Proceedings of Machine Learning Research vol vvv:1–16, 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, 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 aims to reduce the likelihood of acting on erroneous predictions. In a centralized Bayesian learning setting, we study the problem of selecting a minimum cost information set, while ensuring that the maximum penalty of misclassifying the true state is bounded. We prove that this combinatorial optimization problem is submodular, and establish high-probability guarantees for near-optimal performance of a standard greedy algorithm, with the associated finite sample convergence rates for the Bayesian beliefs. Next, we consider the dual problem, where one seeks to select a set of information sources under a limited budget, while trying to minimize the maximum penalty for misclassifying the true state, under finite observation samples from the sources. We prove that this problem is submodular and establish high-probability near-optimal guarantees for a standard greedy algorithm. Finally, we validate our theoretical results through some numerical simulations, and show that the greedy algorithm works well in practice.

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 resources or compute power, one can only query (and process) 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 a variety of 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 certain learning performance (e.g., minimal misclassification penalty) 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 an asymmetric 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 utmost importance. One many have to pay a penalty for misclassifying the true state as another state, for instance, misclassifying a drone (a potential 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 has 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 evidential deep learning classifier to reduce the costs associated with misclassification errors, and present empirical results to show the effectiveness of their algorithm. In Elkan (2001), the authors propose a cost-sensitive learning framework for balancing the optimal number of samples for each class, in order to improve the quality of predictions. In this paper, we study the problem of selecting a minimal cost information set which ensures minimal misclassification penalty, in a finite sample 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*.

There is a substantial body of work focused on the study of submodular optimization and greedy techniques for feature selection in sparse learning (Krause and Cevher (2010); Chepuri and Leus (2014)); sensor selection for estimation in networks (Mo et al. (2011)), Kalman filtering of linear systems (Ye et al. (2020)), and mixed-observable Markov decision processes (Bhargav et al. (2023)).

The closest paper to our work is Ye et al. (2021), in which the authors studied minimal-cost data source selection for Bayesian learning, where the goal was to ensure that the total variation error between the asymptotic belief and the true state distribution is bounded. However, we consider a non-asymptotic setting, where the goal is to ensure that the maximum penalty of misclassifying the true state remains bounded. Building upon the results in Ye et al. (2021), we establish theoretical guarantees for greedy information selection to ensure misclassification penalties are bounded.

1.2. Contributions

First, we consider the problem where a central designer has to select a minimal cost information subset while ensuring the maximum misclassification penalty for the true state is bounded. We prove that this combinatorial optimization problem is submodular, and present a standard greedy algorithm with high-probability near-optimal performance guarantees, along with the finite sample convergence rates for the Bayesian beliefs. Next, we consider the dual problem, where the central designer can only select a subset of information sources due to certain budget constraints and aims to minimize the maximum penalty of misclassification for the true state. We prove that this problem is submodular and establish high-probability near-optimal guarantees for a standard greedy algorithm.

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 classes (hypotheses), of which one of the hypothesis 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 $\ell(\cdot|\theta)$. Similar to Ye et al. (2021); Liu et al. (2014) and Lalitha et al. (2014), we make the following assumption.

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 $\theta \in \Theta$, the observations are independent of each other conditioned on the true state and over time, i.e., $\{o_{i,1}, o_{i,2}, \ldots\}$ is a sequence of independent identically distributed (i.i.d.) random variables.*

It directly follows that the joint likelihood function satisfies $\ell(\cdot|\theta) = \Pi_{i=1}^{n} \ell_i(\cdot|\theta)$ for all $\theta \in \Theta$.

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, using the observation samples from the selected sources. Each information source $i \in \mathcal{D}$ is assumed to have a selection cost $c_i \geq \mathbb{R}_{>0}$. 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 $\mathcal{I}$ denoted as $c(\mathcal{I})$ is the sum of the costs of the selected sources, i.e., $c(\mathcal{I}) = \sum_{s_i \in \mathcal{I}} c_{s_i}$. The joint observation profile 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 likelihood function $\ell_{\mathcal{I}}(\cdot|\theta) = \Pi_{i=1}^{k} \ell_{s_i}(\cdot|\theta)$. By Assumption 1, the central designer knows $\ell_{\mathcal{I}}(\cdot|\theta)$ for all $\mathcal{I} \subseteq \mathcal{D}$ and for all $\theta \in \Theta$.

Let $\theta_p \in \Theta$ be the true state of the world. We define a probability tuple $\left(\Omega, \mathcal{F}, \mathbb{P}^{\theta_p}\right)$, where $\Omega = \{\omega : \omega = (o_1, o_2, \ldots), o_t \in \mathcal{O}, t \in \mathbb{N}_+\}$, $\mathcal{F}$ is the σ-algebra generated by the observation likelihood functions, and $\mathbb{P}^{\theta_p}$ is the probability measure induced by sample space Ω. Specifically, $\mathbb{P}^{\theta_p} = \prod_{t=1}^{\infty} \ell\left(\cdot \mid \theta_p\right)$. For the sake of brevity, we will say that an event occurs almost surely (a.s.) to mean that it occurs almost surely with respect to the probability measure $\mathbb{P}^{\theta_p}$.

