Proceedings of Machine Learning Research vol 242:1–12, 2024

Submodular Information Selection for Hypothesis Testing with Misclassification Penalties

Jayanth Bhargav JBHARGAV@PURDUE.EDU
Mahsa Ghasemi MAHSA@PURDUE.EDU
Shreyas Sundaram SUNDARA2@PURDUE.EDU
Elmore Family School of Electrical & Computer Engineering, Purdue University, IN 47907 USA

Abstract

We consider the problem of selecting an optimal subset of information sources for a hypothesis testing/classification task where the goal is to identify the true state of the world from a finite set of hypotheses, based on finite observation samples from the sources. In order to characterize the learning performance, we propose a misclassification penalty framework, which enables non-uniform treatment of different misclassification errors. In a centralized Bayesian learning setting, we study two variants of the subset selection problem: (i) selecting a minimum cost information set to ensure that the maximum penalty of misclassifying the true hypothesis is below a desired bound and (ii) selecting an optimal information set under a limited budget to minimize the maximum penalty of misclassifying the true hypothesis. Under certain assumptions, we prove that the objective (or constraints) of these combinatorial optimization problems are weak (or approximate) submodular, and establish high-probability performance guarantees for greedy algorithms. Further, we propose an alternate metric for information set selection which is based on the total penalty of misclassification. We prove that this metric is submodular and establish near-optimal guarantees for the greedy algorithms for both the information set selection problems. Finally, we present numerical simulations to validate our theoretical results over several randomly generated instances.[1]

Keywords: Combinatorial Optimization, Bayesian Classification, Submodularity, Greedy Algorithms, Finite Sample Convergence

1. Introduction

In many autonomous systems, agents depend on predictions made by classifiers for making decisions (or taking actions), and may have to pay a high cost for acting on erroneous predictions. An example of this is an incident of an autonomous vehicle crash caused due to the vision system misclassifying a white truck as a bright sky (NHTSA (2016)). In such scenarios, one needs to ensure minimal risk associated with misclassification. In order to improve the quality of predictions, one may need to select an optimal set of features (or observations), often provided by information sources (or sensors), that can best describe the true state. In many practical scenarios, due to limitations on communication or compute resources, one can only query data from a small subset of information sources (Krause and Cevher (2010); Chepuri and Leus (2014); Hashemi et al. (2020)). Moreover, one may also need to pay a certain cost in order to obtain measurements from information sources (Krause et al. (2008)). Thus, a fundamental problem that arises in such scenarios is to select a subset of information sources with minimal cost or under a limited budget, while ensuring

1. An extended version of this paper that includes all the omitted proofs can be found on arXiv as Bhargav et al. (2024).

certain learning performance using the observations provided by the selected sources. In order to characterize the quality of an information set, we propose a framework based on misclassification penalties, specified by a penalty matrix. The goal is to select an information set that minimizes the maximum penalty of misclassifying the true state. As a motivating example, consider a surveillance task, where identifying a target of interest is of importance. One many have to pay a penalty for misclassifying the true state, for instance, misclassifying a drone (an intruder) as a bird. However, the event of misclassifying a bird as a drone may have a different penalty associated with it. The penalty matrix captures the fact that different misclassification errors incur different penalties.

1.1. Related Work

Misclassification risk and uncertainty quantification for various types of classifiers have been very well studied in the literature (Adams and Hand (1999); Pendharkar (2017); Hou et al. (2013)). In Sensoy et al. (2021), the authors propose a risk-calibrated classifier to reduce the costs associated with misclassification errors, and empirically show the effectiveness of their algorithm, in a deep learning framework. In Elkan (2001), the authors study cost-sensitive learning for class balancing in order to improve the quality of predictions in decision tree learning methods. In our work, we consider a hypothesis testing (or classification) task in a Bayesian learning framework.

A subset of the literature has addressed the problem of sequential information gathering within a limited budget (Hollinger and Sukhatme (2013); Chen et al. (2015)). The authors of Golovin et al. (2010) study data source selection for a monitoring application, where the sources are selected sequentially in order to estimate certain parameters of an environment. In Ghasemi and Topcu (2019), the authors study sequential information gathering under a limited budget for a robotic navigation task. In contrast, we consider the scenario where the information set is selected *a priori*.

A substantial body of work focuses on the study of submodularity (and/or weak submodularity) and greedy techniques with provable guarantees for feature selection in sparse learning (Krause and Cevher (2010); Chepuri and Leus (2014)); sensor selection for estimation (Mo et al. (2011); Hashemi et al. (2020)), Kalman filtering (Ye et al. (2020)), and mixed-observable Markov decision processes (Bhargav et al. (2023)). Along the lines of these works, we leverage the weak submodularity property of the performance metric and present greedy algorithms with performance guarantees.

The closest paper to our work is Ye et al. (2021), in which the authors studied data source selection for Bayesian learning, where the learning performance was characterized by a total variation error metric based on the asymptotic belief. However, we consider a non-asymptotic setting, where the learning performance is characterized by misclassification penalties. Building upon the results in Ye et al. (2021) and Das and Kempe (2018), we establish theoretical guarantees for greedy information selection algorithms presented in this paper.

1.2. Contributions

We consider two variants of an information subset selection problem for a hypothesis testing task (i) selecting a minimum cost information set to ensure the maximum penalty for misclassifying the true hypothesis is below a desired bound and (ii) optimal information set selection under a limited budget to minimize the maximum penalty of misclassifying the true hypothesis. First, we prove that the maximum penalty metric is weak submodular by characterizing its submodularity ratio, and establish high-probability guarantees for greedy algorithms for both the problems, along with the associated finite sample convergence rates for the Bayesian beliefs. Next, we propose an alternate

metric based on the total penalty of misclassification. We prove that this metric is submodular, and establish near-optimal guarantees for the greedy algorithms. Finally, we evaluate the empirical performance of the proposed greedy algorithms over several randomly generated problem instances.

