_p = I(S^*;Y) - I(S_p;Y)$ as follows:

$$
\begin{aligned}
\delta_p &\leq (1 - \frac{1}{q})\delta_{p-1} + \epsilon\\
&\leq (1 - \frac{1}{q})((1 - \frac{1}{q})\delta_{p-2} + \epsilon) + \epsilon\\
&\leq (1 - \frac{1}{q})^p \delta_0 + (1 + (1 - \frac{1}{q}) + ... + (1 - \frac{1}{q})^{p-1})\epsilon && (18)\\
&= (1 - \frac{1}{q})^p \delta_0 + (\frac{1 - (1 - \frac{1}{q})^{p-1+1}}{1 - (1 - \frac{1}{q})})\epsilon\\
&= (1 - \frac{1}{q})^p \delta_0 + (q - q(1 - \frac{1}{q})^p)\epsilon && (19)\\
&\leq (1 - \frac{1}{q})^p \delta_0 + q\epsilon\\
&\leq e^{-\frac{p}{q}} \delta_0 + q\epsilon && (20)
\end{aligned}
$$

Eq. (18) to Eq. (19) is through the summation of the geometric series $1 + (1 - \frac{1}{q}) + ... + (1 - \frac{1}{q})^{p-1}$. Eq. (20) is by the inequality $1 - x \leq e^{-x}$ for all $x \in \mathbb{R}$. Substitute the definitions of δ_p and δ_0 into Eq. (20) completes the proof. ■

Corollary 4.1. *Assume conditions in Theorem 4.1 hold, there exists optimal predictor* $h^*(S_p) = \Pr(Y \mid S_p)$ *such that:*

$$\mathbb{E}[\ell_{01}(Y, h^*(S_p))] \leq \mathbb{E}[\ell_{ce}(Y, h^*(S_p))] \leq H(Y) - (1 - e^{-\frac{p}{q}})I(S^*; Y) + q\epsilon \tag{7}$$

Proof Denote the quantity $(1 - e^{-\frac{p}{q}}) \max_{S:|S|\leq q} I(S; Y) - q\epsilon$ from Theorem 4.1 as letter b. By the definition of mutual information, we have $H(Y \mid S_p) \leq H(Y) - b$. Following Proposition A.1, $\inf_{h:S_p \to [0,1]} \mathbb{E}[\ell_{ce}(Y, h(S_p))] \leq H(Y) - b$. In other words, $\exists h^* = \Pr(Y \mid S_p)$ *s.t.* $\mathbb{E}[\ell_{ce}(Y, h^*(S_p))] \leq H(Y) - b$.

When the predictor is probabilistic (i.e., $h(X) = 0$ if and only if $h(X) \leq 0.5$), $\ell_{01}(Y, \hat{Y}) = \mathbb{1}(Y \neq \hat{Y})$ naturally extends to $Y\mathbb{1}(\hat{Y} \leq 0.5) + (1 - Y)\mathbb{1}(\hat{Y} > 0.5)$, which is upper bounded by $\ell_{ce}(Y, \hat{Y})$ for all $(Y, \hat{Y})$. Therefore, for the same h^* as above, we have:

$$\mathbb{E}[\ell_{01}(Y, h^*(S_p))] \leq \mathbb{E}[\ell_{ce}(Y, h^*(S_p))] \leq H(Y) - b$$

■

Corollary 4.2. *Assume conditions in Theorem 4.1 hold. There exists optimal predictors* $h_1^*(S_p) = \Pr(Y \mid S_p)$, $h_2^*(S^*) = \Pr(Y \mid S^*)$ *such that:*

$$\mathbb{E}[\ell_{ce}(Y, h_1^*(S_p))] - \mathbb{E}[\ell_{ce}(Y, h_2^*(S^*))] \leq e^{-\frac{p}{q}} I(S^*; Y) + q\epsilon \tag{8}$$