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

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

The constant L is a measure of the maximum log-likelihood ratio of over the possible set of observations and hypotheses, which we will use later in our analyses. The central node maintains a belief over the set of possible hypotheses. The central node updates this belief using the observations obtained from the selected information sources 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 class at time step t, 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

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

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

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

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

Since $D_{KL}(\ell_{\mathcal{I}}(\cdot|\theta)||\ell_{\mathcal{I}}(\cdot|\theta)) = 0$ and $D_{KL}(\cdot)$ is always positive, we have $\theta \in F_\theta(\mathcal{I})$ for all $\theta \in \Theta$ and for all $\mathcal{I} \subseteq \mathcal{D}$. 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 states (or classes) that cannot be distinguished based on the observations obtained by the information sources in $\mathcal{I}$. Furthermore, by Assumption 1 and Equation (4), we have:

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

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 true state of the world (assumed to be θ_p) based on the belief $\mu^{\mathcal{I}}_{t,\theta_p}$ generated by the information set $\mathcal{I}$. In order to characterize the learning performance, we consider a penalty-based classification framework. We define the *penalty matrix* $\Xi = [\xi_{ij}] \in \mathbb{R}^{m\times m}$, 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., $\sum_{j=1}^m \xi_{ij} = 1$ and $\xi_{ii} = 0,\ \forall i \in \{1, 2, \ldots, m\}$. Note that $\xi_{ii} = 0$ means that there is no penalty when the predicted class is the true class. We wish to keep the maximum penalty of misclassifying the true class bounded under a specified threshold. Since the Bayesian belief may contain non-zero probabilities for all states $\theta \in \Theta$, we consider a belief threshold rule in order to rule out classes that do not have a high likelihood of being predicted as the true class. In Nedić et al. (2017), the authors present finite sample convergence rates for agents' beliefs, in a non-Bayesian social learning setting, where a group of agents aim to collectively agree on a hypothesis (which may not be the true hypothesis). Here, we present finite sample convergence rates for the belief that the central node holds over the set of hypotheses, using the Bayesian update rule in (2).. We consider the case of uniform-prior, but the results can be easily extended to non-uniform priors.

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 $\delta \in [0, 1]$, $0 < \epsilon < 1$ 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) $\mu^{\mathcal{I}}_t(\theta_q) = \mu^{\mathcal{I}}_t(\theta_p)\ \ \forall \theta_q \in F_{\theta_p}(\mathcal{I})$, *and*

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

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 $N(\delta, \epsilon, L) = \left\lceil \frac{2L^2}{\epsilon^2} \log \frac{1}{\delta} \right\rceil$.

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 a $\delta \in (0,1)$ and $\epsilon > 0$, 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*

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

Remark 4 *The number of samples required to ensure that the belief over the states $\theta_q \notin F_{\theta_p}(\mathcal{I})$ remains bounded by μ_{th} may be larger than $N(\delta,\epsilon,L)$, if μ_{th} is smaller than the bound presented in Theorem 2. Equation (6) obtained using the bounds in Theorem 2 and μ_{th}, captures this fact.*

From Corollary 3, we have the following: With probability at least $1-\delta$, after any $t > \tilde{N}$, the central node will predict one of the hypothesis $\theta_q \in F_{\theta_p}(\mathcal{I})$ to be the true state of the world. 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 of misclassification. 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*

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

2.1. Submodularity and Greedy Algorithm

In this section, we first describe the complexity of Problem 1 and then, leveraging the idea of submodularity, we provide a greedy algorithm with high-probability near-optimal guarantees. Similar to the reduction to the well-known Set Cover problem (Wolsey (1982)) presented in Ye et al. (2021), the combinatorial optimization in (10) can be shown to be NP-hard. Greedy algorithms, which iteratively and myopically select items with the highest immediate benefit, have been extensively studied in the literature for approximating NP-hard problems. For maximizing monotone submodular functions, the greedy algorithm enjoys near-optimal performance guarantees (Nemhauser et al. (1978); Wolsey (1982)). We show that the MCIS Problem is submodular in the set of information sources selected by transforming it into an instance of the submodular set covering problem studied in Wolsey (1982). We then present a greedy algorithm for minimum cost information set selection with performance guarantees. We begin with the following definitions.

Definition 5 (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 6 (Submodular Set Function) *(Nemhauser et al. (1978)) A set function $f : 2^\Omega \to \mathbb{R}$ is submodular if it satisfies $f(X\cup\{j\}) - f(X) \geq f(Y\cup\{j\}) - f(Y)$, $\forall X \subseteq Y \subseteq \Omega$ and $\forall j \in \Omega\setminus Y$.*

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 $g_{\theta_p}(\mathcal{I}) = \max_{\theta_j \in F_{\theta_p}(\mathcal{I})} \xi_{p,j}$. It is easy to verify that $g_{\theta_p}(\mathcal{I})$ is a monotone *non-increasing* set function with $g_{\theta_p}(\emptyset) = \max_{\theta_j \in \Theta} \xi_{pj}$.

Lemma 7 *The function* $\left(1 - g_{\theta_p}(\mathcal{I})\right) : 2^{\mathcal{D}} \to \mathbb{R}_{\geq 0}$ *is submodular for all* $\theta_p \in \Theta$.

