Proceedings of Machine Learning Research vol 144:1–12, 2021

Near-Optimal Data Source Selection for Bayesian Learning

Lintao Ye LYE2@ND.EDU
Department of Electrical Engineering, University of Notre Dame, IN 46556, USA

Aritra Mitra AMITRA20@SEAS.UPENN.EDU
Department of Electrical and Systems Engineering, University of Pennsylvania, PA 19104, USA

Shreyas Sundaram SUNDARA2@PURDUE.EDU
School of Electrical and Computer Engineering, Purdue University, IN 47907, USA

Abstract

We study a fundamental problem in Bayesian learning, where the goal is to select a set of data sources with minimum cost while achieving a certain learning performance based on the data streams provided by the selected data sources. First, we show that the data source selection problem for Bayesian learning is NP-hard. We then show that the data source selection problem can be transformed into an instance of the submodular set covering problem studied in the literature, and provide a standard greedy algorithm to solve the data source selection problem with provable performance guarantees. Next, we propose a fast greedy algorithm that improves the running times of the standard greedy algorithm, while achieving performance guarantees that are comparable to those of the standard greedy algorithm. Finally, we validate the results using numerical examples, and show that the greedy algorithms work well in practice.[1]

Keywords: Bayesian Learning, Combinatorial Optimization, Approximation Algorithms, Greedy Algorithms

1. Introduction

The problem of learning the true state of the world based on streams of data has been studied by researchers from different fields. A classical method to tackle this task is Bayesian learning, where we start with a prior belief about the true state of the world and update our belief based on the data streams from the data sources (e.g., Gelman et al. (2013)). In particular, the data streams can come from a variety of sources, including experiment outcomes (Chaloner and Verdinelli, 1995), medical tests (Kononenko, 1993), and sensor measurements (Krause et al., 2008), etc. In practice, we need to pay a cost in order to obtain the data streams from the data sources; for example, conducting certain experiments or installing a particular sensor incurs some cost that depends on the nature of the corresponding data source. Thus, a fundamental problem that arises in Bayesian learning is to select a subset of data sources with the smallest total cost, while ensuring a certain level of the learning performance based on the data streams provided by the selected data sources.

In this paper, we focus on a standard Bayesian learning rule that updates the belief on the true state of the world recursively based on the data streams. The learning performance is then characterized by an error given by the difference between the steady-state belief obtained from the learning rule and the true state of the world. Moreover, we consider the scenario where the data sources are selected a priori before running the Bayesian learning rule, and the set of selected data sources is fixed over

1. An extended version of this paper that includes all the omitted proofs can be found on arXiv as Ye et al. (2020).

time. We then formulate and study the Bayesian Learning Data Source Selection (BLDS) problem, where the goal is to minimize the cost spent on the selected data sources while ensuring that the error of the learning process is within a prescribed range.

1.1. Related Work

In Dasgupta (2005) and Golovin et al. (2010), the authors studied the data source selection problem for Bayesian active learning. They considered the scenario where the data sources are selected in a sequential manner with a single data source selected at each time step in the learning process. The goal is then to find a policy on sequentially selecting the data sources with minimum cost, while the true state of the world can be identified based on the selected data sources. In contrast, we consider the scenario where a subset of data sources are selected a priori. Moreover, the selected data sources may not necessarily lead to the learning of the true state of the world. Thus, we characterize the performance of the learning process via its steady-state error.

The problem studied in this paper is also related but different from the problem of ensuring sparsity in learning, where the goal is to identify the fewest number of features in order to explain a *given* set of data (Palmer et al., 2004; Krause and Cevher, 2010).

Finally, our problem formulation is also related to the sensor placement problem that has been studied for control systems (e.g., Mo et al. (2011) and Ye et al. (2021)), signal processing (e.g., Chepuri and Leus (2014) and Ye and Sundaram (2019)), and machine learning (e.g., Krause et al. (2008)). In general, the goal of these problems is either to optimize certain (problem-specific) performance metrics of the estimate associated with the measurements of the placed sensors while satisfying the sensor placement budget constraint, or to minimize the cost spent on the placed sensors while ensuring that the estimation performance is within a certain range.

1.2. Contributions

First, we formulate the Bayesian Learning Data Source Selection (BLDS) problem, and show that the BLDS problem is NP-hard. Next, we show that the BLDS problem can be transformed into an instance of the submodular set covering problem studied in Wolsey (1982). The BLDS problem can then be solved using a standard greedy algorithm with approximation guarantees, where the query complexity of the greedy algorithm is $O(n^2)$, with n to be the number of all candidate data sources. In order to improve the running times of the greedy algorithm, we further propose a fast greedy algorithm with query complexity $O(\frac{n}{\epsilon}\ln\frac{n}{\epsilon})$, where $\epsilon \in (0,1)$. The fast greedy algorithm achieves comparable performance guarantees to those of the standard greedy algorithm, and can also be applied to solve the general submodular set covering problem with performance guarantees. Finally, we provide illustrative examples to interpret the performance bounds obtained for the greedy algorithms applied to the BLDS problem, and give simulation results.

2. The Bayesian Learning Data Source Selection Problem

In this section, we formulate the data source selection problem for Bayesian learning that we will study in this paper. Let $\Theta \triangleq \{\theta_1, \theta_2, \ldots, \theta_m\}$ be a finite set of possible states of the world, where $m \triangleq |\Theta|$. We consider a set $[n] \triangleq \{1, 2, \ldots, n\}$ of data sources that can provide data streams of the state of the world. At each discrete time step $k \in \mathbb{Z}_{\geq 1}$, the signal (or observation) provided by source $i \in [n]$ is denoted as $\omega_{i,k} \in S_i$, where S_i is the signal space of source i. Conditional on the state of the world $\theta \in \Theta$, an observation profile of the n sources at time k, denoted as $\omega_k \triangleq (\omega_{1,k}, \ldots, \omega_{n,k}) \in S_1 \times \cdots \times S_n$, is generated by the likelihood function $\ell(\cdot|\theta)$. Let $\ell_i(\cdot|\theta)$

denote the i-th marginal of $\ell(\cdot|\theta)$, which is the signal structure of data source $i \in [n]$. We make the following assumption on the observation model (e.g., see Jadbabaie et al. (2012); Liu et al. (2014); Lalitha et al. (2014); Nedić et al. (2017)).