Proof Following Theorem 4.1, and denote $\arg\max_{S:|S|\leq q} I(S; Y)$ as S^*, we have:

$$\begin{aligned}
& I(S_p; Y) \geq (1 - e^{-\frac{p}{q}}) \max_{S:|S|\leq q} I(S; Y) - q\epsilon \\
\implies & H(Y) - H(Y \mid S_p) \geq (1 - e^{-\frac{p}{q}})(H(Y) - H(Y \mid S^*)) - q\epsilon \\
\implies & H(Y \mid S_p) - H(Y \mid S^*) \leq e^{-\frac{p}{q}}(H(Y) - H(Y \mid S^*)) + q\epsilon \\
\implies & H(Y \mid S_p) - H(Y \mid S^*) \leq e^{-\frac{p}{q}}(I(S^*; Y)) + q\epsilon
\end{aligned}$$

Using Proposition A.1 completes the proof. ■

Corollary 4.3. *Assume conditions in Theorem 4.2 hold. There exists an optimal predictor* $h^*(S_p) = \mathbb{E}[Y \mid S_p]$ *such that:*

$$\mathbb{E}[(Y - h^*(S_p))^2] \leq \mathrm{Var}(Y) - (1 - e^{-\frac{p}{q}})\mathrm{Var}(\mathbb{E}[Y \mid S^*]) \tag{10}$$

Proof Apply the law of total variance and Proposition A.2 to Theorem 4.2,

$$\begin{aligned}
& \mathrm{Var}(\mathbb{E}[Y \mid S_p]) \geq (1 - e^{-\frac{p}{q}})\mathrm{Var}(\mathbb{E}[Y \mid S^*]) \\
\implies & \mathrm{Var}(Y) - \mathbb{E}[\mathrm{Var}(Y \mid S_q)] \geq (1 - e^{-\frac{p}{q}})\mathrm{Var}(\mathbb{E}[Y \mid S^*]) \\
\implies & \inf_h \mathbb{E}[(Y - h(X))^2] \leq \mathrm{Var}(Y) - (1 - e^{-\frac{p}{q}})\mathrm{Var}(\mathbb{E}[Y \mid S^*]) \\
\implies & \mathbb{E}[(Y - h^*(X))^2] \leq \mathrm{Var}(Y) - (1 - e^{-\frac{p}{q}})\mathrm{Var}(\mathbb{E}[Y \mid S^*])
\end{aligned}$$

■

Corollary 4.4. *Assume conditions in Theorem 4.2 hold. There exists optimal predictors* $h_1^*(S_p) = \mathbb{E}[Y \mid S_p]$, $h_2^*(S^*) = \mathbb{E}[Y \mid S^*]$ *such that:*

$$\mathbb{E}[(Y - h_1^*(S_p))^2] - \mathbb{E}[(Y - h_2^*(S^*))^2] \leq e^{-\frac{p}{q}} \mathrm{Var}(\mathbb{E}[Y \mid S^*])$$

Proof Apply the law of total variance and Proposition A.2 to Theorem 4.2,

$$\begin{aligned}
&\mathrm{Var}(\mathbb{E}[Y \mid S_p]) \geq (1 - e^{-\frac{p}{q}}) \mathrm{Var}(\mathbb{E}[Y \mid S^*]) \\
&\implies \mathrm{Var}(Y) - \mathbb{E}[\mathrm{Var}(Y \mid S_q)] \geq (1 - e^{-\frac{p}{q}})(\mathrm{Var}(Y) - \mathbb{E}[\mathrm{Var}(Y \mid S^*)]) \\
&\implies \mathbb{E}[\mathrm{Var}(Y \mid S^*)] - \mathbb{E}[\mathrm{Var}(Y \mid S_q)] \geq -e^{-\frac{p}{q}}(\mathrm{Var}(Y) - \mathbb{E}[\mathrm{Var}(Y \mid S^*)]) \\
&\implies \mathbb{E}[(Y - h_1^*(S_p))^2] - \mathbb{E}[(Y - h_2^*(S^*))^2] \leq e^{-\frac{p}{q}} \mathrm{Var}(\mathbb{E}[Y \mid S^*])
\end{aligned}$$

■

Proposition 5.1. *Under Assumption 3.1,* $I(S;Y)$ *is* ϵ*-approximately sub-additive for any* $S \subseteq V$*, i.e.,* $I(S \cup S';Y) \leq I(S;Y) + I(S';Y) + \epsilon$*.*