For all $\mathcal{I} \subseteq \mathcal{D}$, let us define

$$f_{\theta_p}(\mathcal{I}) = 1 - g_{\theta_p}(\mathcal{I}) = 1 - \max_{\theta_i \in F_{\theta_p}(\mathcal{I})} \xi_{pi},\ \forall \theta_p \in \Theta. \tag{8}$$

It follows from (8) and (5) that $f_{\theta_p} : 2^{\mathcal{D}} \to \mathbb{R}_{\geq 0}$ is a monotone non-decreasing set function. 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

$$f^{'}_{\theta_p}(\mathcal{I}) = \min\{f_{\theta_p}(\mathcal{I}), 1 - R_{\theta_p}\} \quad \forall \theta_p \in \Theta. \tag{9}$$

The function $f^{'}_{\theta_p}(\mathcal{I})$ captures the sufficient condition for the penalty constraints corresponding to each state to be satisfied.

We now define, for all $\mathcal{I} \subseteq \mathcal{D}$,

$$z(\mathcal{I}) = \sum_{\theta_p \in \Theta} f'_{\theta_p}(\mathcal{I}) = \sum_{\theta_p \in \Theta} \min\left\{f_{\theta_p}(\mathcal{I}), 1 - R_{\theta_p}\right\}, \tag{10}$$

where $f_{\theta_p}(\cdot)$ is defined in (8). 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 $\mathcal{I}$, i.e., an information 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 submodular and non-decreasing with $f_{\theta_p}(\emptyset) = 0$ and $f_{\theta_p}(\mathcal{D}) \geq 1 - R_{\theta_p}$, it is easy to show that $f'_{\theta_p}(\mathcal{I})$ is also submodular and non-decreasing with $f'_{\theta_p}(\emptyset) = 0$ and $f'_{\theta_p}(\mathcal{D}) = 1 - R_{\theta_p}$. Noting that the sum of submodular functions remains submodular, we have that $z(\cdot)$ is submodular and non-decreasing.

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* $\sum_{\theta_p \in \Theta} f'_{\theta_p}(\mathcal{I}) = \sum_{\theta_p \in \Theta} f'_{\theta_p}(\mathcal{D})$*, where* $f'_{\theta_p}(\cdot)$ *is as defined in (9).*

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

$$\begin{aligned} &\min_{\mathcal{I} \subseteq \mathcal{D}} c(\mathcal{I}) \\ &\text{s.t. } z(\mathcal{I}) = z(\mathcal{D}), \end{aligned} \tag{11}$$

where $z(\cdot)$ is a monotone (non-decreasing) submodular set function with $z(\emptyset) = 0$.

Problem (11) can then be viewed as the submodular set covering problem studied in Wolsey (1982), where the submodular set covering problem is solved using a greedy algorithm (Algorithm 1) with performance guarantees. In each iteration of Algorithm 1, the information set (which has

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

not been selected previously) having the highest gain to cost ratio is selected until the constraints encoded in the $z(\cdot)$ function (as in Equation (10)) are satisfied. The algorithm maintains a sequence of sets $\mathcal{I}_g^0, \mathcal{I}_g^1, \ldots, \mathcal{I}_g^T$ containing the selected elements from $\mathcal{D}$, where $T \in \mathbb{Z}_{\geq 1}$. Based on the results presented in Wolsey (1982) and Ye et al. (2021), we have the following result, which characterizes the performance guarantees for Algorithm 1 when applied to the MCIS problem.

Theorem 9 *Let $\mathcal{I}^*$ be an optimal solution to the MCIS problem. 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, where $\tilde{N}$ is specified in (6), and $\mathcal{I}_g^1, \ldots, \mathcal{I}_g^{T-1}$ are specified in Algorithm 1:*

(a) $c(\mathcal{I}_g) \leq \left(1 + \log \max_{i \in \mathcal{D}, \zeta \in [T-1]} \left\{ \frac{z(i) - z(\emptyset)}{z\left(\mathcal{I}_g^\zeta \cup \{i\}\right) - z\left(\mathcal{I}_g^\zeta\right)} : z\left(\mathcal{I}_g^\zeta \cup \{i\}\right) - z\left(\mathcal{I}_g^\zeta\right) > 0 \right\}\right) c(\mathcal{I}^*),$

(b) $c(\mathcal{I}_g) \leq \left(1 + \log \frac{c_{j_T}(z(j_1) - z(\emptyset))}{c_{j_1}\left(z\left(\mathcal{I}_g^{T-1} \cup \{j_T\}\right) - z\left(\mathcal{I}_g^{T-1}\right)\right)}\right) c(\mathcal{I}^*),$

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

(d) if $z(\mathcal{I}) \in \mathbb{Z}_{\geq 0}$ for all $\mathcal{I} \subseteq \mathcal{D}$, $c(\mathcal{I}_g) \leq \left(\sum_{i=i}^{M} \frac{1}{i}\right) c(\mathcal{I}^*) \leq (1 + \log M) c(\mathcal{I}^*)$, where $M = \max_{j \in \mathcal{D}} z(j)$.