Assumption 1 *For each source $i \in [n]$, the signal space S_i is finite, and the likelihood function $\ell_i(\cdot|\theta)$ satisfies $l_i(s_i|\theta) > 0$ for all $s_i \in S_i$ and for all $\theta \in \Theta$. Furthermore, for all $\theta \in \Theta$, the observations are independent over time, i.e., $\{\omega_{i,1}, \omega_{i,2}, \dots\}$ is a sequence of independent identically distributed (i.i.d.) random variables. The likelihood function is assumed to satisfy $\ell(\cdot|\theta) = \prod_{i=1}^{n} \ell_i(\cdot|\theta)$ for all $\theta \in \Theta$, where $\ell_i(\cdot|\theta)$ is the i-th marginal of $\ell(\cdot|\theta)$.*

Consider the scenario where there is a (central) designer who needs to select a subset of data sources in order to learn the true state of the world based on the observations from the selected sources. Specifically, each data source $i \in [n]$ is assumed to have an associated selection cost $h_i \in \mathbb{R}_{>0}$. Considering any $\mathcal{I} \triangleq \{n_1, n_2, \dots, n_\tau\}$ with $\tau = |\mathcal{I}|$, we let $h(\mathcal{I})$ denote the sum of the costs of the selected sources in $\mathcal{I}$, i.e., $h(\mathcal{I}) \triangleq \sum_{n_i \in \mathcal{I}} h_{n_i}$. Let $\omega_{\mathcal{I},k} \triangleq (\omega_{n_1,k}, \dots, \omega_{n_\tau,k}) \in S_{n_1} \times \cdots \times S_{n_\tau}$ be the observation profile (conditioned on $\theta \in \Theta$) generated by the likelihood function $\ell_{\mathcal{I}}(\cdot|\theta)$, where $\ell_{\mathcal{I}}(\cdot|\theta) = \prod_{i=1}^{\tau} \ell_{n_i}(\cdot|\theta)$. We assume that the designer knows $\ell_i(\cdot|\theta)$ for all $\theta \in \Theta$ and for all $i \in [n]$, and thus knows $\ell_{\mathcal{I}}(\cdot|\theta)$ for all $\mathcal{I} \subseteq [n]$ and for all $\theta \in \Theta$. After the data sources are selected, the designer updates its belief of the state of the world using the following standard Bayes' rule:

$$\mu^{\mathcal{I}}_{k+1}(\theta) = \frac{\mu_0(\theta) \prod_{j=0}^{k} \ell_{\mathcal{I}}(\omega_{\mathcal{I},j+1}|\theta)}{\sum_{\theta_p \in \Theta} \mu_0(\theta_p) \prod_{j=0}^{k} \ell_{\mathcal{I}}(\omega_{\mathcal{I},j+1}|\theta_p)} \ \forall \theta \in \Theta, \tag{1}$$

where $u^{\mathcal{I}}_{k+1}(\theta)$ is the belief of the designer that θ is the true state at time step $k+1$, and $\mu_0(\theta)$ is the initial (or prior) belief of the designer that θ is the true state. We take $\sum_{\theta \in \Theta} \mu_0(\theta) = 1$ and $\mu_0(\theta) \in \mathbb{R}_{\geq 0}$ for all $\theta \in \Theta$. Note that $\sum_{\theta \in \Theta} \mu^{\mathcal{I}}_k(\theta) = 1$ for all $\mathcal{I} \subseteq [n]$ and for all $k \in \mathbb{Z}_{\geq 1}$, where $0 \leq \mu^{\mathcal{I}}_k(\theta) \leq 1$ for all $\theta \in \Theta$. In other words, $\mu^{\mathcal{I}}_k(\cdot)$ is a probability distribution over Θ for all $k \in \mathbb{Z}_{\geq 1}$ and for all $\mathcal{I} \subseteq [n]$. Rule (1) is also equivalent to the following recursive rule:

$$\mu^{\mathcal{I}}_{k+1}(\theta) = \frac{\mu^{\mathcal{I}}_k(\theta) \ell_{\mathcal{I}}(\omega_{\mathcal{I},k+1}|\theta)}{\sum_{\theta_p \in \Theta} \mu^{\mathcal{I}}_k(\theta_p) \ell_{\mathcal{I}}(\omega_{\mathcal{I},k+1}|\theta_p)} \ \forall \theta \in \Theta, \tag{2}$$

with $\mu^{\mathcal{I}}_0(\theta) \triangleq \mu_0(\theta)$ for all $\mathcal{I} \subseteq [n]$. For a given state $\theta \in \Theta$ and a given $\mathcal{I} \subseteq [n]$, we define the set of *observationally equivalent* states to θ as

$$F_\theta(\mathcal{I}) \triangleq \arg\min_{\theta_p \in \Theta} D_{KL}(\ell_{\mathcal{I}}(\cdot|\theta_p) \| \ell_{\mathcal{I}}(\cdot|\theta)),$$

