Submodular Observation Selection and Information Gathering for Quadratic Models

Abolfazl Hashemi [1] **Mahsa Ghasemi** [1] **Haris Vikalo** [1] **Ufuk Topcu** [1]

Abstract

We study the problem of selecting most informative subset of a large observation set to enable accurate estimation of unknown parameters. This problem arises in a variety of settings in machine learning and signal processing including feature selection, phase retrieval, and target localization. Since for quadratic measurement models the moment matrix of the optimal estimator is generally unknown, majority of prior work resorts to approximation techniques such as linearization of the observation model to optimize the alphabetical optimality criteria of an approximate moment matrix. Conversely, by exploiting a connection to the classical Van Trees' inequality, we derive new alphabetical optimality criteria without distorting the relational structure of the observation model. We further show that under certain conditions on parameters of the problem these optimality criteria are monotone and (weak) submodular set functions. These results enable us to develop an efficient greedy observation selection algorithm uniquely tailored for quadratic models, and provide theoretical bounds on its achievable utility.

1. Introduction

In many machine learning applications, one needs to efficiently collect the most informative observations from a potentially significantly larger set of uncertain observations. The goal of such selection is to reduce the burden on computational and communication resources while still providing accurate inference of unknown parameters. For instance, to reduce the computational and communication costs in sensor and feature selection applications (Joshi and Boyd, 2009; Shamaiah et al., 2010; Bhaskara et al., 2016), the goal is to select a small representative subset of observations gathered by different sensing units within a network of autonomous systems. In sensor placement (Krause et al., 2008b), the problem is to select a subset of locations in an environment such that the measurements collected at those locations enable effective detection of unknown targets. In Bayesian experimental design, the objective is to select a subset of experiments from the set of all possible experiments to optimize a statistical metric or expected utility defined for a specific task such as inference or prediction of some parameters (Chaloner and Verdinelli, 1995; Wang et al., 2017; Chamon and Ribeiro, 2017). This selection is motivated by the desire to reduce experimental cost and abide to limited resources.

The example applications above belong to a family of problems referred to as observation selection or information gathering (Krause and Guestrin, 2007). They can be cast as optimization problems with an objective of maximizing the information of the selected observations subject to cardinality constraints. Since such optimization problems are generally NP-hard (Williamson and Shmoys, 2011; Krause and Golovin, 2014), one often resorts to heuristic schemes that find a sub-optimal subset of observations. It has been shown that, when the measurement model is linear, typical objective functions (e.g., expected utility) are scalarization of the error covariance matrix and possess a diminishing return property known as submodularity or weak (i.e. approximate) submodularity. For such objectives, a simple greedy approximation scheme achieves near-optimal observation selections with provable performance guarantees (Nemhauser et al., 1978; Krause and Golovin, 2014). Furthermore, since expected utilities are scalarizations of the error covariance matrix, the selection criteria for linear models have a strong connection to mean square error (MSE), the desired performance metric in many applications, including those considered in this paper. These two properties, that is (weak) submodularity of typical objective functions as well as a strong connection to MSE, makes linear models appealing. In fact when the measurement model is nonlinear, one typically resort to linearization of the model or Monte-Carlo methods to find an approximate utility prior to the actual observation selection step (Chaloner and Verdinelli, 1995; Sebastiani and Wynn, 2000; Flaherty et al., 2006; Krause

[1]University of Texas at Austin. Correspondence to: Abolfazl Hashemi <abolfazl@utexas.edu>, Mahsa Ghasemi <mahsa.ghasemi@utexas.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

et al., 2008a; Wang et al., 2016; Davidian, 2017). However, theoretical guarantees for the performance of greedy algorithms hold only for the linearized model, i.e., for the linear approximation of the actual nonlinear model. More importantly, due to approximation, the connection to MSE (if any) becomes less explicit and hence the selected subset of observations is generally not necessarily the most informative collection of observations.