2. Minimum-Cost Information Set Selection Problem

In this section, we formulate the minimum-cost information set selection problem. Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$, where $m = |\Theta|$, be a finite set of possible hypotheses (also referred to as classes or states), of which one of them is the true state of the world. We consider a set $\mathcal{D} = \{1, 2, \ldots, n\}$ of information sources (or data streams) from which we need to select a subset $\mathcal{I} \subseteq \mathcal{D}$. At each time step $t \in \mathbb{Z}_{\geq 1}$, the observation provided by the information source $i \in \mathcal{D}$ is denoted as $o_{i,t} \in O_i$, where O_i is the observation space of the source i. Each information source $i \in \mathcal{D}$ is associated with an observation likelihood function $\ell_i(\cdot|\theta)$, which is conditioned on the state of the world $\theta \in \Theta$. At any time t, conditioned on the true state of the world $\theta \in \Theta$, a joint observation profile of n information sources, denoted as $o_t = (o_{1,t}, \ldots, o_{n,t}) \in \mathcal{O}$ where $\mathcal{O} = O_1 \times \ldots \times O_n$, is generated by the joint likelihood function $\ell(\cdot|\theta)$. We make the following assumption on the observation model (e.g., see Jadbabaie et al. (2012); Liu et al. (2014); Lalitha et al. (2014) for detailed discussions).

Assumption 1: *The observation space O_i associated with each information source $i \in \mathcal{D}$ is finite, and the likelihood function $\ell_i(\cdot|\theta)$ satisfies $\ell_i(\cdot|\theta) > 0$ for all $o_i \in O_i$ and for all $\theta \in \Theta$. We assume that the designer knows $\ell_i(\cdot|\theta)$ for all $\theta \in \Theta$ and all $i \in \mathcal{D}$. For all $\theta \in \Theta$, conditioned on the true state, the observations are independent of each other over time, i.e., $\{o_{i,1}, o_{i,2}, \ldots\}$ is a sequence of independent identically distributed (i.i.d.) random variables, given a true state $\theta \in \Theta$.*

Consider the scenario where a designer at a central node needs to select a subset of information sources in order to identify the true state of the world. Each source $i \in \mathcal{D}$ has a selection cost $c_i \in \mathbb{R}_{>0}$. For any subset $\mathcal{I} \subseteq \mathcal{D}$ with $|\mathcal{I}| = k$, let $\{s_1, s_2, \ldots, s_k\}$ denote the set of information sources. The cost of the information set $\mathcal{I}$ is given by $c(\mathcal{I}) = \sum_{s_i \in \mathcal{I}} c_{s_i}$. The joint observation conditioned on the $\theta \in \Theta$ of this information set at time t is defined as $o_{\mathcal{I},t} = \{o_{s_1,t}, \ldots, o_{s_k,t}\} \in O_{s_1} \times \ldots \times O_{s_k}$, and is generated by the joint likelihood function $\ell_{\mathcal{I}}(\cdot|\theta) = \Pi_{i=1}^{k} \ell_{s_i}(\cdot|\theta)$ (by Assumption 1), and the central designer knows $\ell_{\mathcal{I}}(\cdot|\theta)$ for all $\mathcal{I} \subseteq \mathcal{D}$ and for all $\theta \in \Theta$.

Assumption 1 also implies the existence of a constant L[2] $\in (0, \infty)$ such that:

$$\max_{i \in \mathcal{D}} \max_{o_i \in O_i} \max_{\theta_p, \theta_q \in \Theta} \left| \log \frac{\ell_i(o_i \mid \theta_p)}{\ell_i(o_i \mid \theta_q)} \right| \leq L. \tag{1}$$

For a true state $\theta_p \in \Theta$, we define $\mathbb{P}^{\theta_p} = \prod_{t=1}^{\infty} \ell(\cdot|\theta_p)$ to be the probability measure. For the sake of brevity, we will say that an event occurs almost surely to mean that it occurs almost surely w.r.t. the probability measure $\mathbb{P}^{\theta_p}$. As the data comes in, the central node updates its belief over the set of possible hypotheses using the standard Bayes' rule. Let $\mu_t^{\mathcal{I}}(\theta)$ denote the belief of the central designer (or node) that θ is the true hypothesis at time step t based on the information sources in $\mathcal{I}$, and let $\mu_0(\theta)$ denote the initial belief (or prior) of the central node that θ is the true state of the world, with $\sum_{\theta \in \Theta} \mu_0(\theta) = 1$. The Bayesian update rule is given by

$$\mu_{t+1}^{\mathcal{I}}(\theta) = \frac{\mu_0(\theta) \prod_{j=0}^{t} \ell_{\mathcal{I}}(o_{\mathcal{I},j+1}|\theta)}{\sum_{\theta_i \in \Theta} \mu_0(\theta_i) \prod_{j=0}^{t} \ell_{\mathcal{I}}(o_{\mathcal{I},j+1}|\theta_i)} \quad \forall \theta \in \Theta. \tag{2}$$

2. The constant L is an upper bound on the maximum difference between the log-likelihood of an observation from an information source under any two hypotheses, which we will use later in our analyses.

For a hypothesis $\theta \in \Theta$ and an information set $\mathcal{I} \subseteq \mathcal{D}$, we have the following.

Definition 1 (Observationally Equivalent Set) *(Ye et al. (2021)) For a given hypothesis (or class) $\theta \in \Theta$ and a given $\mathcal{I} \subseteq \mathcal{D}$, the observationally equivalent set of classes to θ is defined as*

$$F_\theta(\mathcal{I}) = \{\theta_i \in \Theta \mid D_{KL}(\ell_\mathcal{I}(\cdot|\theta_i)||\ell_\mathcal{I}(\cdot|\theta)) = 0\}, \tag{3}$$

where $D_{KL}(\ell_\mathcal{I}(\cdot|\theta_i)||\ell_\mathcal{I}(\cdot|\theta))$ is the Kullback-Leibler divergence measure $\ell_\mathcal{I}(\cdot|\theta_i)$ and $\ell_\mathcal{I}(\cdot|\theta)$.