Proof

$$\begin{aligned}
I(S \cup S';Y) &= I(S;Y) + I(S';Y \mid S) \\
&= I(S;Y) + I(S \cup Y;S') - I(S;S') \\
&= I(S;Y) + I(S';Y) + I(S;S' \mid Y) - I(S;S') && (21) \\
&\leq I(S;Y) + I(S';Y) + \epsilon && (22)
\end{aligned}$$

Eq. (21) to Eq. (22) because $I(S;S' \mid Y) \leq \epsilon$ by Assumption 3.1, and $I(S;S')$ is always non-negative. ■

Proposition 5.2. *If Assumption 5.1 holds,* $I(S;Y)$ *is* ϵ*-approximately super-additive for any* $S \subseteq V$*, i.e.,* $\forall S, S' \subseteq V, S \cap S' = \emptyset$*,* $I(S \cup S';Y) \geq I(S;Y) + I(S';Y) - \epsilon$*.*

Proof Similarly to the proof of Proposition 5.1, we have:

$$\begin{aligned}
I(S \cup S';Y) &= I(S;Y) + I(S';Y) + I(S;S' \mid Y) - I(S;S') && (23) \\
&\geq I(S;Y) + I(S';Y) - \epsilon && (24)
\end{aligned}$$

Eq. (23) to Eq. (24) because $I(S;S') \leq \epsilon$ by Assumption 5.1, and $I(S;S' \mid Y)$ is non-negative. ■

Proposition 5.3. *If* $I(S;Y)$ *is both* ϵ*-approximately sub- and super-additive for any* $S \subseteq V$*, we have* $I(X_i;Y) - \epsilon \leq \phi_{I,X_i} \leq I(X_i;Y) + \epsilon$ *for any* $X_i \in V$*.*

Proof By Proposition 5.1 and Proposition 5.2, for any $X_i \in V$ and $S \subseteq V$, we have:

$$I(X_i;Y) - \epsilon \leq I(S \cup \{X_i\};Y) - I(S;Y) \leq I(X_i;Y) + \epsilon$$

Let's first apply the right inequality in Appendix B to Definition 2.2. Because $I(X_i;Y)+\epsilon$ is independent of S, we can simplify the calculation of the upper bound of ϕ_{I,X_i} as follows.

$$\begin{aligned}
\phi_{I,X_i} &= \sum_{S \subseteq V \setminus \{X_i\}} \frac{|S|!(|V|-|S|-1)!}{|V|!}(I(S \cup \{i\};Y) - I(S;Y)) \\
&\leq \sum_{S \subseteq V \setminus \{i\}} \frac{|S|!(|V|-|S|-1)!}{|V|!}(I(X_i;Y) + \epsilon) \\
&= \sum_{|S|=0}^{|V|-1} \binom{|V|-1}{|S|} \frac{|S|!(|V|-|S|-1)!}{|V|!}(I(X_i;Y) + \epsilon) \\
&= \sum_{|S|=0}^{|V|-1} \frac{(|V|-1)!}{|S|(|F|-1-|S|)!} \frac{|S|!(|V|-|S|-1)!}{|V|!}(I(X_i;Y) + \epsilon) \\
&= \sum_{|S|=0}^{|V|-1} \frac{1}{|V|}(I(X_i;Y) + \epsilon) \\
&= I(X_i;Y) + \epsilon
\end{aligned}$$

Applying the same procedure to the left inequality in Appendix B to Definition 2.2, we have $\phi_{I,X_i} \geq I(X_i;Y) - \epsilon$. Combining both results completes the proof. ■

Proposition 5.4. *Under Assumption 3.1, we have $I(X_i;Y) \leq \phi^{mci}_{I,X_i} \leq I(X_i;Y) + \epsilon$ for any $X_i \in V$.*