We now have the following result, which characterizes 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.a.s.).*

3. Minimum-Penalty Information Set Selection

In this section, we consider the dual problem, where the central node (designer) has a fixed budget for selecting information sources and seeks to minimize the maximum penalty of misclassifying

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: $\mathcal{I}_K$

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

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, subject to the budget constraint. Thus, we have an instance of a multi-objective constrained optimization problem. We consider an a priori method, by scalarizing the multi-objective optimization into a single-objective optimization problem, subject to the budget constraint. The optimal solution to this single-objective optimization problem is a Pareto optimal solution to the multi-objective optimization problem (Hwang and Masud (2012)). We now formalize the Minimum-Penalty Information Set Selection (MPIS) Problem as follows.

Problem 2 (MPIS) *Consider a set $\Theta = \{\theta_1, \dots, \theta_m\}$ 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 MPIS Problem is to find a set of selected information sources $\mathcal{I} \subseteq \mathcal{D}$ that solves*

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

We now define an optimization problem equivalent to Problem 2, in order to leverage submodularity and establish theoretical guarantees for a greedy algorithm. Consider the following problem:

$$\max_{\mathcal{I}\subseteq\mathcal{D}} \sum_{\theta_p\in\Theta} \left(1 - \max_{\theta_j\in F_{\theta_p}(\mathcal{I})} \xi_{pj} \right); \;\; \textit{s.t.} \; \sum_{i\in\mathcal{I}} c_i \leq K. \tag{13}$$

It is easy to verify that the problem defined in (13) is equivalent to the problem defined in (12), i.e., the information set $\mathcal{I} \subseteq \mathcal{D}$ that optimizes the problem in Equation (13) is also the optimal solution to the Problem 2. Denote $\Lambda(\mathcal{I}) = \sum_{\theta_p\in\Theta} \left(1 - g_{\theta_p}(\mathcal{I})\right)$. From Lemma 7, it follows that $\Lambda(\mathcal{I})$ is a monotone non-decreasing submodular set function. Since $F_{\theta_p}(\emptyset) = \Theta \;\; \forall \theta_p \in \Theta$, we have $\Lambda(\emptyset) = \sum_{\theta_p\in\Theta} \left(1 - g_{\theta_p}(\emptyset)\right) = m - \sum_{\theta_p\in\Theta} \max_{\theta_i\in\Theta} \xi_{pi}$.

Based on the performance guarantees provided for greedy maximization of monotone, non-decreasing submodular set functions subject to Knapsack constraints in Sviridenko (2004), we have the following result that characterizes the performance of Algorithm 2 for Problem 2.

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 Problem 2. 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., (12)) that satisfies $\Lambda(\mathcal{I}_K) \geq \left(1 - e^{-1}\right) \Lambda(\mathcal{I}_K^*) + c$, where $c = \Lambda(\emptyset)/e$ and $\tilde{N}$ is specified in (6).*

We now have the following result, which characterizes the asymptotic performance of the greedy algorithm.

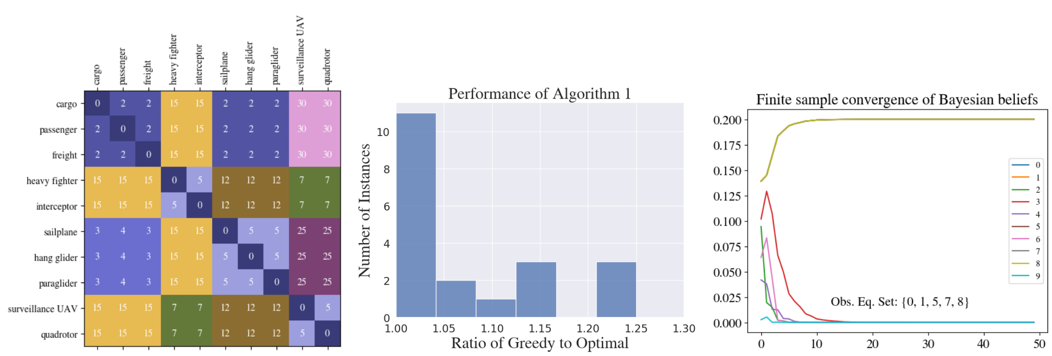
Corollary 12 *Instate the hypothesis and notation of Theorem 2. As* $t \to \infty$*, we have the following: (a)* $\mu_\infty^{\mathcal{I}}(\theta_q) = 0 \quad \forall \theta_q \notin F_{\theta_p}(\mathcal{I})$*, and (b)* $\mu_\infty^{\mathcal{I}}(\theta_q) = \frac{1}{|F_{\theta_p}(\mathcal{I})|} \quad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

4. Empirical Evaluation

In this section, we validate the theoretical results presented in this paper through some numerical simulations.

4.1. Case Study: Aerial Vehicle Classification

We consider a classification task (Problem 1) where a central designer has to select a minimum cost information set to classify an aerial vehicle into one of the following 10 classes: *cargo, passenger, freight, heavy fighter, interceptor, sailplane, hang glider, paraglider, surveillance UAV,* and *quadrotor*. The penalty matrix is as shown in Figure 1 (a). The penalty values for each true state are normalized by 100. We set $|\mathcal{D}| = 10$, the costs c_i for $i \in \mathcal{D}$ are sampled uniformly from $\{1, \ldots, 10\}$, and the observationally equivalent set $F_{\theta_p}(i)$ for each $\theta_p \in \Theta$ and $i \in \mathcal{D}$ is randomly generated. 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 20 randomly generated instances, we plot the ratio $c(\mathcal{I}_g)/c(\mathcal{I}^*)$ in Figure 1 (b). The plot shows near-optimal performance of the greedy algorithm, which aligns with the theoretical guarantees presented earlier.