where $D_{KL}(\ell_{\mathcal{I}}(\cdot|\theta_p) \| \ell_{\mathcal{I}}(\cdot|\theta))$ is the Kullback-Leibler (KL) divergence between the likelihood functions $\ell_{\mathcal{I}}(\cdot|\theta_p)$ and $\ell_{\mathcal{I}}(\cdot|\theta)$. Noting that $D_{KL}(\ell_{\mathcal{I}}(\cdot|\theta) \| \ell_{\mathcal{I}}(\cdot|\theta)) = 0$ and that the KL divergence is always nonnegative, we have $\theta \in F_\theta(\mathcal{I})$ for all $\theta \in \Theta$ and for all $\mathcal{I} \subseteq [n]$. Equivalently, we can write

$$F_\theta(\mathcal{I}) = \{\theta_p \in \Theta : \ \ell_{\mathcal{I}}(s_{\mathcal{I}}|\theta_p) = \ell_{\mathcal{I}}(s_{\mathcal{I}}|\theta), \forall s_{\mathcal{I}} \in S_{\mathcal{I}}\}, \tag{3}$$

where $S_{\mathcal{I}} \triangleq S_{n_1} \times \cdots \times S_{n_\tau}$. Note that $F_\theta(\mathcal{I})$ is the set of states that cannot be distinguished from θ based on the data streams provided by the data sources indicated by $\mathcal{I}$. Moreover, we define $F_\theta(\emptyset) \triangleq \Theta$. Noting that $\ell_{\mathcal{I}}(\cdot|\theta) = \prod_{i=1}^{\tau} \ell_{n_i}(\cdot|\theta)$ under Assumption 1, we can further obtain from Eq. (3) the following:

$$F_\theta(\mathcal{I}) = \bigcap_{n_i \in \mathcal{I}} F_\theta(n_i), \tag{4}$$

for all $\mathcal{I} \subseteq [n]$ and for all $\theta \in \Theta$. Using similar arguments to those for Lemma 1 in Mitra et al. (2020), one can show the following result.

Lemma 1 *Suppose the true state of the world is θ^*, and $\mu_0(\theta) > 0$ for all $\theta \in \Theta$. For all $\mathcal{I} \subseteq [n]$, the rule given in (1) ensures: (a) $\lim_{k\to\infty} \mu_k^{\mathcal{I}}(\theta_p) = 0$ almost surely (a.s.) for all $\theta_p \notin F_{\theta^*}(\mathcal{I})$; and (b) $\lim_{k\to\infty} \mu_k^{\mathcal{I}}(\theta_q) = \frac{\mu_0(\theta_q)}{\sum_{\theta\in F_{\theta^*}(\mathcal{I})} \mu_0(\theta)}$ a.s. for all $\theta_q \in F_{\theta^*}(\mathcal{I})$, where $F_{\theta^*}(\mathcal{I})$ is given by Eq. (4).*

Consider a true state $\theta^* \in \Theta$ and a set $\mathcal{I} \subseteq [n]$ of selected sources. In order to characterize the (steady-state) learning performance of rule (1), we will use the following error metric (e.g., Jadbabaie et al. (2013)):

$$e_{\theta^*}(\mathcal{I}) \triangleq \frac{1}{2} \lim_{k\to\infty} \|\mu_k^{\mathcal{I}} - \mathbf{1}_{\theta^*}\|_1, \tag{5}$$

where $\mu_k^{\mathcal{I}} \triangleq \begin{bmatrix}\mu_k^{\mathcal{I}}(\theta_1) & \cdots & \mu_k^{\mathcal{I}}(\theta_m)\end{bmatrix}'$, and $\mathbf{1}_{\theta^*} \in \mathbb{R}^m$ is a (column) vector where the element that corresponds to θ^* is 1 and all the other elements are zero. Note that $\frac{1}{2}\|\mu_k^{\mathcal{I}} - \mathbf{1}_{\theta^*}\|_1$ is also known as the total variation distance between the two distributions $\mu_k^{\mathcal{I}}$ and $\mathbf{1}_{\theta^*}$ (e.g., Brémaud (2013)). Also note that $e_{\theta^*}(\mathcal{I})$ exists (a.s.) due to Lemma 1. We then obtain from Lemma 1 that $e_{\theta^*}(\mathcal{I}) = 1 - \frac{\mu_0(\theta^*)}{\sum_{\theta\in F_{\theta^*}(\mathcal{I})} \mu_0(\theta)}$ holds almost surely. Since the true state is not known a priori to the designer, we further define

$$e^s_{\theta_p}(\mathcal{I}) \triangleq 1 - \frac{\mu_0(\theta_p)}{\sum_{\theta\in F_{\theta_p}(\mathcal{I})} \mu_0(\theta)} \; \forall \theta_p \in \Theta, \tag{6}$$

which represents the (steady-state) total variation distance between the designer's belief $\mu_k^{\mathcal{I}}$ and $\mathbf{1}_{\theta_p}$, when θ_p is assumed to be the true state of the world. We then define the Bayesian Learning Data Source Selection (BLDS) problem as follows.

Problem 1 *(BLDS) Consider a set $\Theta = \{\theta_1, \ldots, \theta_m\}$ of possible states of the world; a set $[n]$ of data sources providing data streams, where the signal space of source $i \in [n]$ is S_i and the observation from source $i \in [n]$ under state $\theta \in \Theta$ is generated by $\ell_i(\cdot|\theta)$; a selection cost $h_i \in \mathbb{R}_{>0}$ of each source $i \in [n]$; an initial belief $\mu_0(\theta) \in \mathbb{R}_{>0}$ for all $\theta \in \Theta$ with $\sum_{\theta\in\Theta} \mu_0(\theta) = 1$; and prescribed error bounds $0 \leq R_{\theta_p} \leq 1$ ($R_{\theta_p} \in \mathbb{R}$) for all $\theta_p \in \Theta$. The BLDS problem is to find a set of selected data sources $\mathcal{I} \subseteq [n]$ that solves*

$$\begin{aligned} &\min_{\mathcal{I}\subseteq[n]} h(\mathcal{I}) \\ &s.t.\ e^s_{\theta_p}(\mathcal{I}) \leq R_{\theta_p} \; \forall \theta_p \in \Theta, \end{aligned} \tag{7}$$

where $e^s_{\theta_p}(\mathcal{I})$ is defined in (6).