Proof By Proposition 3.3, $I(\cdot;Y)$ would be approximately submodular under Assumption 3.1, thus:

$$\begin{aligned}
I(X_i;Y) + \epsilon &= I(\emptyset \cup X_i;Y) - I(\emptyset;Y) + \epsilon \\
&\geq \max_{S \subseteq V} I(S \cup X_i;Y) - I(S;Y) = \phi^{mci}_{I,X_i}
\end{aligned}$$

On the other hand, if $\arg\max_{S \subseteq V} I(S \cup X_i;Y) - I(S;Y) = \emptyset$, we have $\phi^{mci}_{I,X_i} = I(\emptyset \cup X_i;Y) - I(\emptyset;Y) = I(X_i;Y)$. If $\arg\max_{S \subseteq V} I(S \cup X_i;Y) - I(S;Y)$ is some non-empty subset A, we have $\phi^{mci}_{I,X_i} = I(A \cup X_i;Y) - I(A;Y) \geq I(\emptyset \cup X_i;Y) - I(\emptyset;Y)$. In this case, $\phi^{mci}_{I,X_i} \geq I(X_i;Y)$. Combining both inequalities completes the proof. ■

Proposition 5.5. *Under Assumption 3.2 and Assumption 3.3,*

- $\mathrm{Var}(\mathbb{E}[Y \mid S])$ *is additive for any $S \subseteq V'$.*

- $\phi_{\mathrm{Var}(\mathbb{E}[Y|\cdot]),X_i} = \phi^{mci}_{\mathrm{Var}(\mathbb{E}[Y|\cdot]),X_i} = \mathrm{Var}(\mathbb{E}[Y \mid X_i])$ *for any* $X_i \in V'$.

Proof Additivity. It is equivalent to show: $\forall S,\ S' \subseteq V',\ S \cap S' = \emptyset$,

$$\mathrm{Var}(\mathbb{E}[Y \mid S \cup S']) = \mathrm{Var}(\mathbb{E}[Y \mid S]) + \mathrm{Var}(\mathbb{E}[Y \mid S'])$$

First, by the law of total variance, $\mathrm{Var}(\mathbb{E}[Y \mid S \cup S']) = \mathrm{Var}(\mathbb{E}[Y \mid S]) + \mathbb{E}_S[\mathrm{Var}(\mathbb{E}[Y \mid S \cup S'] \mid S)]$. Thus we can show the following instead:

$$\mathbb{E}_S[\mathrm{Var}(\mathbb{E}[Y \mid S \cup S'] \mid S)] = \mathrm{Var}(\mathbb{E}[Y \mid S'])$$

Since $\mathbb{E}[Y \mid V']$ is linear by Proposition 3.6, we can denote $\mathbb{E}[Y \mid V'] \coloneqq \sum_{X_i \in V'} \alpha_i X_i + \alpha$. Accordingly,

$$\begin{aligned}
\mathbb{E}[Y \mid S \cup S'] &= \mathbb{E}[\mathbb{E}[Y \mid V'] \mid S \cup S'] \\
&= \mathbb{E}[\sum_{X_i \in V'} \alpha_i X_i + \alpha \mid S \cup S'] \\
&= \sum_{X_i \in S \cup S'} \alpha_i \mathbb{E}[X_i \mid S \cup S'] + \sum_{X_i \in V' \setminus (S \cup S')} \alpha_i \mathbb{E}[X_i \mid S \cup S'] + \mathbb{E}[\alpha \mid S \cup S']
\end{aligned}$$

Each distinct pair of $X_i,\ X_j \in V'$ are independent by the definition of V'. Thus, for any $X_i \in S \cup S'$, $\mathbb{E}[X_i \mid S \cup S'] = \mathbb{E}[X_i \mid X_i] = X_i$; and for any $X_i \in V \setminus (S \cup S')$, $\mathbb{E}[X_i \mid S \cup S'] = \mathbb{E}[X_i]$. $\mathbb{E}[Y \mid S \cup S']$ can be further simplified accordingly:

$$\mathbb{E}[Y \mid S \cup S'] = \sum_{X_i \in S} \alpha_i X_i + \sum_{X_i \in S'} \alpha_i X_i + \sum_{X_i \in V' \setminus (S \cup S')} \alpha_i \mathbb{E}[X_i] + \alpha$$

where α and each $\mathbb{E}[X_i],\ X_i \in V' \setminus (S \cup S')$ are constants.