(a) (b) (c)

Figure 1: Numerical Example for Problem 1 (a) Penalty Matrix, (b) Histogram of ratio of greedy to optimal of Algorithm 2, (c) Finite sample convergence of Bayesian beliefs.

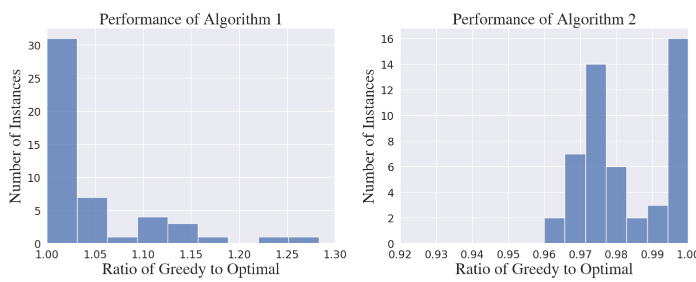
4.2. Finite Sample Convergence of Bayesian beliefs

We consider an instance of Problem 1 with number of states $\Theta = 10$ and show the finite sample convergence of beliefs. We randomly sample the observationally equivalent set $F_{\theta_p}(\mathcal{I})$ for an information set $\mathcal{I} \in \mathcal{D}$. We start with a uniform prior and apply the Bayesian update rule as in (2).

Figure 1(c) shows convergence of beliefs under 50 observation samples. It can be verified that the beliefs over the states not in the observationally equivalent set of the true state $\theta_p = 0$ get arbitrarily close to zero and the beliefs over the states which are observationally equivalent to the true state $\theta_p = 0$, i.e., $\{1, 5, 7, 8\}$ are all equal to 0.2, validating the results presented in Theorem 2.

4.3. Evaluation on Random Instances

We now evaluate the performance of the proposed greedy algorithms over random instances. For Problem 1, we generate 50 random instances, where for each instance, we set the total number of data sources $\mathcal{D}$ to be 10 and for each source $i \in \mathcal{D}$, the selection cost c_i is drawn uniformly from $\{1, 2, \ldots, 10\}$. We generate a random row-stochastic penalty matrix $\Xi \in \mathbb{R}^{|\Theta| \times |\Theta|}$ with $|\Theta| = 20$. We consider a uniform prior $\mu_0(\theta) = 1/|\Theta|$ and set a penalty threshold of $R/|\Theta|$, where R is drawn randomly from $\{1, \ldots, |\Theta| - 1\}$ for all $\theta_p \in \Theta$. Note that the constraints in Equation (11) can be completely specified by $F_{\theta_p}(i)$ for all $\theta_p \in \Theta$ and for all $i \in \mathcal{D}$, which capture the underlying likelihood functions $\ell_i(\cdot|\theta_p)$. We randomly generate the set $F_{\theta_p}(i)$ for all $i \in \mathcal{D}$ and for all $\theta_p \in \Theta$. In Figure 2(a), we plot the ratio of greedy to optimal cost, i.e., $c(\mathcal{I}_g)/c(\mathcal{I}^*)$. Similarly, we generate 50 random instances of Problem 2. Here, we set uniform costs for the data sources. For each instance, we set the total number of data sources $\mathcal{D}$ to be 10 and the selection budget to be $K = 4$. In Figure 2(b), we plot the ratio of $\Lambda(\mathcal{I}_K)/\Lambda(\mathcal{I}^*)$. We observe from Figure 2 that the greedy algorithms exhibit a near-optimal performance, which aligns with the respective theoretical bounds provided in Theorem 9 (for Algorithm 1) and Theorem 11 (for Algorithm 2).



(a) (b)

Figure 2: Histogram of ratio of greedy to optimal (a) Alg. 1 (Problem 1), (b) Alg. 2 (Problem 2).

5. Conclusion

In this work, we considered the scenario where a central designer seeks to select a minimum-cost information set while ensuring the maximum penalty for misclassifying the true hypothesis remains bounded, under finite observation samples from the sources. We proved that is problem is submodular and presented a greedy algorithm with high-probability near-optimal performance guarantees. We also provided the associated sample complexity for the non-asymptotic convergence rates of Bayesian beliefs. Next, we considered the dual problem, where the central designer has a fixed budget for information set selection and aims to minimize the maximum penalty of misclassifying the true hypothesis and provided high-probability near-optimal performance guarantees for a greedy algorithm, under finite observation samples. Finally, we validated the theoretical results through some numerical simulations, and showed that the greedy algorithm provides near optimal solutions.

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.

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

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.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

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*, 15(2):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.

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.

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.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, 1978.

NHTSA. PE16-007 Technical Report: Tesla crash preliminary evaluation report Jan 2017. *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.