From the definition above, we have $\theta \in F_\theta(\mathcal{I})$ for all $\theta \in \Theta$ and all for $\mathcal{I} \subseteq \mathcal{D}$. We can write the set $F_\theta(\mathcal{I})$ equivalently as

$$F_\theta(\mathcal{I}) = \{\theta_i \in \Theta : \ell_\mathcal{I}(o_\mathcal{I}|\theta_i) = \ell_\mathcal{I}(o_\mathcal{I}|\theta), \forall o_\mathcal{I} \in \mathcal{O}_\mathcal{I}\}, \tag{4}$$

where $\mathcal{O}_\mathcal{I} = O_{s_1} \times \ldots \times O_{s_k}$ is the joint observation space of the information set $\mathcal{I}$. In other words, $F_\theta(\mathcal{I})$ is the set of hypotheses (or classes) that cannot be distinguished from θ based on the observations obtained by the information sources in $\mathcal{I}$. Furthermore, by Assumption 1 and Equation (4), we have the following (see Section 2 in Ye et al. (2021)):

$$F_\theta(\mathcal{I}) = \bigcap_{s_i \in \mathcal{I}} F_\theta(s_i), \forall \mathcal{I} \in \mathcal{D}, \forall \theta \in \Theta. \tag{5}$$

Define $F_\theta(\emptyset) = \Theta$, i.e., when there is no information set, all classes are observationally equivalent.

At time t, the central designer predicts the state of the world based on the belief $\mu_t^\mathcal{I}$ generated by the information set $\mathcal{I}$. In order to characterize the learning performance, we consider a penalty-based classification framework. Let $\Xi = [\xi_{ij}] \in \mathbb{R}^{m \times m}$ denote the *penalty matrix*, where $0 \leq \xi_{ij} \leq 1$ is the penalty associated with predicting the class to be θ_j, given that the true class is θ_i. The penalty matrix is assumed to be *row stochastic*, i.e., $\sum_{j=1}^m \xi_{ij} = 1$. We have $\xi_{ii} = 0,\ \forall i \in \{1, 2, \ldots, m\}$, i.e., there is no penalty when the predicted hypothesis is the true hypothesis. Similar to analyses presented in Nedić et al. (2017) and Mitra et al. (2020), we present finite sample convergence rates for the Bayesian belief over the set of hypotheses. In this paper, we consider the case of a uniform prior, but the results can be extended to non-uniform priors (as in Lemma 1 of Mitra et al. (2020)). We defer all the proofs to Appendix A of the extended version of this paper (Bhargav et al. (2024)).

Theorem 2 *Let the true state of the world be θ_p and let $\mu_0(\theta) = \frac{1}{m}\ \forall \theta \in \Theta$ (i.e., uniform prior). Under Assumption 1, for any $\delta, \epsilon \in [0, 1]$, and L as defined in Equation (1), and for an information set $\mathcal{I} \subseteq \mathcal{D}$, the Bayesian update rule in Equation (2) has the following property: there is an integer $N(\delta, \epsilon, L)$, such that with probability at least $1 - \delta$, for all $t > N(\delta, \epsilon, L)$ we have:*

(a) $\mu_t^\mathcal{I}(\theta_q) = \mu_t^\mathcal{I}(\theta_p)\ \forall \theta_q \in F_{\theta_p}(\mathcal{I})$, *and*

(b) $\mu_t^\mathcal{I}(\theta_q) \leq \exp\left(-t(|K(\theta_p, \theta_q) - \epsilon|)\right)\ \forall \theta_q \notin F_{\theta_p}(\mathcal{I})$;

where $K(\theta_p, \theta_q) = D_{KL}(\ell_\mathcal{I}(\cdot|\theta_p)||\ell_\mathcal{I}(\cdot|\theta_q))$ is the Kullback-Leibler divergence measure between the likelihood functions $\ell_\mathcal{I}(\cdot|\theta_p)$ and $\ell_\mathcal{I}(\cdot|\theta_q)$, $F_{\theta_p}(\mathcal{I})$ is defined in (4), and $N(\delta, \epsilon, L) = \left\lceil \frac{2L^2}{\epsilon^2} \log \frac{2}{\delta} \right\rceil$.

We consider a belief threshold rule in order to rule out hypotheses that do not have a high likelihood of being predicted as the true hypothesis. Let μ_{th} be the threshold chosen by the central designer. Corollary 3 presents the sample complexity for the observations in order to ensure that the beliefs over the states $\theta_q \notin F_{\theta_p}(\mathcal{I})$ remain bounded under the specified threshold.

Corollary 3 *Instate the hypothesis and notation of Theorem 2. For a specified threshold $\mu_{th} \in (0,1)$ for the belief over any class $\theta_q \notin F_{\theta_p}(\mathcal{I})$, there exists $\delta, \epsilon \in [0,1]$, for which one can guarantee with probability at least $1-\delta$ that $\mu_t^{\mathcal{I}}(\theta_q) \leq \mu_{th}$ for all $\theta_q \notin F_{\theta_p}$ and for all $t > \tilde{N}$, where*

$$\tilde{N} = \left\lceil \max \left\{ \frac{2L^2}{\epsilon^2} \log \frac{2}{\delta}, \frac{1}{\min\limits_{\theta_p, \theta_q \in \Theta} |K(\theta_p, \theta_q) - \epsilon|} \log \frac{1}{\mu_{th}} \right\} \right\rceil . \tag{6}$$

From Corollary 3, we have the following: After any $t > \tilde{N}$, the central node will predict one of $\theta_q \in F_{\theta_p}(\mathcal{I})$ to be the true hypothesis, with probability at least $1-\delta$. Therefore, it is sufficient to consider the penalties associated with the states $\theta_q \in F_{\theta_p}(\mathcal{I})$ for finding the maximum penalty. We now formalize the Minimum-Cost Information Set Selection (MCIS) Problem as follows:

Problem 1 (MCIS) *Consider a set $\Theta = \{\theta_1, \ldots, \theta_m\}$ of possible states of the world, a set $\mathcal{D}$ of information sources, a selection cost $c_i \in \mathbb{R}_{>0}$ of each source $i \in \mathcal{D}$, a row-stochastic penalty matrix $\Xi = [\xi_{ij}] \in \mathbb{R}^{m \times m}$, and prescribed penalty bounds $0 \leq R_{\theta_p} \leq 1$ for all $\theta_p \in \Theta$. The MCIS Problem is to find a set of selected information sources $\mathcal{I} \subseteq \mathcal{D}$ that solves*

$$\min_{\mathcal{I} \subseteq \mathcal{D}} \; c(\mathcal{I}); \;\; \textit{s.t.} \max_{\theta_i \in F_{\theta_p}(\mathcal{I})} \xi_{pi} \leq R_{\theta_p} \quad \forall \theta_p \in \Theta. \tag{7}$$

2.1. Weak Submodularity and Greedy Algorithm

The combinatorial optimization in (7) can be shown to be NP-hard (based on similar arguments as in Theorem 3 of Ye et al. (2021)). In this section, we propose a greedy algorithm with performance guarantees to efficiently approximate the solution to the MCIS Problem. We first begin by transforming the MCIS problem into the minimum cost set cover problem studied in Wolsey (1982).

Definition 4 (Monotonicity) *A set function $f : 2^{\Omega} \to \mathbb{R}$ is monotone non-decreasing if $f(X) \leq f(Y)$ for all $X \subseteq Y \subseteq \Omega$ and monotone non-increasing if $f(X) \geq f(Y)$ for all $X \subseteq Y \subseteq \Omega$.*

Definition 5 (Submodularity Ratio) [3] *Given a set Ω, the submodularity ratio of a non-negative function $f : 2^{\Omega} \to \mathbb{R}_{\geq 0}$ is the largest $\gamma \in \mathbb{R}$ that satisfies for all $A, B \subseteq \Omega$, the following:*

$$\sum_{a \in A \setminus B} (f(\{a\} \cup B) - f(B)) \geq \gamma (f(A \cup B) - f(B)).$$

Remark 6 *For a non-negative and non-decreasing function $f(\cdot)$ with submodularity ratio γ, we have $\gamma \in [0,1]$. If γ is closer to 1, the function is closer to being submodular. $f(\cdot)$ is submodular if and only if $\gamma \geq 1$. Das and Kempe (2018) provide guarantees for greedy optimization of weak submodular functions, which depend on the submodularity ratio γ. Thus, in order to characterize the performance of greedy, one has to give a (non-zero) lower bound on γ.*

The constraint in (7) can be equivalently written as: $1 - \max_{\theta_i \in F_{\theta_p}(\mathcal{I})} \xi_{pi} \geq 1 - R_{\theta_p}, \; \forall \theta_p \in \Theta$. For all $\mathcal{I} \subseteq \mathcal{D}$ and for a true state $\theta_p \in \Theta$, let us define $f_{\theta_p}(\mathcal{I}) = 1 - \max_{\theta_i \in F_{\theta_p}(\mathcal{I})} \xi_{pi}$. It follows from (5) that $f_{\theta_p}(\cdot)$ is a monotone non-decreasing set function with $f_{\theta_p}(\emptyset) = 1 - \max_{\theta_j \in \Theta} \xi_{pj}$.

3. There are several notions of submodularity ratio. We consider $\gamma_{U,k}$ as defined in Das and Kempe (2018), where U is the universal set and $k \geq 1$ is a parameter, and drop the dependence on k by defining $\gamma = \min_k \gamma_{U,k}$.

In order to establish the approximate submodularity property, we make the following assumption on the misclassification penalties.

Assumption 2: The misclassification penalties are unique, i.e., $\xi_{pi} \neq \xi_{pj}$ for all $i \neq j, \forall \theta_p \in \Theta$.

Note that the above assumption requires that no two misclassification events have the same penalty associated with them, which is often a reasonable assumption in many applications.

Lemma 7 *Under Assumption 2, the function $f_{\theta_p}(\mathcal{I}) : 2^{\mathcal{D}} \rightarrow \mathbb{R}_{\geq 0}$ is approximately submodular for all $\theta_p \in \Theta$, with a submodularity ratio $\gamma = \underline{\xi}/\bar{\xi}$, where*

$$\underline{\xi} = \min_{\theta_p \in \Theta} \left(\min_{\theta_i, \theta_j \in \Theta} |\xi_{pi} - \xi_{pj}| \right); \quad \bar{\xi} = \max_{\theta_p \in \Theta} \left(\max_{\theta_i, \theta_j \in \Theta} |\xi_{pi} - \xi_{pj}| \right). \tag{8}$$

In order to ensure that there exists a feasible solution $\mathcal{I} \subseteq \mathcal{D}$ that satisfies the constraints, we assume that $f_{\theta_p}(\mathcal{D}) \geq 1 - R_{\theta_p}$ for all $\theta_p \in \Theta$. For any $\mathcal{I} \subseteq \mathcal{D}$, we define $f'_{\theta_p}(\mathcal{I}) = \min\{f_{\theta_p}(\mathcal{I}), 1 - R_{\theta_p}\} \quad \forall \theta_p \in \Theta$. The function $f'_{\theta_p}(\mathcal{I})$ captures the sufficient condition for satisfying the penalty constraints corresponding to each state. We now define, for all $\mathcal{I} \subseteq \mathcal{D}$,

$$z(\mathcal{I}) = \sum_{\theta_p \in \Theta} f'_{\theta_p}(\mathcal{I}) = \sum_{\theta_p \in \Theta} \min\left\{f_{\theta_p}(\mathcal{I}), 1 - R_{\theta_p}\right\}. \tag{9}$$

The expression $z(\mathcal{I})$ combines all the constraints (corresponding to each hypothesis $\theta \in \Theta$), which we wish to satisfy, while selecting the information set. In other words, $z(\cdot)$ is used to find the optimal set $\mathcal{I}$, i.e., a set $\mathcal{I} \subseteq \mathcal{D}$ with minimal $c(\mathcal{I})$, satisfying $z(\mathcal{I}) = z(\mathcal{D})$. Since $F_{\theta_p}(\emptyset) = \Theta$, we have $z(\emptyset) = m - \sum_{\theta_p \in \Theta} \max_{\theta_i \in \Theta} \xi_{pi}$. Since $f_{\theta_p}(\mathcal{I})$ is approximately submodular and non-decreasing, we have that $f'_{\theta_p}(\mathcal{I})$ is also approximately submodular and non-decreasing. Noting that the non-negative sum of approximately submodular functions is approximately submodular (Lemma 3.12 of Borodin et al. (2014)), we have that $z(\cdot)$ is also approximately submodular. We have the following result, which follows from the existence of a feasible solution for Problem 1.