A practically appealing class of inference tasks are those described by quadratic measurement models and inverse problems that occur in many natural phenomena and real-world applications. For instance, in object tracking and localization in robotics and autonomous systems, the range measurements gathered by radar systems follow a quadratic relation (Skolnik, 1970; Hightower and Borriello, 2001). In phase retrieval applications, where the goal is to recover an unknown object from its magnitude measurements, the relation between the unknown parameter and the magnitude measurements is described by a quadratic model (Fienup, 1982; Shechtman et al., 2015).

In this paper, we consider observation selection under models where the relation between the unknown parameters and the measurements follows a quadratic mapping with additive noise. By establishing a connection between the classical Van Trees' inequality (Van Trees, 2004) and alphabetical optimality criteria (Chaloner and Verdinelli, 1995), we devise new objective functions that exploit the quadratic relation of the observation model. Since the proposed objectives are built upon Van Trees' inequality, they enjoy an intuitive connection to MSE. In particular, by optimizing the proposed functions, one attempts to minimize a lower bound on MSE of any estimator of the unknown parameters. We further prove that these utility functions are monotone and (weak) submodular set functions under mild conditions on the statistics of the problem and parameters of the model. These results allow us to develop a simple greedy scheme for observation selection with theoretical bounds on its achievable utility without requiring any a priori approximation step. To demonstrate efficacy of the proposed framework, we consider two applications – multi-object tracking and phase retrieval – and empirically verify that the subsets selected by the proposed greedy algorithm outperform approaches based on random selection, and greedy selection of observations that relies on a linearized model.

2. Submodular Observation Selection

Consider the linear observation model $y_i = \mathbf{x}_i^\top \boldsymbol{\theta} + \nu_i$ where $\mathbf{X} \in \mathbb{R}^{n\times m}$ is the model parameter matrix, $\mathbf{y} \in \mathbb{R}^n$ is the collection of n observations, $\boldsymbol{\theta} \in \mathbb{R}^m$ denotes the unknown parameters, and $\nu_i \sim \mathcal{N}(\mathbf{0}, \sigma_i^2)$ is the additive Gaussian noise. Let $\mathcal{S}$ denote a subset of selected observations. Assuming a normal prior $p_{\boldsymbol{\theta}}(\boldsymbol{\Theta}) = \mathcal{N}(\mathbf{0}, \mathbf{P})$ on the unknown parameters, the minimum variance unbiased estimator of the parameters has a closed-form expression (Kay, 2013)

$$\hat{\boldsymbol{\theta}}_{\mathcal{S}} = \mathbf{M}_{\mathcal{S}} \left(\sum_{i\in\mathcal{S}} y_i \mathbf{x}_i \right), \tag{1}$$

where

$$\begin{aligned} \mathbf{M}_{\mathcal{S}} &:= \mathbb{E}[(\hat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta})^\top] \\ &= \left(\mathbf{P}^{-1} + \sum_{i\in\mathcal{S}} \frac{1}{\sigma_i^2} \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \end{aligned} \tag{2}$$

is the error covariance matrix of the estimated parameters $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$. Therefore, since the error covariance matrix for any subset of observations is known, one can use a suitable scalarization of the error covariance (also referred to as the moment matrix) to choose a subset of observations leading to the minimal estimation error (i.e., providing the highest information) for the inference task at hand. Typical utility functions are derived from the so-called alphabetical design criteria including A-optimality, D-optimality, E-optimality, and T-optimality (Chaloner and Verdinelli, 1995):

- In A-optimality, we minimize the trace of the error covariance matrix $\mathbf{M}_{\mathcal{S}}$, or equivalently maximize

$$f^A(\mathcal{S}) = \mathrm{Tr}\left(\mathbf{P}\right) - \mathrm{Tr}\left(\mathbf{M}_{\mathcal{S}}\right). \tag{3}$$

- In D-optimality, we minimize the log-det of the error covariance matrix $\mathbf{M}_{\mathcal{S}}$, or equivalently maximize

$$f^D(\mathcal{S}) = \log\det\left(\mathbf{M}_{\mathcal{S}}^{-1}\right) - \log\det\left(\mathbf{P}^{-1}\right). \tag{4}$$