Maxim Sviridenko. A note on maximizing a submodular set function subject to a Knapsack constraint. *Operations Research Letters*, 32(1):41–43, 2004.

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.

Appendix A.

A.1. Proof of Theorem 2

Theorem *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 $\delta \in [0,1]$, $0 < \epsilon < 1$ 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 the following:*

(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

$$N(\delta, \epsilon, L) = \left\lceil \frac{2L^2}{\epsilon^2} \log \frac{1}{\delta} \right\rceil. \tag{14}$$

Proof Consider a class $\theta_q = \Theta \setminus \theta_p$, which is not the true state of the world. We now define, for all $\theta_q \in \Theta \setminus \theta_p$ and for all $k \in \mathbb{N}_+$, the following:

$$\phi_k^{\mathcal{I}}(\theta_q) = \log \frac{\mu_k^{\mathcal{I}}(\theta_q)}{\mu_k^{\mathcal{I}}(\theta_p)} \text{ and } \eta_k^{\mathcal{I}}(\theta_q) = \log \frac{\ell_{\mathcal{I}}(o_k^{\mathcal{I}}|\theta_q)}{\ell_{\mathcal{I}}(o_k^{\mathcal{I}}|\theta_p)}, \tag{15}$$

where $o_k^{\mathcal{I}} \in \mathcal{O}_{\mathcal{I}}$ is the joint observation and $\ell_{\mathcal{I}}(\cdot)$ is the joint likelihood function of the information set $\mathcal{I} \subseteq \mathcal{D}$. From the Bayesian update rule in Equation (2), we have:

$$\phi_t^{\mathcal{I}}(\theta_q) = \phi_0^{\mathcal{I}}(\theta_q) + \sum_{k=1}^{t} \eta_k^{\mathcal{I}}(\theta_q). \tag{16}$$

Since we assume a uniform prior, i.e., $\mu_0(\theta) = \frac{1}{m}$ $\forall \theta \in \Theta$, we have that $\phi_0^{\mathcal{I}}(\theta_q) = 0$ $\forall \theta_q \in \Theta \setminus \theta_p$ and thus $\phi_t^{\mathcal{I}}(\theta_q) = \sum_{k=1}^{t} \eta_k^{\mathcal{I}}(\theta_q)$, $\forall t \in \mathbb{N}_+$. Note that $\{\eta_k^{\mathcal{I}}(\theta_q)\}$ is a sequence of *i.i.d.* random variables which are bounded and have a finite mean (by Equation 1). Each random variable $\eta_k^{\mathcal{I}}(\theta_q)$ has a mean given by $-K(\theta_p, \theta_q)$, where the mean is obtained by using the expectation operator $\mathbb{E}^{\theta p}[\cdot]$ associated with the probability measure $\mathbb{P}^{\theta_p}$ as defined in Section 2. By the strong law of large numbers, we have that $\frac{1}{t}\sum_{k=1}^{t} \eta_k^{\mathcal{I}}(\theta_q) \to -K(\theta_p, \theta_q)$ asymptotically almost surely (a.a.s.).

If $\theta_q \in F_{\theta_p}(\mathcal{I})$, we know from (4) that $\eta_k^{\mathcal{I}}(\theta_q) = 0$ for all $k \in \{1, \ldots, t\}$, and thus we have $\phi_t^{\mathcal{I}}(\theta_q) = 0$ for all $t \in \mathbb{N}_+$. This directly implies that $\mu_t^{\mathcal{I}}(\theta_q) = \mu_t^{\mathcal{I}}(\theta_p)$ $\forall \theta_q \in F_{\theta_p}(\mathcal{I})$, establishing part (a) of the result. Let $\hat{K}(\theta_p, \theta_q) = -\frac{1}{t}\sum_{k=1}^{t} \eta_k^{\mathcal{I}}(\theta_q)$ denote the sample mean (estimated KL divergence). Now consider a state $\theta_q \notin F_{\theta_p}(\mathcal{I})$, Equation (16) can be equivalently written as

$$\mu_t^{\mathcal{I}}(\theta_q) = \mu_t^{\mathcal{I}}(\theta_p) \exp\left(t.\frac{1}{t}\sum_{k=1}^{t} \eta_k^{\mathcal{I}}(\theta_q)\right) = \mu_t^{\mathcal{I}}(\theta_p) \exp\left(-t\hat{K}(\theta_p, \theta_q)\right). \tag{17}$$

By Equation (1) and Hoeffding's Inequality (Hoeffding (1994)), for all $\epsilon > 0$, we have the following:

$$\mathbb{P}\left(\left|\frac{1}{t}\sum_{k=1}^{t} \eta_k^{\mathcal{I}}(\theta_q) - (-K(\theta_p, \theta_q))\right| \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2 t}{2L^2}\right). \tag{18}$$

This condition is equivalent to:

$$\mathbb{P}\left(\left|\frac{1}{t}\sum_{k=1}^{t}\eta_k^{\mathcal{I}}(\theta_q) - (-K(\theta_p,\theta_q))\right| \leq \epsilon\right) \geq 1 - \exp\left(-\frac{\epsilon^2 t}{2L^2}\right). \tag{19}$$