Note that the constraints in (7) also capture the fact that the true state of the world is unknown to the designer a priori. In other words, for any set $\mathcal{I} \subseteq [n]$ and for any $\theta_p \in \Theta$, the constraint $e^s_{\theta_p}(\mathcal{I}) \leq R_{\theta_p}$ requires the (steady-state) learning error $e^s_{\theta_p}(\mathcal{I})$ to be upper bounded by R_{θ_p} when the true state of the world is assumed to be θ_p. Moreover, the interpretation of R_{θ_p} for $\theta_p \in \Theta$ is as follows. When $R_{\theta_p} = 0$, we see from (6) and the constraint $e^s_{\theta_p}(\mathcal{I}) \leq R_{\theta_p}$ that $F_{\theta_p}(\mathcal{I}) = \{\theta_p\}$. In other words, the constraint $e^s_{\theta_p}(\mathcal{I}) \leq 0$ requires that any $\theta_q \in \Theta \setminus \{\theta_p\}$ is *not* observationally equivalent to θ_p, based on the observations from the data sources indicated by $\mathcal{I} \subseteq [n]$. Next, when $R_{\theta_p} = 1$, we know from (6) that the constraint $e^s_{\theta_p}(\mathcal{I}) \leq 1$ is satisfied for all $\mathcal{I} \subseteq [n]$. Finally, when

$0 < R_{\theta_p} < 1$ and $\mu_0(\theta) = \frac{1}{m}$ for all $\theta \in \Theta$, where $m = |\Theta|$, we see from (6) that the constraint $e^s_{\theta_p}(\mathcal{I}) \leq R_{\theta_p}$ is equivalent to $|F_{\theta_p}(\mathcal{I})| \leq \frac{1}{1-R_{\theta_p}}$, i.e., the number of states that are observationally equivalent to θ_p should be less than or equal to $\frac{1}{1-R_{\theta_p}}$, based on the observations from the data source indicated by $\mathcal{I} \subseteq [n]$. In summary, the value of R_{θ_p} in the constraints represents the requirements of the designer on distinguishing state θ_p from other states in Θ; a smaller value of R_{θ_p} would imply that the designer wants to distinguish θ_p from more states in Θ and vice versa.

Remark 2 *The above problem formulation can be extended to distributed non-Bayesian learning (e.g., Nedić et al. (2017)). See the Appendix in Ye et al. (2020) for details about this extension.*

Next, we show that the BLDS problem is NP-hard via a reduction from the set cover problem, which is known to be NP-hard (e.g., Garey and Johnson (1979), Feige (1998)).

Theorem 3 *The BLDS problem is NP-hard even when all the data sources have the same cost.*

3. Greedy Algorithms for the BLDS Problem

In this section, we first show that the BLDS problem can be transformed into an instance of the submodular set covering problem studied in Wolsey (1982). We then consider two greedy algorithms for the BLDS problem and study their performance guarantees when applied to the problem. We start with the following definition.

Definition 4 *(Nemhauser et al. (1978)) A set function $f : 2^{[n]} \to \mathbb{R}$ is submodular if for all $X \subseteq Y \subseteq [n]$ and for all $j \in [n] \setminus Y$, $f(X \cup \{j\}) - f(X) \geq f(Y \cup \{j\}) - f(Y)$.*

To proceed, note that the constraint corresponding to θ_p in Problem 1 (i.e., (7)) is satisfied for all $\mathcal{I} \subseteq [n]$ if $R_{\theta_p} = 1$. Since $\mu_0(\theta) > 0$ for all $\theta \in \Theta$, we can equivalently write the constraints as

$$\sum_{\theta \in F_{\theta_p}(\mathcal{I})} \mu_0(\theta) \leq \frac{\mu_0(\theta_p)}{1 - R_{\theta_p}}, \ \forall \theta_p \in \Theta \text{ with } R_{\theta_p} < 1. \tag{8}$$

Define $F^c_{\theta_p}(\mathcal{I}) \triangleq \Theta \setminus F_{\theta_p}(\mathcal{I})$ for all $\theta_p \in \Theta$ and for all $\mathcal{I} \subseteq [n]$, where $F_{\theta_p}(\mathcal{I})$ is given by Eq. (4). Note that $F^c_{\theta_p}(\mathcal{I})$ is the set of states that can be distinguished from θ_p, given the data sources indicated by $\mathcal{I}$. Using the fact $\sum_{\theta \in \Theta} \mu_0(\theta) = 1$, (8) can be equivalently written as

$$\sum_{\theta \in F^c_{\theta_p}(\mathcal{I})} \mu_0(\theta) \geq 1 - \frac{\mu_0(\theta_p)}{1 - R_{\theta_p}}, \ \forall \theta_p \in \Theta \text{ with } R_{\theta_p} < 1. \tag{9}$$

Moreover, we note that the constraint corresponding to θ_p in (9) is satisfied for all $\mathcal{I} \subseteq [n]$ if $1 - \frac{\mu_0(\theta_p)}{1-R_{\theta_p}} \leq 0$, i.e., $R_{\theta_p} \geq 1 - \mu_0(\theta_p)$. Hence, we can equivalently write (9) as

$$\sum_{\theta \in F^c_{\theta_p}(\mathcal{I})} \mu_0(\theta) \geq 1 - \frac{\mu_0(\theta_p)}{1 - R_{\theta_p}}, \ \forall \theta_p \in \bar{\Theta},$$

where $\bar{\Theta} \triangleq \{\theta_p \in \Theta : 0 \leq R_{\theta_p} < 1 - \mu_0(\theta_p)\}$. For all $\mathcal{I} \subseteq [n]$, let us define

$$f_{\theta_p}(\mathcal{I}) \triangleq \sum_{\theta \in F^c_{\theta_p}(\mathcal{I})} \mu_0(\theta), \ \forall \theta_p \in \bar{\Theta}. \tag{10}$$