By the independence of variables in V' again,

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[Y \mid S \cup S'] \mid S) &= \mathrm{Var}(\sum_{X_i \in S} \alpha_i X_i \mid S) + \mathrm{Var}(\sum_{X_i \in S'} \alpha_i X_i \mid S) \\
&\quad + \mathrm{Var}(\sum_{X_i \in V' \setminus (S \cup S')} \alpha_i \mathbb{E}[X_i] \mid S) + \mathrm{Var}(\alpha \mid S) \\
&= \mathrm{Var}(\sum_{X_i \in S} \alpha_i X_i \mid X_i) + \mathrm{Var}(\sum_{X_i \in S'} \alpha_i X_i \mid S) \\
&= \sum_{X_i \in S'} \alpha_i^2 \mathrm{Var}(X_i)
\end{aligned}$$

Since $S \cap S' = \emptyset$, we have $\mathbb{E}_S[\mathrm{Var}(\mathbb{E}[Y \mid S \cup S'] \mid S)] = \mathbb{E}_S[\sum_{X_i \in S'} \alpha_i^2 \mathrm{Var}(X_i)] = \sum_{X_i \in S'} \alpha_i^2 \mathrm{Var}(X_i)$.

Finally, we can also derive the following for $\mathrm{Var}(\mathbb{E}[Y \mid S'])$ by Assumption 3.2 and the independence between any $X_i,\ X_j \in V',\ i \neq j$.

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[Y \mid S']) &= \mathrm{Var}(\sum_{X_i \in S'} \alpha_i X_i + \sum_{X_i \in V' \setminus S'} \alpha_i \mathbb{E}[X_i] + \alpha) \\
&= \sum_{X_i \in S'} \alpha_i^2 \mathrm{Var}(X_i)
\end{aligned}$$

Thus, $\mathbb{E}_S[\text{Var}(\mathbb{E}[Y \mid S \cup S'] \mid S)] = \text{Var}(\mathbb{E}[Y \mid S'])$. And $\text{Var}(\mathbb{E}[Y \mid S])$ is additive for any $S \subseteq V'$.

Shapley. By additivity, for any $X_i \in V'$ and $S \subseteq V'$, we have $\text{Var}(\mathbb{E}[Y \mid S \cup \{X_i\}]) - \text{Var}(\mathbb{E}[Y \mid S]) = \text{Var}(\mathbb{E}[Y \mid X_i])$. Follow similar steps as the proof of Proposition 5.3, we can derive that $\phi_{\text{Var}(\mathbb{E}[Y|\cdot]),X_i} = \text{Var}(\mathbb{E}[Y \mid X_i])$ and thus complete the proof.

MCI. Denote $v_1(\cdot) = \text{Var}(\mathbb{E}[Y \mid \cdot])$. By Definition 2.3 and additivity, for any $X_i \in V'$ and $S \subseteq V'$,

$$\phi^{mci}_{v_1,X_i} = \max_{S' \subseteq V'}(v_1(S \cup \{X_i\}) - v_1(S)) = \max_{S' \subseteq V'} v_1(X_i) = v_1(X_i)$$

■

Appendix C. Experimental Details

Regression Synthetic Data Set. To construct the synthetic data set, the key part is to construct a covariance matrix with desired block-diagonal and off-diagonal values. We control the block-diagonal values by the matrix A and the off-diagonal values by the matrix B. For matrix B, we fill in each entry by uniformly sampling from $[-1, 1]$, then multiply it with its transpose to ensure B is positive semi-definite (PSD). Then, we normalize B by its row sum to ensure that each entry is still between -1 and 1 after multiplication. Similarly, for matrix A, we construct 10 PSD matrices with desired block sizes, and fill them in the diagonal. If we want the block-diagonal values to range from $[-1, 1]$ and the off-diagonal values to range from $[-\epsilon, \epsilon]$, the covariance matrix will be constructed as

$$cov = (1 - \epsilon)A + \epsilon B$$

Due to the fact that the addition of two PSD matrices is still PSD, the matrix *cov* is a valid covariance matrix.