- In E-optimality, we minimize the largest eigenvalue of the error covariance matrix $\mathbf{M}_{\mathcal{S}}$, λ_{max}, or equivalently maximize

$$f^E(\mathcal{S}) = \lambda_{min}\left(\mathbf{M}_{\mathcal{S}}^{-1}\right) - \lambda_{min}\left(\mathbf{P}^{-1}\right). \tag{5}$$

- In T-optimality, we maximize the trace of inverse of the error covariance matrix $\mathbf{M}_{\mathcal{S}}^{-1}$, or equivalently maximize

$$f^T(\mathcal{S}) = \mathrm{Tr}\left(\mathbf{M}_{\mathcal{S}}^{-1}\right) - \mathrm{Tr}\left(\mathbf{P}^{-1}\right). \tag{6}$$

If, for instance, one considers A-optimality as the utility, the task of selecting a subset of k out of n available observations can be formally cast as the following optimization problem

$$\max_{\mathcal{S}} \quad f^A(\mathcal{S}) \quad \text{s.t.} \qquad |\mathcal{S}| \leq k. \tag{7}$$

By a reduction to the set cover problem (Williamson and Shmoys, 2011), it has been shown that finding an optimal solution to (7) is NP-hard. Nonetheless, as we mentioned in Section 1, it has been shown that observation selection and

information gathering under linear models using the above optimality measures can be cast as the problem of maximizing a *monotone, (weak) submodular* function subject to cardinality (or uniform matroid) constraints (Nemhauser et al., 1978). These properties and their implications on solving the optimization problem (7) are stated next.

Definition 1. *Set function $f : 2^{\mathcal{X}} \to \mathbb{R}$ is monotone if $f(\mathcal{S}) \leq f(\mathcal{T})$ for all $\mathcal{S} \subseteq \mathcal{T} \subseteq \mathcal{X}$.*

Definition 2. *Set function $f : 2^{\mathcal{X}} \to \mathbb{R}$ is submodular if*

$$f(\mathcal{S} \cup \{j\}) - f(\mathcal{S}) \geq f(\mathcal{T} \cup \{j\}) - f(\mathcal{T}) \tag{8}$$

for all subsets $\mathcal{S} \subseteq \mathcal{T} \subset \mathcal{X}$ and $j \in \mathcal{X} \backslash \mathcal{T}$. The term $f_j(\mathcal{S}) = f(\mathcal{S} \cup \{j\}) - f(\mathcal{S})$ is the marginal value of adding element j to set $\mathcal{S}$.

Definition 3. *The multiplicative weak-submodularity constant of a monotone non-decreasing function f is defined as*

$$c_f = \max_{(\mathcal{S},\mathcal{T},i) \in \tilde{\mathcal{X}}} f_i(\mathcal{T}) / f_i(\mathcal{S}), \tag{9}$$

where $\tilde{\mathcal{X}} = \{(\mathcal{S}, \mathcal{T}, i) | \mathcal{S} \subseteq \mathcal{T} \subset \mathcal{X}, i \in \mathcal{X} \backslash \mathcal{T}\}$.

The multiplicative weak-submodularity constant (Zhang and Vorobeychik, 2016; Chamon and Ribeiro, 2017) is a closely related concept to submodularity and essentially quantifies how close the set function is to being submodular. It is worth noting that a set function $f(\mathcal{S})$ is submodular if and only if its multiplicative weak-submodularity constant satisfies $c_f \leq 1$ (Das and Kempe, 2011; Elenberg et al., 2018; Horel and Singer, 2016).

A similar notion of weak submodularity is the additive weak-submodularity constant as defined below (Zhang and Vorobeychik, 2016; Chamon and Ribeiro, 2017).

Definition 4. *The additive weak-submodularity constant of a monotone non-decreasing function f is defined as*

$$\epsilon_f = \max_{(\mathcal{S},\mathcal{T},i) \in \tilde{\mathcal{X}}} f_i(\mathcal{T}) - f_i(\mathcal{S}), \tag{10}$$

where $\tilde{\mathcal{X}} = \{(\mathcal{S}, \mathcal{T}, i) | \mathcal{S} \subseteq \mathcal{T} \subset \mathcal{X}, i \in \mathcal{X} \backslash \mathcal{T}\}$.