Noting that $F_{\theta_p}(\emptyset) = \Theta$, i.e., $F^c_{\theta_p}(\emptyset) = \emptyset$, we let $f_{\theta_p}(\emptyset) = 0$. It then follows directly from (10) that $f_{\theta_p} : 2^{[n]} \to \mathbb{R}_{\geq 0}$ is a monotone nondecreasing set function.[2]

Remark 5 *Note that in order to ensure that there exists $\mathcal{I} \subseteq [n]$ that satisfies the constraints in (9), we assume that $f_{\theta_p}([n]) \geq 1 - \frac{\mu_0(\theta_p)}{1-R_{\theta_p}}$ for all $\theta_p \in \bar{\Theta}$, since $f_{\theta_p}(\cdot)$ is nondecreasing for all $\theta_p \in \bar{\Theta}$.*

Lemma 6 *The set function $f_{\theta_p} : 2^{[n]} \to \mathbb{R}_{\geq 0}$ defined in (10) is submodular for all $\theta_p \in \bar{\Theta}$.*

Moreover, considering any $\mathcal{I} \subseteq [n]$, we define

$$f'_{\theta_p}(\mathcal{I}) \triangleq \min\{f_{\theta_p}(\mathcal{I}), 1 - \frac{\mu_0(\theta_p)}{1 - R_{\theta_p}}\} \; \forall \theta_p \in \bar{\Theta}, \tag{11}$$

where $f_{\theta_p}(\mathcal{I})$ is defined in (10). Since $f_{\theta_p}(\cdot)$ is submodular and nondecreasing with $f_{\theta_p}(\emptyset) = 0$ and $f_{\theta_p}([n]) \geq 1 - \frac{\mu_0(\theta_p)}{1-R_{\theta_p}}$, one can show that $f'_{\theta_p}(\cdot)$ is also submodular and nondecreasing with $f'_{\theta_p}(\emptyset) = 0$ and $f'_{\theta_p}([n]) = 1 - \frac{\mu_0(\theta_p)}{1-R_{\theta_p}}$. Since the sum of submodular functions remains submodular, $\sum_{\theta_p \in \bar{\Theta}} f'_{\theta_p}(\cdot)$ is submodular and nondecreasing. We also have the following result.

Lemma 7 *Consider any $\mathcal{I} \subseteq [n]$. The constraint $\sum_{\theta \in F^c_{\theta_p}(\mathcal{I})} \mu_0(\theta) \geq 1 - \frac{\mu_0(\theta_p)}{1-R_{\theta_p}}$ holds for all $\theta_p \in \bar{\Theta}$ if and only if $\sum_{\theta_p \in \bar{\Theta}} f'_{\theta_p}(\mathcal{I}) = \sum_{\theta_p \in \bar{\Theta}} f'_{\theta_p}([n])$, where $f'_{\theta_p}(\cdot)$ is defined in (11).*

Based on the above arguments, for all $\mathcal{I} \subseteq [n]$, we further define

$$z(\mathcal{I}) \triangleq \sum_{\theta_p \in \bar{\Theta}} f'_{\theta_p}(\mathcal{I}) = \sum_{\theta_p \in \bar{\Theta}} \min\{f_{\theta_p}(\mathcal{I}), 1 - \frac{\mu_0(\theta_p)}{1 - R_{\theta_p}}\}, \tag{12}$$

where $f_{\theta_p}(\mathcal{I})$ is defined in (10). We then see from Lemma 7 that (7) in Problem 1 can be equivalently written as

$$\begin{aligned} &\min_{\mathcal{I} \subseteq [n]} \; h(\mathcal{I}) \\ &s.t. \; z(\mathcal{I}) = z([n]), \end{aligned} \tag{13}$$

where one can show that $z(\cdot)$ defined in Eq. (12) is a nondecreasing and submodular set function with $z(\emptyset) = 0$. Now, considering an instance of the BLDS problem, for any $\mathcal{I} \subseteq [n]$ and for any $\theta \in \Theta$, one can obtain $F_\theta(\mathcal{I})$ (and $F^c_\theta(\mathcal{I})$) in $O(S|\mathcal{I}||\Theta|)$ time, where $S \triangleq \max_{n_i \in \mathcal{I}} |S_i|$ with S_i to be the signal space of source $n_i \in \mathcal{I}$. Therefore, we see from (10) and (12) that for any $\mathcal{I} \subseteq [n]$, one can compute the value of $z(\mathcal{I})$ in $O(Sn|\Theta|^2)$ time.

Problem (13) 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 with performance guarantees. Specifically, we consider the greedy algorithm defined in Algorithm 1 for the BLDS problem. The algorithm maintains a sequence of sets $\mathcal{I}^0_g, \mathcal{I}^1_g, \ldots, \mathcal{I}^T_g$ containing the selected elements from $[n]$, where $T \in \mathbb{Z}_{\geq 1}$. Note that Algorithm 1 requires $O(n^2)$ evaluations of function $z(\cdot)$, where $z(\mathcal{I})$ can be computed in $O(Sn|\Theta|^2)$ time for any $\mathcal{I} \subseteq [n]$ as argued above. In other words, the query complexity of Algorithm 1 is $O(n^2)$. We then have the following result from the arguments above (i.e., Lemmas 6-7) and Theorem 1 in Wolsey (1982), which characterizes the performance guarantees for the greedy algorithm (Algorithm 1) when applied to the BLDS problem.

2. A set function $f : 2^{[n]} \to \mathbb{R}$ is monotone nondecreasing if $f(X) \leq f(Y)$ for all $X \subseteq Y \subseteq [n]$.