Now let $\delta = \exp\left(-\frac{\epsilon^2 t}{2L^2}\right)$ which yields $t = \frac{2L^2}{\epsilon^2}\log\frac{1}{\delta}$. The condition in Equation (19) means that, with probability at least $1-\delta$, we have that $\left|-\hat{K}(\theta_p,\theta_q) - (-K(\theta_p,\theta_q))\right| \leq \epsilon$, for all $t \geq \frac{2L^2}{\epsilon^2}\log\frac{1}{\delta}$. We know that $\mu_t^{\mathcal{I}}(\theta_p) \leq 1$ for any $t \in \mathbb{N}_+$. Now, by combining Equations (17) and (19), with probability at least $1-\delta$, for all $t > N(\delta,\epsilon,L)$, we have

$$\mu_t^{\mathcal{I}}(\theta_q) \leq \exp\left(-t(|K(\theta_p,\theta_q) - \epsilon|)\right) \qquad \forall\theta_q \notin F_{\theta_p}(\mathcal{I}), \tag{20}$$

where $N(\delta,\epsilon,L) = \left\lceil\frac{2L^2}{\epsilon^2}\log\frac{1}{\delta}\right\rceil$, establishing part (b) of the result. ∎

A.2. Proof of Corollary 3

Corollary *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 a $\delta \in (0,1)$ and $\epsilon > 0$, 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*

$$\tilde{N} = \left\lceil \max\left\{\frac{2L^2}{\epsilon^2}\log\frac{1}{\delta}, \frac{1}{\min\limits_{\theta_p,\theta_q\in\Theta}|K(\theta_p,\theta_q)-\epsilon|}\log\frac{1}{\mu_{th}}\right\}\right\rceil. \tag{21}$$

Proof From Theorem 2, we know that with probability at least $1-\delta$, for all $t > N(\delta,\epsilon,L)$, we have

$$\mu_t^{\mathcal{I}}(\theta_q) \leq \exp\left(-t(|K(\theta_p,\theta_q) - \epsilon|)\right) \qquad \forall\theta_q \notin F_{\theta_p}(\mathcal{I}), \tag{22}$$

where $N(\delta,\epsilon,L) = \left\lceil\frac{2L^2}{\epsilon^2}\log\frac{1}{\delta}\right\rceil$. Since we require $\mu_t^{\mathcal{I}}(\theta_q) \leq \mu_{th}$ for all $\theta_q \notin F_{\theta_p}$, we let

$$\mu_{th} = \exp\left(-t(|K(\theta_p,\theta_q) - \epsilon|)\right) \qquad \forall\theta_q \notin F_{\theta_p}(\mathcal{I}). \tag{23}$$

Re-arranging the terms in the above equation, we get

$$t = \frac{1}{|K(\theta_p,\theta_q)-\epsilon|}\log\frac{1}{\mu_{th}} \tag{24}$$

Since the true state is not known a priori, in order to ensure $\mu_t^{\mathcal{I}}(\theta_q) \leq \mu_{th}\forall\theta_q \notin F_{\theta_p}(\mathcal{I})$, for any pair of hypotheses θ_p,θ_q, we have

$$t > \frac{1}{\min\limits_{\theta_p,\theta_q\in\Theta}|K(\theta_p,\theta_q)-\epsilon|}\log\frac{1}{\mu_{th}} \tag{25}$$

∎

A.3. Proof of Lemma 7

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

Proof Consider any $\mathcal{I}_1 \subseteq \mathcal{I}_2 \subseteq \mathcal{D}$ and any $j \in \mathcal{D} \backslash \mathcal{I}_2$. We then have the following:

$$
\begin{aligned}
& g_{\theta_p}\left(\mathcal{I}_1\right) - g_{\theta_p}\left(\mathcal{I}_1 \cup \{j\}\right) \\
&= \max_{\theta_i \in F_{\theta_p}(\mathcal{I}_1)} \xi_{pi} - \max_{\theta_i \in F_{\theta_p}(\mathcal{I}_1 \cup \{j\})} \xi_{pi} \\
&= \max_{\theta_i \in F_{\theta_p}(\mathcal{I}_1)} \xi_{pi} - \max_{\theta_i \in F_{\theta_p}(\mathcal{I}_1) \cap F_{\theta_p}(j)} \xi_{pi} \\
&= \max_{\theta_i \in F_{\theta_p}(\mathcal{I}_1) \backslash \left(F_{\theta_p}(\mathcal{I}_1) \cap F_{\theta_p}(j)\right)} \xi_{pi} = \max_{\theta_i \in F_{\theta_p}(\mathcal{I}_1) \backslash F_{\theta_p}(j)} \xi_{pi}.
\end{aligned}
$$

Note that the above arguments follow from De Morgan's laws. Similarly, we also have

$$
g_{\theta_p}\left(\mathcal{I}_2\right) - g_{\theta_p}\left(\mathcal{I}_2 \cup \{j\}\right) = \max_{\theta_i \in F_{\theta_p}(\mathcal{I}_2) \backslash F_{\theta_p}(j)} \xi_{pi}.
$$