Note that when $f(\mathcal{S})$ is submodular, its additive weak-submodularity constant satisfies $\epsilon_f \leq 0$.

For a monotone function with bounded additive and multiplicative weak-submodularity constants (WSCs) we have the following proposition. [1]

Proposition 1. *Let c_f and ϵ_f be the multiplicative and additive weak-submodularity constants of $f(\mathcal{S})$, a monotone non-decreasing function with $f(\emptyset) = 0$. Let $\mathcal{S}$ and $\mathcal{T}$ be any subsets such that $\mathcal{S} \subset \mathcal{T} \subseteq \mathcal{X}$ with $|\mathcal{T} \backslash \mathcal{S}| = r$. Then, it holds that*

$$f(\mathcal{T}) - f(\mathcal{S}) \leq \frac{1}{r} \left(1 + (r-1)c_f\right) \sum_{j \in \mathcal{T} \backslash \mathcal{S}} f_j(\mathcal{S}), \tag{11}$$

and

$$f(\mathcal{T}) - f(\mathcal{S}) \leq (r-1)\epsilon_f + \sum_{j \in \mathcal{T} \backslash \mathcal{S}} f_j(\mathcal{S}). \tag{12}$$

[1] Detailed proofs of all lemmas, propositions, and theorems are stated in the supplementary.

Algorithm 1 Greedy Observation Selection

1: **Input:** Utility function $f(\mathcal{S})$, set of all observations $\mathcal{X}$, number of selected observations k.
2: **Output:** Subset $\mathcal{S}_g \subseteq \mathcal{X}$ with $|\mathcal{S}_g| = k$.
3: Initialize $\mathcal{S}_g = \emptyset$
4: **for** $i = 0, \ldots, k-1$ **do**
5: $\quad j_s = \text{argmax}_{j \in \mathcal{X} \backslash \mathcal{S}_g} f_j(\mathcal{S}_g)$
6: $\quad \mathcal{S}_g \leftarrow \mathcal{S}_g \cup \{j_s\}$
7: **end for**

It has been shown in (Nemhauser et al., 1978) that if a set function is monotone and submodular, a simple greedy algorithm that iteratively select an observation with the highest marginal gain (see Algorithm 1) satisfies a $1 - 1/e$ approximation factor. Using Proposition 1, one can extend these theoretical results to the case of weak submodular functions, as illustrated in the following proposition (Das and Kempe, 2011; Chamon and Ribeiro, 2017; Elenberg et al., 2018; Hashemi et al., 2018).

Proposition 2. *Let c_f and ϵ_f be the multiplicative and additive weak-submodularity constants of $f(\mathcal{S})$, a monotone non-decreasing function with $f(\emptyset) = 0$. Let $\mathcal{S}_g \subseteq \mathcal{X}$ with $|\mathcal{S}_g| \leq k$ be the subset selected when maximizing $f(\mathcal{S})$ subject to a cardinality constraint via the greedy observation selection scheme, and let $\mathcal{S}^\star$ denote the optimal subset. Then*

$$f(\mathcal{S}_g) \geq \left(1 - e^{-\frac{1}{c}}\right) f(\mathcal{S}^\star), \tag{13}$$

where $c = \max\{c_f, 1\}$ and

$$f(\mathcal{S}_g) \geq \left(1 - \frac{1}{e}\right) \left(f(\mathcal{S}^\star) - (k-1)\epsilon_f\right). \tag{14}$$

The results of Propositions 1 and 2 imply that if the objective function of observation selection task (see (7)) is monotone and (weak) submodular, the greedy selection scheme that in each iteration selects an observation with the highest marginal gain satisfies the approximation bounds given in Proposition 2. Indeed, it has been shown that when the observation model is linear, D-optimality criterion is submodular (Krause et al., 2008b; Shamaiah et al., 2010), T-optimality criterion is modular (Krause et al., 2008b; Summers et al.,

2016) while A-optimality and E-optimality measures are weak submodular (Bian et al., 2017; Chamon and Ribeiro, 2017; Hashemi et al., 2018).