Lemma 8 *For any $\mathcal{I} \subseteq \mathcal{D}$, the constraint $1 - \max_{\theta_i \in F_{\theta_p}(\mathcal{I})} \xi_{pi} \geq 1 - R_{\theta_p}$ holds for all $\theta_p \in \Theta$ if and only if $\sum_{\theta_p \in \Theta} f'_{\theta_p}(\mathcal{I}) = \sum_{\theta_p \in \Theta} f'_{\theta_p}(\mathcal{D})$.*

We now have from Lemma 8 that the constraint (7) in Problem 1 can be equivalently written as

$$\min_{\mathcal{I} \subseteq \mathcal{D}} c(\mathcal{I}); \;\; \textit{s.t. } z(\mathcal{I}) = z(\mathcal{D}). \tag{10}$$

Problem (10) can then be viewed as the set covering problem studied in Wolsey (1982). In Das and Kempe (2018), the authors present performance guarantees for the weak submodular version of the set covering problem studied in Wolsey (1982). Based on Theorem 9 in Das and Kempe (2018), we have the following performance guarantees for Algorithm 1 when applied to the MCIS problem.

Theorem 9 *Let $\mathcal{I}^*$ be an optimal solution to the MCIS problem having a submodularity ratio γ. For a specified threshold $\mu_{th} \in (0, 1)$ and $0 \leq \delta \leq 1$, with probability at least $1 - \delta$, Algorithm 1 under $\tilde{N}$ observation samples returns a solution $\mathcal{I}_g$ to the MCIS problem (i.e., (7)) that satisfies the following:*

$$c\left(\mathcal{I}_g\right) \leq \left(1 + \frac{1}{\gamma} \log \frac{z(\mathcal{D}) - z(\emptyset)}{z(\mathcal{D}) - z\left(\mathcal{I}_g^{T-1}\right)}\right) c\left(\mathcal{I}^*\right),$$

where $\tilde{N}$ is specified in (6), and $\mathcal{I}_g^1, \ldots, \mathcal{I}_g^{T-1}$ are specified in Algorithm 1.

Algorithm 1 Greedy Algorithm for MCIS

Input: $\mathcal{D}, z : 2^{\mathcal{D}} \to \mathbb{R}_{\geq 0}, c_i \in \mathbb{R}_{>0}\ \forall i \in \mathcal{D}$
Output: $\mathcal{I}_g$

$k \leftarrow 0, \mathcal{I}_g^0 \leftarrow \emptyset$
while $z(\mathcal{I}_g^t) < z(\mathcal{D})$ **do**
  $j_t \in \arg\max_{i \in \mathcal{D} \setminus \mathcal{I}_g^t} \frac{z(\mathcal{I}_g^t \cup \{i\}) - z(\mathcal{I}_g^t)}{c_i}$
  $\mathcal{I}_g^{t+1} \leftarrow \mathcal{I}_g^t \cup \{j_t\}, k \leftarrow k + 1$
end while
$T \leftarrow k, \mathcal{I}_g \leftarrow \mathcal{I}_g^T$
return $\mathcal{I}_g$

We have the following result characterizing the asymptotic performance of the greedy algorithm.

Corollary 10 *Instate the hypothesis and notation of Theorem 2. As $t \to \infty$, we have the following: (a) $\mu_\infty^{\mathcal{I}}(\theta_q) = 0 \quad \forall \theta_q \notin F_{\theta_p}(\mathcal{I})$, and (b) $\mu_\infty^{\mathcal{I}}(\theta_q) = \frac{1}{|F_{\theta_p}(\mathcal{I})|} \quad \forall \theta_q \in F_{\theta_p}(\mathcal{I})$. The near-optimal guarantees provided in Theorem 9 for Problem 1 hold with probability 1 (a.s.).*

3. Minimum-Penalty Information Set Selection

In this section, we consider the problem where the central designer has a fixed budget for selecting information sources and seeks to minimize the maximum penalty of misclassifying the true state. Since the true state is not known a priori, the central designer has to minimize the maximum penalty for each possible true state, which is a multi-objective optimization problem under a budget constraint. We scalarize the multi-objective optimization into a single-objective optimization problem. The optimal solution to this single-objective problem is a Pareto optimal solution to the multi-objective problem (Hwang and Masud (2012)). We now formalize the Minimum-Penalty Information Set Selection (MPIS) Problem as follows.

Problem 2 (MPIS) *Consider a set $\Theta = \{\theta_1, \ldots, \theta_m\}$ of possible states of the world; a set $\mathcal{D}$ of information sources, with each source $i \in \mathcal{D}$ having a cost $c_i \in \mathbb{R}_{\geq 0}$; a row-stochastic penalty matrix $\Xi = [\xi_{ij}] \in \mathbb{R}^{m \times m}$; and a selection budget $K \in \mathbb{R}_{\geq 0}$. The MPIS Problem is to find a set of selected information sources $\mathcal{I} \subseteq \mathcal{D}$ that solves*

$$\min_{\mathcal{I} \subseteq \mathcal{D}} \sum_{\theta_p \in \Theta} \left(\max_{\theta_j \in F_{\theta_p}(\mathcal{I})} \xi_{pj} \right); \ \text{s.t.} \ \sum_{i \in \mathcal{I}} c_i \leq K. \tag{11}$$

Consider the following equivalent optimization problem:

$$\max_{\mathcal{I} \subseteq \mathcal{D}} \sum_{\theta_p \in \Theta} \left(1 - \max_{\theta_j \in F_{\theta_p}(\mathcal{I})} \xi_{pj} \right); \ \text{s.t.} \ \sum_{i \in \mathcal{I}} c_i \leq K. \tag{12}$$

It is easy to verify that the problem defined in (12) is equivalent to the problem defined in (11), i.e., the information set $\mathcal{I} \subseteq \mathcal{D}$ that optimizes the problem in Equation (12) is also the optimal solution to the Problem 2. We note that $f_{\theta_p}(\mathcal{I}) = 1 - \max_{\theta_j \in F_{\theta_p}(\mathcal{I})} \xi_{pj}$. We denote $\Lambda(\mathcal{I}) = \sum_{\theta_p \in \Theta} f_{\theta_p}(\mathcal{I})$.