Algorithm 1 Greedy Algorithm for BLDS

Input: $[n]$, $z : 2^{[n]} \to \mathbb{R}_{\geq 0}$, $h_i \; \forall i \in [n]$
Output: $\mathcal{I}_g$

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

Theorem 8 *Let $\mathcal{I}^*$ be an optimal solution to the BLDS problem. Algorithm 1 returns a solution $\mathcal{I}_g$ to the BLDS problem (i.e., (13)) that satisfies the following, where $\mathcal{I}_g^1, \ldots, \mathcal{I}_g^{T-1}$ are specified in Algorithm 1, and $[T-1] \triangleq \{1, 2, \ldots, T-1\}$.*

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

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

(c) $h(\mathcal{I}_g) \leq \left(1 + \ln \frac{z([n]) - z(\emptyset)}{z([n]) - z(\mathcal{I}_g^{T-1})}\right) h(\mathcal{I}^*)$,

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

Note that the bounds in Theorem 8(a)-(c) depend on $\mathcal{I}_g^t$ from the greedy algorithm. We can compute the bounds in Theorem 8(a)-(c) in parallel with the greedy algorithm, in order to provide a performance guarantee on the output of the algorithm. The bound in Theorem 8(d) does not depend on $\mathcal{I}_g^t$, and can be computed using $O(n)$ evaluations of function $z(\cdot)$.

3.1. Fast greedy algorithm

We now give an algorithm (Algorithm 2) for BLDS that achieves $O(\frac{n}{\epsilon} \ln \frac{n}{\epsilon})$ query complexity for any $\epsilon \in (0, 1)$, which is significantly smaller than $O(n^2)$ as n scales large. In line 3 of Algorithm 2, $h_{\max} \triangleq \max_{j \in [n]} h_j$ and $h_{\min} \triangleq \min_{j \in [n]} h_j$. While achieving faster running times, we will show that the solution returned by Algorithm 2 has slightly worse performance bounds compared to those of Algorithm 1 provided in Theorem 8, and potentially slightly violates the constraint of the BLDS problem given in (13). Specifically, a larger value of ϵ in Algorithm 2 leads to faster running times of Algorithm 2, but yields worse performance guarantees. Moreover, note that Algorithm 1 adds a single element to $\mathcal{I}_g$ in each iteration of the while loop in lines 2-4. In contrast, Algorithm 2 considers multiple candidate elements in each iteration of the for loop in lines 3-9, and adds elements that satisfy the threshold condition given in line 5, which leads to the faster running times. Formally, we have the following result.

Theorem 9 *Suppose $\frac{h_{\max}}{h_{\min}} \leq nH$ holds in the BLDS instances, where $h_{\max} = \max_{j \in [n]} h_j$, $h_{\min} = \min_{j \in [n]} h_j$, and $H \in \mathbb{R}_{\geq 1}$ is a fixed constant. Let $\mathcal{I}^*$ be an optimal solution to the BLDS problem. For any $\epsilon \in (0, 1)$, Algorithm 2 returns a solution $\mathcal{I}_f$ to the BLDS problem (i.e., (13)) in query complexity $O(\frac{n}{\epsilon} \ln \frac{n}{\epsilon})$ that satisfies $z(\mathcal{I}_f) \geq (1 - \epsilon) z([n])$, and has the following performance bounds, where $\mathcal{I}_f^{T-1}$ is given in Algorithm 2.*

(a) $h(\mathcal{I}_f) \leq \frac{1}{1-\epsilon}\Big(1+\ln\frac{z([n])}{z([n])-z(\mathcal{I}_f^{T-1})}\Big)h(\mathcal{I}^*)$,
(b) if $z(\mathcal{I}) \in \mathbb{Z}_{\geq 0}$ *for all* $\mathcal{I} \subseteq [n]$, $h(\mathcal{I}_f) \leq \frac{1}{1-\epsilon}\big(1+\ln z([n])\big)h(\mathcal{I}^*)$.

Algorithm 2 Fast Greedy Algorithm for DSSL

Input: $[n]$, $z : 2^{[n]} \to \mathbb{R}_{\geq 0}$, $h_i \; \forall i \in [n]$, $\epsilon \in (0,1)$
Output: $\mathcal{I}_f$

$t \leftarrow 0, \mathcal{I}_f^0 \leftarrow \emptyset$
$d \leftarrow \max_{i\in[n]} \frac{z(i)-z(\emptyset)}{h_i}$
for $(\tau = d; \tau \geq \frac{\epsilon h_{\min}}{n h_{\max}} d; \tau \leftarrow \tau(1-\epsilon))$ **do**
  for $j \in [n]$ **do**
    if $\frac{z(\mathcal{I}_f^t \cup \{j\}) - z(\mathcal{I}_f^t)}{h_j} \geq \tau$ **then**
      $\mathcal{I}_f^{t+1} \leftarrow \mathcal{I}_f^t \cup \{j\}, t \leftarrow t+1$
    if $z(\mathcal{I}_f^t) = z([n])$ **then**
      $T \leftarrow t, \mathcal{I}_f \leftarrow \mathcal{I}_f^T$
      return $\mathcal{I}_f$
$T \leftarrow t, \mathcal{I}_f \leftarrow \mathcal{I}_f^T$
return $\mathcal{I}_f$

Remark 10 *The threshold-based greedy algorithm has also been proposed for the problem of maximizing a monotone nondecreasing submodular function subject to a cardinality constraint (e.g., Badanidiyuru and Vondrák (2014)). The threshold-based greedy algorithm proposed in Badanidiyuru and Vondrák (2014) improves the running times of the standard greedy algorithm proposed in Nemhauser et al. (1978), and achieves a comparable performance guarantee to that of the standard greedy algorithm in Nemhauser et al. (1978). Here, we propose a threshold-based greedy algorithm (Algorithm 2) to solve the submodular set covering problem, which improves the running times of the standard greedy algorithm for the submodular set covering problem proposed in Wolsey (1982) (i.e., Algorithm 2), and achieves comparable performances guarantees as we showed in Theorem 9.*