When the model is nonlinear, which as we discussed before is frequently encountered in many applications including phase retrieval and localization, finding the posterior distribution of the unknown parameter (and hence the error covariance matrix $\mathbf{M}_{\mathcal{S}}$ associated with an estimator $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$) becomes intractable in general. Existing approaches mainly rely on heuristic methods to approximate the expected utility, for instance, by linearizing the model or employing Monte-Carlo methods (Chaloner and Verdinelli, 1995; Sebastiani and Wynn, 2000; Flaherty et al., 2006; Krause et al., 2008a; Davidian, 2017). For instance, in locally-optimal observation selection approach in experimental design (Flaherty et al., 2006; Krause et al., 2008a) for a nonlinear model $y_i = g_i(\boldsymbol{\theta}) + \nu_i$, one linearizes the model around an initial guess $\boldsymbol{\theta}_0$ (e.g. $\boldsymbol{\theta}_0 = \mathbb{E}[\boldsymbol{\theta}]$) to obtain

$$\tilde{y}_i := y_i - g_i(\boldsymbol{\theta}_0) = \nabla g_i(\boldsymbol{\theta}_0)^\top \boldsymbol{\theta} + \nu_i, \tag{15}$$

find an approximate moment

$$\hat{\mathbf{M}}_{\mathcal{S}} = \left(\mathbf{P}^{-1} + \sum_{i \in \mathcal{S}} \frac{1}{\sigma_i^2} \nabla g_i(\boldsymbol{\theta}_0) \nabla g_i(\boldsymbol{\theta}_0)^\top \right)^{-1}, \tag{16}$$

and use alphabetical scalarization of $\hat{\mathbf{M}}_{\mathcal{S}}$ as the optimality measure to select the subset $\mathcal{S}_g$ of observation via the greedy selection scheme. Notice that $\hat{\mathbf{M}}_{\mathcal{S}}$ is no longer equivalent to the error covariance (i.e. moment) matrix. Therefore, optimizing scalarization of $\hat{\mathbf{M}}_{\mathcal{S}}$ does not necessarily result in selecting a low MSE subset. Additionally, existing theoretical results established for the performance of the greedy algorithm hold for the approximate, linearized model and not for the original nonlinear observation model.

In contrast to the existing observation selection methods for nonlinear models that rely on finding an approximate moment matrix, our proposed framework builds upon the idea of optimizing alphabetical scalarizations of the Van Trees' bound (Van Trees, 2004) on the moments of a weakly biased estimator. As we will see in the subsequent sections, these *surrogate utilities* have an intuitive connection to MSE while enjoying (weak) submodularity. The Van Trees' inequality is outlined in the following theorem.

Theorem 1. *Let $\boldsymbol{\theta}$ be a collection of random unknown parameters, and let $\mathbf{y}_{\mathcal{S}} = \{y_i\}_{i \in \mathcal{S}}$ denote the collection of measurements indexed by the subset $\mathcal{S}$. For any estimator $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$ that satisfies*

$$\int_{-\infty}^{+\infty} \nabla_{\boldsymbol{\Theta}} \left(p_{\boldsymbol{\theta}}(\boldsymbol{\Theta}) \, \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}}[\hat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\Theta}] \right) d\boldsymbol{\Theta} = \mathbf{0}, \tag{17}$$

it holds that

$$\mathbf{M}_{\mathcal{S}} \succeq \mathbb{E}_{\mathbf{y}_{\mathcal{S}},\boldsymbol{\theta}} \left[(\nabla_{\boldsymbol{\Theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{\Theta}))(\nabla_{\boldsymbol{\Theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{\Theta}))^\top \right]^{-1}, \tag{18}$$

where $q_{\boldsymbol{\theta}}(\boldsymbol{\Theta}) = p_{\boldsymbol{\theta},\mathbf{y}_{\mathcal{S}}}(\boldsymbol{\Theta}; \mathbf{y})$ is the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{y}_{\mathcal{S}}$.