Algorithm 2 Greedy Algorithm for MPIS

Input: Data sources: $\mathcal{D}$, Penalties: $\Xi \in \mathbb{R}^{m\times m}$, Selection costs: $c_i \; \forall i \in \mathcal{D}$, Budget: $K \in \mathbb{R}_{>0}$
Output: $\mathcal{I}_K$

$t \leftarrow 0, \mathcal{I}_K \leftarrow \emptyset$
while $t \leq K$ **do**
  $j \leftarrow \arg\max_{i\in\mathcal{D}\setminus\mathcal{I}_K} \frac{\Lambda(\mathcal{I}_K\cup\{i\})-\Lambda(\mathcal{I}_K)}{c_i}$
  $\mathcal{I}_K \leftarrow \mathcal{I}_K \cup \{j\}, t \leftarrow t + c_j$
end while
return $\mathcal{I}_K$

From Lemma 7 and Lemma 3.12 in Borodin et al. (2014), it follows that the objective function in (12) is approximately submodular with the submodularity ratio γ. Based on the guarantees for greedy maximization of monotone, non-decreasing, approximately submodular functions subject to Knapsack constraints in Theorem 6 of Das and Kempe (2018), we have the following result.

Theorem 11 *Let $\mathcal{I}_K \subseteq \mathcal{D}$ denote the information set selected by Algorithm 2 and let $\mathcal{I}_K^* \subseteq \mathcal{D}$ denote the optimal information set for the MPIS Problem with submodularity ratio γ. For a specified threshold $\mu_{th} \in (0,1)$ and $0 \leq \delta \leq 1$, with probability at least $1-\delta$, Algorithm 2 under $\tilde{N}$ observation samples returns a solution $\mathcal{I}_K$ to the MPIS problem (i.e., (11)) that satisfies $\Lambda(\mathcal{I}_K) \geq (1-e^{-\gamma})\,\Lambda(\mathcal{I}_K^*) + c$, where $c = \Lambda(\emptyset)/e^{\gamma}$ and $\tilde{N}$ is specified in (6).*

We have the following result characterizing the asymptotic performance of the greedy algorithm.

Corollary 12 *Instate the hypothesis and notation of Theorem 2. As $t \to \infty$, we have the following: (a) $\mu_\infty^{\mathcal{I}}(\theta_q) = 0 \quad \forall \theta_q \notin F_{\theta_p}(\mathcal{I})$, and (b) $\mu_\infty^{\mathcal{I}}(\theta_q) = \frac{1}{|F_{\theta_p}(\mathcal{I})|} \quad \forall \theta_q \in F_{\theta_p}(\mathcal{I})$. The near-optimal guarantees provided in Theorem 11 for Problem 2 hold with probability 1 (a.s.).*

4. Alternate Penalty Metric for Information Set Selection

In many practical scenarios, the submodularity ratio of the maximum penalty metric may be arbitrarily small (or zero) when misclassification penalties for two hypotheses are very close to each other (or equal) (see Appendix B of the extended version (Bhargav et al. (2024)) for a detailed discussion). It is also easy to verify that the submodularity ratio γ decreases as the number of hypotheses increase. As a result, the performance bounds for the greedy algorithms become weaker. In such scenarios, one can turn to an alternate metric for optimization, which can provide non-trivial guarantees for the performance of the greedy algorithm. To this end, we present an alternate metric to characterize the quality of an information set, based on the total penalty of misclassification, defined as follows:

$$\rho_{\theta_p}(\mathcal{I}) = \sum_{\theta_i \in F_{\theta_p}(\mathcal{I})} \xi_{pi}. \tag{13}$$

Intuitively, in order to minimize the total penalty ($\rho_{\theta_p}(\mathcal{I})$) (or ensure that it is below a desired bound), one has to select a subset $\mathcal{I} \subseteq \mathcal{D}$ that ensures that the number of hypotheses which are observationally equivalent to the true hypothesis θ_p, i.e., $|F_{\theta_p}(\mathcal{I})|$, is small and/or the hypotheses that are observationally equivalent to the true hypothesis have lower misclassification penalties. Effectively, this results in lower penalty associated with misclassifying the true hypothesis.

We define the Modified Minimum Cost Information Set Selection (M-MCIS) and Modified Minimum Penalty Information Set Selection (M-MPIS) Problems based on this metric as follows.

Problem 3 (M-MCIS) *Consider a set $\Theta = \{\theta_1, \dots, \theta_m\}$ of possible states of the world, a set $\mathcal{D}$ of information sources, a selection cost $c_i \in \mathbb{R}_{>0}$ of each source $i \in \mathcal{D}$, a row-stochastic penalty matrix $\Xi = [\xi_{ij}] \in \mathbb{R}^{m \times m}$, and prescribed penalty bounds $0 \leq R'_{\theta_p} \leq 1$ for all $\theta_p \in \Theta$. The M-MCIS Problem is to find a set of selected information sources $\mathcal{I} \subseteq \mathcal{D}$ that solves*

$$\min_{\mathcal{I} \subseteq \mathcal{D}} \; c(\mathcal{I}); \;\; \textit{s.t.} \; \rho_{\theta_p}(\mathcal{I}) \leq R'_{\theta_p} \quad \forall \theta_p \in \Theta. \tag{14}$$

Note that the penalty bounds R_{θ_p} of the MCIS Problem (Problem 1) differ from the bounds R'_{θ_p} of M-MCIS Problem (Problem 3), as the former is a bound on the maximum penalty, while the latter is a bound on the total penalty. The designer can choose the bounds R'_{θ_p} in order to achieve the desired classification performance.