3.2. Interpretation of Performance Bounds

Here, we give an illustrative example to interpret the performance bounds of Algorithm 1 and Algorithm 2 given in Theorem 8 and Theorem 9, respectively. In particular, we focus on the bounds given in Theorem 8(d) and Theorem 9(b). Consider an instance of the BLDS problem, where we set $\mu_0(\theta_p) = \frac{1}{m}$ for all $\theta_p \in \Theta$ with $m = |\Theta|$. In other words, there is a uniform prior belief over the states in $\Theta = \{\theta_1, \ldots, \theta_m\}$. Moreover, we set the error bounds $R_{\theta_p} = \frac{R}{m}$ for all $\theta_p \in \Theta$, where $R \in \mathbb{Z}_{\geq 0}$ and $R < m-1$. Recalling that $\bar{\Theta} = \{\theta_p \in \Theta : 0 \leq R_{\theta_p} < 1 - \mu_0(\theta_p)\}$ and noting the definition of $z(\cdot)$ in Eq. (12), for all $\mathcal{I} \subseteq [n]$, we define

$$z'(\mathcal{I}) \triangleq m(m-R)z(\mathcal{I}) = m(m-R)\sum_{\theta_p \in \Theta} f'_{\theta_p}(\mathcal{I}). \tag{14}$$

One can check that $z'(\mathcal{I}) \in \mathbb{Z}_{\geq 0}$ for all $\mathcal{I} \subseteq [n]$, and that (13) can be equivalently written as

$$\begin{aligned} &\min_{\mathcal{I}\subseteq[n]} h(\mathcal{I}) \\ &s.t.\; z'(\mathcal{I}) = z'([n]). \end{aligned} \tag{15}$$

Noting that $M' \triangleq \max_{j\in[n]} z'(j) \leq m^2(m-R)$ from (14), we then see from Theorem 8(d) that applying Algorithm 1 to (15) yields the following performance bound:

$$h(\mathcal{I}_g) \leq \Big(\sum_{i=i}^{M'} \frac{1}{i}\Big) h(\mathcal{I}^*) \leq \big(1+\ln M'\big)h(\mathcal{I}^*) \leq \big(1+2\ln m+\ln(m-R)\big)h(\mathcal{I}^*). \tag{16}$$

Similarly, since $z'([n]) \leq m^2(m-R)$ also holds, Theorem 9(b) implies the following performance bound for Algorithm 2 when applied to (15):

$$h(\mathcal{I}_f) \leq \frac{1}{1-\epsilon}\big(1+\ln z'([n])\big)h(\mathcal{I}^*) \leq \frac{1}{1-\epsilon}\big(1+2\ln m+\ln(m-R)\big)h(\mathcal{I}^*), \tag{17}$$

where $\epsilon \in (0,1)$. Again, we note from Theorem 9 that a smaller value of ϵ yields a tighter performance bound for Algorithm 2 (according to (17)) at the cost of slower running times. Thus, supposing m and ϵ are fixed, we see from (16) and (17) that the performance bounds of Algorithm 1 and Algorithm 2 become tighter as R increases, i.e., as the error bound R_{θ_p} increases. On the other hand, supposing R and ϵ are fixed, we see from (16) and (17) that the performance bounds of Algorithm 1 and Algorithm 2 become tighter as m decreases, i.e., as the number of possible states of the world decreases.

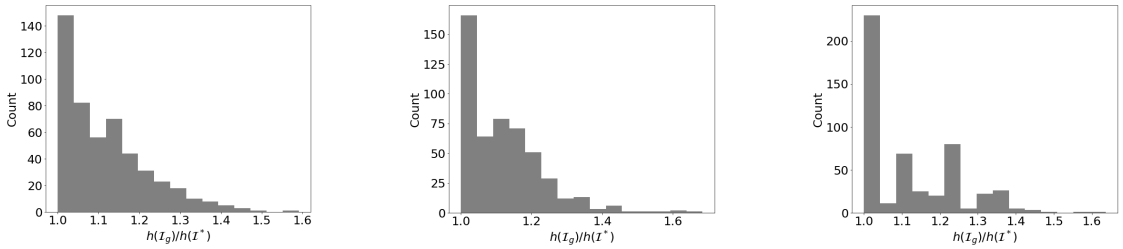
Finally, we note that the performance bounds given in Theorem 8 are worst-case performance bounds for Algorithm 1. Thus, in practice the ratio between a solution returned by the algorithm and an optimal solution can be smaller than the ratio predicted by Theorem 8. Nevertheless, there may also exist instances of the BLDS problem that let Algorithm 1 return a solution that meets the worst-case performance bound. Moreover, instances with tighter performance bounds (given by Theorem 8) potentially imply better performance of the algorithm when applied to those instances, as we can see from the above discussions and the numerical examples that will be provided in the next section. Therefore, the performance bounds given in Theorem 8 also provide insights into how different problem parameters of BLDS influence the actual performance of Algorithm 1. Similar arguments also hold for Algorithm 2 and the corresponding performance bounds given in Theorem 9.