The condition stated in Theorem 1 essentially quantifies to what extend the estimator is biased. Indeed, for an unbiased estimator satisfying $\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}}[\hat{\boldsymbol{\theta}}_{\mathcal{S}}] = \boldsymbol{\theta}$, this condition is met.

The lower bound in the Van Trees' inequality cannot be computed in a closed-form for general nonlinear models. Hence, we focus our study on quadratic models since, as we show in the next section, the Van Trees' bound has a closed-form expression.

3. Observation Selection for Quadratic Models

Consider the quadratic model

$$y_i = \frac{1}{2} \boldsymbol{\theta}^\top \mathbf{X}_i \boldsymbol{\theta} + \mathbf{z}_i^\top \boldsymbol{\theta} + \nu_i \,, \quad i \in [n], \tag{19}$$

where $\mathbf{X}_i$ and $\mathbf{z}_i$ are known parameters (e.g., the features or the design parameters), $\nu_i \sim \mathcal{N}(0, \sigma_i^2)$, and $\boldsymbol{\theta}$ denotes the unknown collection of parameters with prior $p_{\boldsymbol{\theta}}(\boldsymbol{\Theta})$. We further assume the expectation is known and $\mathbb{E}[\boldsymbol{\theta}] = \mathbf{0}$; alternatively, we can make $\boldsymbol{\theta}$ centered.

Our first theoretical result, stated in Theorem 2, demonstrates that for the quadratic model in (19) and for any prior on $\boldsymbol{\theta}$ with covariance matrix $\mathbf{P}$, the Van Trees' bound has a closed-form expression.

Theorem 2. *Let $\mathbf{B}_{\mathcal{S}}$ denote the lower bound in the Van Trees' inequality for the quadratic model* (19)*. Let $p_{\boldsymbol{\theta}}(\boldsymbol{\Theta})$ be the prior on $\boldsymbol{\theta}$ such that $\mathbb{E}[\boldsymbol{\theta}] = \mathbf{0}$, and $\mathrm{Cov}(\boldsymbol{\theta}) = \mathbf{P}$. Then*

$$\mathbf{B}_{\mathcal{S}} = \left(\sum_{i \in \mathcal{S}} \frac{1}{\sigma_i^2} \left(\mathbf{X}_i \mathbf{P} \mathbf{X}_i^\top + \mathbf{z}_i \mathbf{z}_i^\top \right) + \mathbf{I}_x \right)^{-1}, \tag{20}$$

where

$$\mathbf{I}_x = \mathbb{E}_{\boldsymbol{\theta}} \left[(\nabla_{\boldsymbol{\Theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{\Theta}))(\nabla_{\boldsymbol{\Theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{\Theta}))^\top \right] \tag{21}$$

is the Fisher information matrix associated with $p_{\boldsymbol{\theta}}(\boldsymbol{\Theta})$.

Theorem 2 opens a new avenue in the task of observation selection and information gathering for quadratic models which, as we see in our simulation results, enables selection of observations leading to lower estimation error (i.e., higher information) as compared to the approximate method based on linearization. Relying on the result of Theorem 2, we propose to use alphabetical optimality criteria (see (3) – (6)) applied to the Van Trees' lower bound $\mathbf{B}_{\mathcal{S}}$ as the utility function in the observation selection task (effectively replacing $\mathbf{M}_{\mathcal{S}}$ and $\mathbf{P}^{-1}$ with $\mathbf{B}_{\mathcal{S}}$ and $\mathbf{I}_x$, respectively). Since these utility functions aim to minimize a scalarization

of Van Trees' lower bound on MSE, they have an explicit connection to MSE. We note however that similar to locally-optimal observation selection, the proposed utilities are in fact surrogates to MSE and ultimately heuristic in nature. However, due to the connection established by Theorem 1, they result in selection of a more informative subset of observations compared to other heuristics. Nonetheless, the Van Trees lower bound is asymptotically tight, i.e., it is tight in the high signal-to-noise ratio settings or in the case of sufficiently large number of observations. Hence, we expect to select a near-optimal subset by using the proposed selection criteria in such settings.