Problem 4 (M-MPIS) *Consider a set $\Theta = \{\theta_1, \dots, \theta_m\}$ of possible states of the world; a set $\mathcal{D}$ of information sources, with each source $i \in \mathcal{D}$ having a cost $c_i \in \mathbb{R}_{\geq 0}$; a row-stochastic penalty matrix $\Xi = [\xi_{ij}] \in \mathbb{R}^{m \times m}$; and a selection budget $K \in \mathbb{R}_{\geq 0}$. The M-MPIS Problem is to find a set of selected information sources $\mathcal{I} \subseteq \mathcal{D}$ that solves*

$$\min_{\mathcal{I} \subseteq \mathcal{D}} \; \sum_{\theta_p \in \Theta} \rho_{\theta_p}(\mathcal{I}); \;\; \textit{s.t.} \; \sum_{i \in \mathcal{I}} c_i \leq K. \tag{15}$$

Lemma 13 *The function $g_{\theta_p}(\mathcal{I}) = 1 - \rho_{\theta_p}(\mathcal{I}) : 2^{\mathcal{D}} \to \mathbb{R}_{\geq 0}$ is submodular for all $\theta_p \in \Theta$.*

By Lemma 13, we have the following result characterizing the performance of the greedy algorithms for the modified information set selection problems.

Corollary 14 *For Algorithm 1 (respectively Algorithm 2) applied to M-MCIS (respectively M-MPIS) Problem, the near-optimal guarantees provided in Theorem 9 (respectively Theorem 11) hold with $\gamma = 1$.*

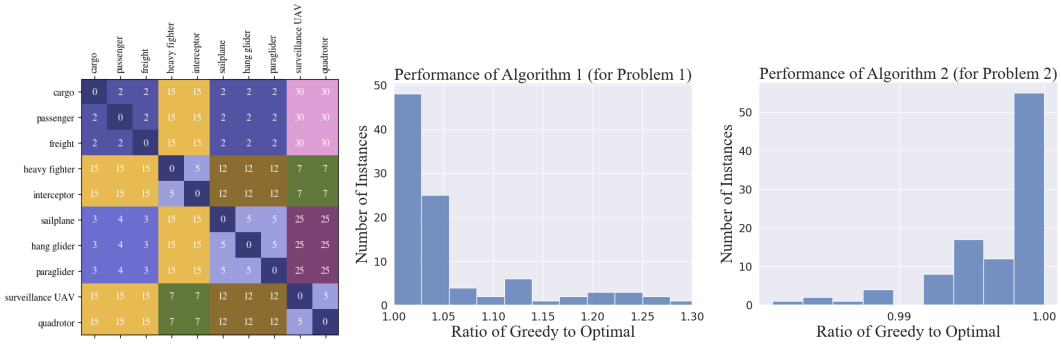
From Corollary 14, we have that the total penalty metric enjoys stronger near-optimal guarantees (due to submodularity) compared to that of the maximum penalty metric (which is weak submodular) for greedy optimization. Moreover, the near-optimal guarantees for the M-MCIS and M-MPIS problems are independent of the misclassification penalties and the number of hypotheses.

5. Empirical Evaluation

In this section, we validate the theoretical results through numerical simulations. We present simulations for varying submodularity ratios, finite sample convergence of the beliefs and the modified information selection problems in the extended version (Bhargav et al. (2024)) (see Appendix C).

We consider a hypothesis testing task where one has to identify (or classify) an aerial vehicle into one of the following 10 classes: $\Theta = \{$*cargo, passenger, freight, heavy fighter, interceptor, sailplane, hang glider, paraglider, surveillance UAV, quadrotor*$\}$. We will refer to this as the Aerial Vehicle Classification task (AVC task). The penalty matrix is as shown in Figure 1 (a). Each row of the penalty matrix is normalized. We set $|\mathcal{D}| = 10$, the costs c_i for $i \in \mathcal{D}$

are sampled uniformly from $\{1, \ldots, 10\}$. We consider the infinite-observation case and randomly generate the observationally equivalent sets $F_{\theta_p}(i)$ for each $\theta_p \in \Theta$ and $i \in \mathcal{D}$. We first consider the minimum cost information set selection problem for the AVC task. The thresholds R_{θ_p} for $\theta_p \in \{$*cargo, passenger, freight, sailplane, hang glider, paraglider*$\}$ are randomly sampled from $[0.7, 1]$ and for $\theta_p \in \{$*heavy fighter, interceptor, surveillance UAV, quadrotor*$\}$ are randomly sampled from $[0.1, 0.4]$. For 100 randomly generated instances, we run Algorithm 1 to find the greedy information set $\mathcal{I}_g$ and find the optimal information set $\mathcal{I}^*$ using brute-force search. We plot the ratio of cost of the greedy information set to that of the optimal, i.e., $c(\mathcal{I}_g)/c(\mathcal{I}^*)$, in Figure 1 (b).



(*a*) (*b*) (*c*)

Figure 1: (a) Penalty Matrix for the Aerial Vehicle Classification (AVC) task, (b) Performance of Algorithm 1 (for Problem 1), (c) Performance of Algorithm 2 (for Problem 2).

Next, we consider the minimum penalty information set selection problem for the AVC task. We generate 100 random instances with varying information source costs and selection budgets. We run Algorithm 2 to find the greedy information set $\mathcal{I}_g$ and find the optimal information set $\mathcal{I}^*$ using brute-force search. We plot the ratio of greedy utility to that of the optimal, i.e., $\Lambda(\mathcal{I}_g)/\Lambda(\mathcal{I}^*)$, in Figure 1(c). These plots show the near-optimal performance of the greedy algorithm. Note that the penalty matrix for these instances does not satisfy Assumption 2 (uniqueness of misclassification penalties). Thus, these problem instances are not guaranteed to exhibit the weak submodularity property. Despite this, we observe that the greedy algorithms provide near-optimal performance.