References

Massih R Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. *Advances in Neural Information Processing Systems*, 2009.

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. mslam: Massively multilingual joint pre-training for speech and text. *arXiv preprint arXiv:2202.01374*, 2022.

Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on neural networks*, 1994.

F Guillaume Blanchet, Pierre Legendre, and Daniel Borcard. Forward selection of explanatory variables. *Ecology*, 2008.

Leo Breiman. Random forests. *Machine learning*, 2001.

Luis M. Candanedo, Véronique Feldheim, and Dominique Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 2017.

Amnon Catav, Boyang Fu, Yazeed Zoabi, Ahuva Libi Weiss Meilik, Noam Shomron, Jason Ernst, Sriram Sankararaman, and Ran Gilad-Bachrach. Marginal contribution feature importance-an axiomatic approach for explaining data. In *International Conference on Machine Learning*, 2021.

Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 2014.

Abhimanyu Das and David Kempe. Algorithms for subset selection in linear regression. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, 2008.

Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv preprint arXiv:1102.3975*, 2011.

Angus Dempster, François Petitjean, and Geoffrey I Webb. Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 2020.

Pedro Domingos. Every model learned by gradient descent is approximately a kernel machine. *arXiv preprint arXiv:2012.00152*, 2020.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, 1998.

Ulrich Faigle and Walter Kern. The shapley value for cooperative games under precedence constraints. *International Journal of Game Theory*, 1992.

Farzan Farnia and David Tse. A minimax approach to supervised learning. *Advances in Neural Information Processing Systems*, 2016.

Shaheen S Fatima, Michael Wooldridge, and Nicholas R Jennings. A linear approximation method for the shapley value. *Artificial Intelligence*, 2008.

Christos A Frantzidis, Charalampos Bratsas, Manousos A Klados, Evdokimos Konstantinidis, Chrysa D Lithari, Ana B Vivas, Christos L Papadelis, Eleni Kaldoudi, Costas Pappas, and Panagiotis D Bamidis. On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications. *IEEE Transactions on Information Technology in Biomedicine*, 2010.

Satoru Fujishige. *Submodular functions and optimization*. 2005.

Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Estimating mutual information for discrete-continuous mixtures. *Advances in Neural Information Processing Systems*, 2017.

Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 2011.

Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data*, 2012.

Peter D Grünwald and A Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *the Annals of Statistics*, 2004.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 2000.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 2021.

Satoru Iwata. Submodular function minimization. *Mathematical Programming*, 2008.

Rishabh K Iyer and Jeff A Bilmes. Submodular optimization with submodular cover and submodular knapsack constraints. *Advances in Neural Information Processing Systems*, 2013.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Yoshinobu Kawahara, Kiyohito Nagano, Koji Tsuda, and Jeff A Bilmes. Submodularity cuts and applications. *Advances in Neural Information Processing Systems*, 2009.

Rajiv Khanna, Ethan Elenberg, Alex Dimakis, Sahand Negahban, and Joydeep Ghosh. Scalable greedy feature selection via weak submodularity. In *Artificial Intelligence and Statistics*, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

Andreas Krause and Daniel Golovin. Submodular function maximization. *Tractability*, 2014.

Andreas Krause and Carlos E Guestrin. Near-optimal nonmyopic value of information in graphical models. *arXiv preprint arXiv:1207.1394*, 2012.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 2008.

Andreas Krause, Carlos Guestrin, Anupam Gupta, and Jon Kleinberg. Robust sensor placements at informative and communication-efficient locations. *ACM Transactions on Sensor Networks*, 2011.

I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, 2020.

Matt Kusner, Wenlin Chen, Quan Zhou, Zhixiang Eddie Xu, Kilian Weinberger, and Yixin Chen. Feature-cost sensitive learning with submodular trees of classifiers. In *Conference on Artificial Intelligence*, 2014.

Yann LeCun and Corinna Cortes. Mnist handwritten digit database. 1998.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. *Advances in Neural Information Processing Systems*, 1989.

Jon Lee, Maxim Sviridenko, and Jan Vondrák. Submodular maximization over multiple matroids via generalized exchange properties. *Mathematics of Operations Research*, 2010.

Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *International Conference on Knowledge Discovery and Data Mining*, 2007.

Yuzong Liu, Kai Wei, Katrin Kirchhoff, Yisong Song, and Jeff Bilmes. Submodular feature selection for high-dimensional acoustic score spaces. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

Zhentao Liu, Min Wu, Weihua Cao, Luefeng Chen, Jianping Xu, Ri Zhang, Mengtian Zhou, and Junwei Mao. A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA Journal of Automatica Sinica*, 2017.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018.

Markus Löning, Anthony Bagnall, Sajaysurya Ganesh, Viktor Kazakov, Jason Lines, and Franz J Király. sktime: A unified interface for machine learning with time series. *arXiv preprint arXiv:1909.07872*, 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Tomasz P Michalak, Karthik V Aadithya, Piotr L Szczepanski, Balaraman Ravindran, and Nicholas R Jennings. Efficient computation of the shapley value for game-theoretic network centrality. *Journal of Artificial Intelligence Research*, 2013.

Baharan Mirzasoleiman, Amin Karbasi, Rik Sarkar, and Andreas Krause. Distributed submodular maximization: Identifying representative elements in massive data. *Advances in Neural Information Processing Systems*, 2013.

Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrák, and Andreas Krause. Lazier than lazy greedy. In *Conference on Artificial Intelligence*, 2015.

Piyushkumar A Mundra and Jagath C Rajapakse. Svm-rfe with mrmr filter for gene selection. *IEEE transactions on nanobioscience*, 2009.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 1978.

OpenAI. GPT-4 Technical Report. `https://cdn.openai.com/papers/gpt-4.pdf`, 2023.

Johannes Pittermann, Angela Pittermann, and Wolfgang Minker. Emotion recognition and adaptation in spoken dialogue systems. *International Journal of Speech Technology*, 2010.

Pavel Pudil, Jana Novovičová, and Josef Kittler. Floating search methods in feature selection. *Pattern recognition letters*, 1994.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Michael A Redmond and Timothy Highley. Empirical analysis of case-editing approaches for numeric prediction. In *Innovations in Computing Sciences and Software Engineering*, 2010.

Alvin E Roth. *The Shapley value: essays in honor of Lloyd S. Shapley*. 1988.

Lloyd S. Shapley. *A Value for N-Person Games*. 1952.

Xinwei Sun, Yilun Xu, Peng Cao, Yuqing Kong, Lingjing Hu, Shanghang Zhang, and Yizhou Wang. Tcgm: An information-theoretic framework for semi-supervised multi-modality learning. In *European conference on computer vision*, 2020.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1996.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning*, 2015.

Jason Weston, André Elisseeff, Bernhard Schölkopf, and Mike Tipping. Use of the zero norm with linear models and kernel methods. *Journal of Machine Learning Research*, 2003.

Martha White, Xinhua Zhang, Dale Schuurmans, and Yao-liang Yu. Convex multi-view subspace learning. *Advances in Neural Information Processing Systems*, 2012.

Eyal Winter. The shapley value. *Handbook of game theory with economic applications*, 2002.

Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nuwa: Visual synthesis pre-training for neural visual world creation. *arXiv preprint arXiv:2111.12417*, 2021.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 2018.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

Changqing Zhang, Zongbo Han, Huazhu Fu, Joey Tianyi Zhou, Qinghua Hu, et al. Cpm-nets: Cross partial multi-view networks. *Advances in Neural Information Processing Systems*, 2019.

Mingtian Zhang, Tim Z Xiao, Brooks Paige, and David Barber. Improving vae-based representation learning. *arXiv preprint arXiv:2205.14539*, 2022.

Yizhe Zhu, Martin Renqiang Min, Asim Kadav, and Hans Peter Graf. S3vae: Self-supervised sequential vae for representation disentanglement and data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Honglei Zhuang, Yihan Sun, Jie Tang, Jialin Zhang, and Xiaoming Sun. Influence maximization in dynamic social networks. In *2013 IEEE 13th International Conference on Data Mining*, 2013.