3.3. Numerical examples

In this section, we focus on validating Algorithm 1 and the performance bounds provided in Theorem 8 using numerical examples. Numerical results for Algorithm 2 and Theorem 9 can be found in the extended version (Ye et al., 2020). First, the total number of data sources is set to be 10, and the selection cost h_i is drawn uniformly from $\{1,2,\ldots,10\}$ for all $i \in [n]$. The cost structure is then fixed in the sequel. Similarly to Section 3.2, we consider BLDS instances where $\mu_0(\theta_p) = \frac{1}{m}$ for all $\theta_p \in \Theta$ with $m = |\Theta|$, and $R_{\theta_p} = \frac{R}{m}$ for all $\theta_p \in \Theta$ with $R \in \mathbb{Z}_{>0}$ and $R < m-1$. Specifically, we set $m = 15$ and range R from 0 to 13. For each $R \in \{0,1,\ldots,13\}$, we further consider 500 corresponding randomly generated instances of the BLDS problem, where for each BLDS instance we randomly generate the set $F^c_{\theta_p}(i)$ (i.e., the set of states that can be distinguished from θ_p given data source i) for all $i \in [n]$ and for all $\theta_p \in \Theta$.[3] In Fig. 1, we plot histograms of the ratio $h(\mathcal{I}_g)/h(\mathcal{I}^*)$

3. Note that in the BLDS problem (Problem 1), the signal structure of each data source $i \in [n]$ is specified by the likelihood functions $\ell_i(\cdot|\theta_p)$ for all $\theta_p \in \Theta$. As we discussed in previous sections, (7) in Problem 1 can be equivalently written as (13), where one can further note that the function $z(\cdot)$ does not directly depend on any likelihood function $\ell_i(\cdot|\theta_p)$, and can be (fully) specified given $F^c_{\theta_p}(i)$ for all $i \in [n]$ and for all $\theta_p \in \Theta$. Thus, when constructing the BLDS instances in this section, we directly construct $F^c_{\theta_p}(i)$ for all $i \in [n]$ and for all $\theta_p \in \Theta$ in a random manner.

for $R = 1$, $R = 5$ and $R = 10$, where $\mathcal{I}_g$ is the solution returned by Algorithm 1 and $\mathcal{I}^*$ is an optimal solution to BLDS. We see from Fig. 1 that Algorithm 1 works well on the randomly generated BLDS instances, as the values of $h(\mathcal{I}_g)/h(\mathcal{I}^*)$ are close to 1. Moreover, we see from Fig. 1 that as R increases, Algorithm 1 yields better overall performance for the 500 randomly generated BLDS instances. Now, from the way we set $\mu_0(\theta_p)$ and R_{θ_p} in the BLDS instances constructed above, we see from the arguments in Section 3.2 that the performance bound for Algorithm 1 given by Theorem 8(d) can be written as $h(\mathcal{I}_g) \leq \big(1 + \ln M'\big)h(\mathcal{I}^*)$, where $M' = \max_{j\in[n]} z'(j)$ and $z'(\cdot)$ is defined in (14). Thus, in Fig. 2, we plot the performance bound of Algorithm 1, i.e., $1 + \ln M'$, for R ranging from 0 to 13. Also note that for each $R \in \{0, 1, \ldots, 13\}$, we obtain the averaged value of $1 + \ln M'$ over 500 random BLDS instances as we constructed above. We then see from Fig. 2 that the value of the performance bound of Algorithm 1 decreases, i.e., the performance bound becomes tighter, as R increases from 0 to 13. The behavior of the performance bound aligns with the actual performance of Algorithm 1 as we presented in Fig. 1, i.e., a tighter performance bound implies a better overall performance of the algorithm on the 500 random BLDS instances.



(a) $R = 1$. (b) $R = 5$. (c) $R = 10$.

Figure 1: Histograms of the ratio $h(\mathcal{I}_g)/h(\mathcal{I}^*)$.

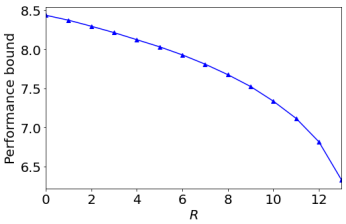


Figure 2: Performance bound for Algorithm 1 given by Theorem 8(d).

4. Conclusion

In this work, we considered the problem of data source selection for Bayesian learning. We first proved that the data source selection problem for Bayesian learning is NP-hard. Next, we showed that the data source selection problem can be transformed into an instance of the submodular set covering problem, and can then be solved using a standard greedy algorithm with provable performance guarantees. We also proposed a fast greedy algorithm that improves the running times of the standard greedy algorithm, while achieving comparable performance guarantees. We showed that the performance bounds provide insights into the actual performances of the algorithms under different instances of the data source selection problem. Finally, we validated our theoretical analysis using numerical examples, and showed that the greedy algorithms work well in practice.

References

Ashwinkumar Badanidiyuru and Jan Vondrák. Fast algorithms for maximizing submodular functions. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, pages 1497–1514, 2014.

Pierre Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer Science & Business Media, 2013.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

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.

Sanjoy Dasgupta. Analysis of a greedy active learning strategy. In *Proc. Advances in Neural Information Processing Systems*, pages 337–344, 2005.

Uriel Feige. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 45(4): 634–652, 1998.

Michael R Garey and David S Johnson. *Computers and intractability: a guide to the theory of NP-Completeness*. Freeman, 1979.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

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.

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

Ali Jadbabaie, Pooya Molavi, and Alireza Tahbaz-Salehi. Information heterogeneity and the speed of learning in social networks. *Columbia Business School Research Paper*, (13-28), 2013.

Igor Kononenko. Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal*, 7(4):317–337, 1993.

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.

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.

Jason Palmer, Bhaskar D Rao, and David P Wipf. Perspectives on sparse Bayesian learning. In *Proc. Advances in Neural Information Processing Systems*, pages 249–256, 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*, 66(5): 2146–2161, 2021.

Lintao Ye and Shreyas Sundaram. Sensor selection for hypothesis testing: Complexity and greedy algorithms. In *Proc. IEEE Conference on Decision and Control*, pages 7844–7849, 2019.

Lintao Ye, Aritra Mitra, and Shreyas Sundaram. Near-optimal data source selection for Bayesian learning. *arXiv preprint arXiv:2011.10712*, 2020.