Since $\mathcal{I}_1 \subseteq \mathcal{I}_2$, we have $F_{\theta_p}(\mathcal{I}_2) \backslash F_{\theta_p}(j) \subseteq F_{\theta_p}(\mathcal{I}_1) \backslash F_{\theta_p}(j)$, which implies

$$
g_{\theta_p}\left(\mathcal{I}_1\right) - g_{\theta_p}\left(\mathcal{I}_1 \cup \{j\}\right) \geq g_{\theta_p}\left(\mathcal{I}_2\right) - g_{\theta_p}\left(\mathcal{I}_2 \cup \{j\}\right).
$$

Thus, we have

$$
(1 - g_{\theta_p}\left(\mathcal{I}_1 \cup \{j\}\right)) - (1 - g_{\theta_p}\left(\mathcal{I}_1\right)) \geq (1 - g_{\theta_p}\left(\mathcal{I}_2 \cup \{j\}\right)) - (1 - g_{\theta_p}\left(\mathcal{I}_2\right)).
$$

Since the above arguments hold for all $\theta_p \in \Theta$, $(1 - g_{\theta_p}(\cdot))$ is submodular for all $\theta_p \in \Theta$. ■

A.4. Proof of Lemma 8

Lemma *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* $\sum_{\theta_p \in \Theta} f'_{\theta_p}(\mathcal{I}) = \sum_{\theta_p \in \Theta} f'_{\theta_p}(\mathcal{D})$*, where* $f'_{\theta_p}(\cdot)$ *is as defined in (9).*

Proof Suppose the constraints $1 - \max_{\theta_i \in F_{\theta_p}(\mathcal{I})} \xi_{pi} \geq 1 - R_{\theta_p}$ hold for all $\theta_p \in \Theta$. It follows that $f'_{\theta_p}(\mathcal{I}) = 1 - R_{\theta_p}$ for all $\theta_p \in \Theta$. Noting that, $f_{\theta_p}(\mathcal{D}) \geq 1 - R_{\theta_p}$, we have $f'_{\theta_p}(\mathcal{D}) = 1 - R_{\theta_p}$ for all $\theta_p \in \Theta$, which implies $\sum_{\theta_p \in \Theta} f'_{\theta_p}(\mathcal{I}) = \sum_{\theta_p \in \Theta} f'_{\theta_p}(\mathcal{D})$. Conversely, suppose $\sum_{\theta_p \in \Theta} f'_{\theta_p}(\mathcal{I}) = \sum_{\theta_p \in \Theta} f'_{\theta_p}(\mathcal{D})$, i.e., $\sum_{\theta_p \in \Theta} \left(f'_{\theta_p}(\mathcal{I}) - \left(1 - R_{\theta_p}\right)\right) = 0$. Noting that $f'_{\theta_p}(\mathcal{I}) \leq 1 - R_{\theta_p}$ for all $\mathcal{I} \subseteq \mathcal{D}$, we have $f'_{\theta_p}(\mathcal{I}) = 1 - R_{\theta_p}$ for all $\theta_p \in \Theta$, i.e., $f_{\theta_p}(\mathcal{I}) \geq 1 - R_{\theta_p}$ for all $\theta_p \in \Theta$. This completes the proof of the lemma. ■

A.5. Proof of Theorem 9

The performance guarantees presented in this result are that of the submodular set covering problem in Wolsey (1982). With probability at least $1-\delta$, Algorithm 1, under $\tilde{N}(\epsilon, \delta, L)$ observation samples, enjoys the same performance guarantees, where δ, ϵ are specified by the central designer.

A.6. Proof of Corollary 10

Corollary *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.a.s.).*

Proof From Theorem 2, the following hold for $t \to \infty$:

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

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

We have $\lim_{t\to\infty} \mu_t^{\mathcal{I}}(\theta_q) = 0 \quad \forall \theta_q \notin F_{\theta_p}(\mathcal{I})$ and $\mu_t^{\mathcal{I}}(\theta_q) = \mu_t^{\mathcal{I}}(\theta_p) \ \forall \theta_q \in F_{\theta_p}(\mathcal{I})$. Since $\sum_{\theta \in \Theta} \mu_t^{\mathcal{I}}(\theta) = 1$ for all t, we have $\lim_{t\to\infty} \mu_t^{\mathcal{I}}(\theta_q) = \frac{1}{|F_{\theta_p}(\mathcal{I})|} \quad \forall \theta_q \in F_{\theta_p}(\mathcal{I})$. Since we have $\lim_{t\to\infty} \mu_t^{\mathcal{I}}(\theta_q) = 0 \quad \forall \theta_q \notin F_{\theta_p}(\mathcal{I})$, the central designer with predict one of the hypotheses $\theta_q \in F_{\theta_p}(\mathcal{I})$ as the true state of the world, with probability 1. Therefore, the guarantees provided in Theorem 9 for Problem 1 hold with probability 1. ■

A.7. Proof of Theorem 11

The performance guarantees presented in this result are that of the submodular maximization under Knapsack constraints in Sviridenko (2004). With probability at least $1 - \delta$, Algorithm 2, under $\tilde{N}(\epsilon, \delta, L)$ observation samples, enjoys the same performance guarantees, where δ, ϵ are specified by the central designer.

A.8. Proof of Corollary 12

The construction of the proof and arguments are similar to those presented in Corollary 10.