6. Conclusion

In this work, we studied two variants of an information set selection problem for hypothesis testing: (i) selecting a minimum cost information set to ensure that the maximum penalty for misclassifying the true hypothesis is below a desired bound and (ii) optimal information set selection under a limited budget to minimize the maximum penalty of misclassifying the true hypothesis. Leveraging the weak submodularity property of the performance metric, we established high-probability guarantees for greedy algorithms for both problems, along with the associated finite sample convergence rates for the Bayesian beliefs. Next, we proposed an alternate metric based on the total penalty of misclassification for information set selection, which enjoys (stronger) near-optimal performance guarantees with high-probability for the greedy algorithms. Finally, we evaluated the empirical performance of the proposed greedy algorithms over several randomly generated problem instances.

Acknowledgments

This material is based upon work supported by the Office of Naval Research (ONR) and Saab, Inc. under the Threat and Situational Understanding of Networked Online Machine Intelligence (TSUNOMI) program (grant no. N00014-23-C-1016). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the ONR and/or Saab, Inc.

References

Niall M Adams and David J Hand. Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, 32(7):1139–1147, 1999.

Jayanth Bhargav, Mahsa Ghasemi, and Shreyas Sundaram. On the Complexity and Approximability of Optimal Sensor Selection for Mixed-Observable Markov Decision Processes. In *2023 American Control Conference (ACC)*, pages 3332–3337. IEEE, 2023.

Jayanth Bhargav, Mahsa Ghasemi, and Shreyas Sundaram. Submodular Information Selection for Hypothesis Testing with Misclassification Penalties. *arXiv preprint arXiv:2405.10930*, 2024.

Allan Borodin, Dai Tri Man Le, and Yuli Ye. Weakly submodular functions. *arXiv preprint arXiv:1401.6697*, 2014.

Yuxin Chen, S Hamed Hassani, Amin Karbasi, and Andreas Krause. Sequential information maximization: When is greedy near-optimal? In *Conference on Learning Theory, 338-363*, 2015.

Sundeep Prabhakar Chepuri and Geert Leus. Sparsity-promoting sensor selection for non-linear measurement models. *IEEE Transactions on Signal Processing*, 63(3):684–698, 2014.

Abhimanyu Das and David Kempe. Approximate submodularity and its applications: Subset selection, sparse approximation and dictionary selection. *Journal of Machine Learning Research*, 19 (3):1–34, 2018.

Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

Mahsa Ghasemi and Ufuk Topcu. Online active perception for partially observable markov decision processes with limited budget. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 6169–6174. IEEE, 2019.

Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal Bayesian active learning with noisy observations. In *Proc. Advances in Neural Information Processing Systems*, pages 766–774, 2010.

Abolfazl Hashemi, Mahsa Ghasemi, Haris Vikalo, and Ufuk Topcu. Randomized greedy sensor selection: Leveraging weak submodularity. *IEEE Transactions on Automatic Control*, 66(1): 199–212, 2020.

Geoffrey A Hollinger and Gaurav S Sukhatme. Sampling-based motion planning for robotic information gathering. In *Robotics: Science and Systems*, volume 3, pages 1–8, 2013.

Yi Hou, Praveen Edara, and Carlos Sun. Modeling mandatory lane changing using bayes classifier and decision trees. *IEEE Transactions on Intelligent Transportation Systems, 647-655*, 2013.

C-L Hwang and Abu Syed Md Masud. *Multiple objective decision making — Methods and Applications: A state-of-the-art survey*, volume 164. Springer Science & Business Media, 2012.

Ali Jadbabaie, Pooya Molavi, Alvaro Sandroni, and Alireza Tahbaz-Salehi. Non-Bayesian social learning. *Games and Economic Behavior*, 76(1):210–225, 2012.

Andreas Krause and Volkan Cevher. Submodular dictionary selection for sparse representation. In *Proc. International Conference on Machine Learning*, pages 567–574, 2010.

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*, 9(Feb):235–284, 2008.

Anusha Lalitha, Anand Sarwate, and Tara Javidi. Social learning and distributed hypothesis testing. In *Proc. IEEE International Symposium on Information Theory*, pages 551–555, 2014.

Qipeng Liu, Aili Fang, Lin Wang, and Xiaofan Wang. Social learning with time-varying weights. *Journal of Systems Science and Complexity*, 27(3):581–593, 2014.

A. Mitra, J. A. Richards, and S. Sundaram. A new approach to distributed hypothesis testing and non-Bayesian learning: Improved learning rate and Byzantine-resilience. *IEEE Transactions on Automatic Control*, 2020.

Yilin Mo, Roberto Ambrosino, and Bruno Sinopoli. Sensor selection strategies for state estimation in energy constrained wireless sensor networks. *Automatica*, 47(7):1330–1338, 2011.

Angelia Nedić, Alex Olshevsky, and César A Uribe. Fast convergence rates for distributed non-Bayesian learning. *IEEE Transactions on Automatic Control*, 62(11):5538–5553, 2017.

NHTSA. PE16-007 Technical Report: Tesla crash report. *U.S. Department of Transportation*, 2016.

Parag C Pendharkar. Bayesian posterior misclassification error risk distributions for ensemble classifiers. *Engineering Applications of Artificial Intelligence*, 65:484–492, 2017.

Murat Sensoy, Maryam Saleki, Simon Julier, Reyhan Aydogan, and John Reid. Misclassification risk and uncertainty quantification in deep classifiers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2484–2492, 2021.

Laurence A Wolsey. An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica*, 2(4):385–393, 1982.

L. Ye, N. Woodford, S. Roy, and S. Sundaram. On the complexity and approximability of optimal sensor selection and attack for Kalman filtering. *IEEE Transactions on Automatic Control*, 2020.

Lintao Ye, Aritra Mitra, and Shreyas Sundaram. Near-optimal data source selection for Bayesian learning. In *Learning for Dynamics and Control*, pages 854–865. PMLR, 2021.