Compared to existing robust methods, (Flaherty et al., 2006) consider robustness with respect to the Jacobian of the unknown parameters in the linearized model and apply semi-definite programming with E-optimality criterion. (Krause et al., 2008a) consider robustness with respect to the initial guess. They do so by performing linearization over multiple starting point and use the saturated algorithm for selection. In large-scale problems one needs to consider many starting points which is impractical. Since for quadratic models we can efficiently find the Van Trees' inequality without any approximation, it is meaningful to utilize such structural property. However, for general nonlinear models, robust methods such as (Flaherty et al., 2006; Krause et al., 2008a) are indeed advantageous.

The question that remains to be answered is whether the greedy observation selection method that provides a provably near-optimal selection for the linear models still enjoys similar theoretical performance guarantees. To this end, we demonstrate such theoretical performance guarantees for the greedy algorithm by showing weak submodularity of the proposed optimality criteria.

4. Near-Optimal Greedy Observation Selection

In the following theorems we consider alphabetical scalarizations of the Van Trees' bound $\mathbf{B}_S$ defined in Theorem 2 and show that they are monotonically increasing as well as either modular, submodular, or weak submodular. These results illustrate not only that the proposed optimality criteria deal with the quadratic model without resorting to any approximations, but also that one can use the greedy observation selection method of Algorithm 1 to find a near-optimal subset of observations with performance guarantees established in Proposition 2. Proofs of the subsequent results are established by employing tools from linear algebra and matrix analysis such as Weyl's inequality, Sylvester's determinant identity, matrix inversion lemma, and Courant–Fischer min-max theorem (Bellman, 1997).

Theorem 3. *Instate the notation and hypothesis of Theorem 2. The T-optimality of the Van Trees' bound, i.e.,*

$$f^T(\mathcal{S}) = \operatorname{Tr}\left(\mathbf{B}_{\mathcal{S}}^{-1}\right) - \operatorname{Tr}\left(\mathbf{I}_x\right), \tag{22}$$

is monotone and modular.

Theorem 4. *Instate the notation and hypothesis of Theorem 2. The D-optimality of the Van Trees' bound, i.e.,*

$$f^D(\mathcal{S}) = \log\det\left(\mathbf{B}_{\mathcal{S}}^{-1}\right) - \log\det\left(\mathbf{I}_x\right), \tag{23}$$

is monotone and submodular.

Theorem 5. *Instate the notation and hypothesis of Theorem 2. The E-optimality of the Van Trees' bound, i.e.,*

$$f^E(\mathcal{S}) = \lambda_{\min}\left(\mathbf{B}_{\mathcal{S}}^{-1}\right) - \lambda_{\min}\left(\mathbf{I}_x\right), \tag{24}$$

is monotone and weak submodular. Further, its additive and multiplicative weak-submodularity constants satisfy

$$\begin{aligned} c_{f^E} &\leq \max_{j\in\mathcal{X}} \frac{\lambda_{\max}\left(\mathbf{I}_j\right)}{\lambda_{\min}\left(\mathbf{I}_j\right)}, \\ \epsilon_{f^E} &\leq \max_{j\in\mathcal{X}} \left(\lambda_{\max}\left(\mathbf{I}_j\right) - \lambda_{\min}\left(\mathbf{I}_j\right)\right), \end{aligned} \tag{25}$$

where $\mathbf{I}_j = \frac{1}{\sigma_j^2}\left(\mathbf{X}_j\mathbf{P}\mathbf{X}_j^\top + \mathbf{z}_j\mathbf{z}_j^\top\right)$.

The term $\mathbf{I}_j$ is reflective of the amount of *information* captured by the j^{th} observation. In this regard, Theorem 5 states that if the difference between the minimum and maximum information of individual observations is small, the E-optimality of $\mathbf{B}_{\mathcal{S}}$ is nearly submodular. Hence, the greedy observation selection scheme is expected to find a good (informative) subset.

Theorem 6. *Instate the notation and hypothesis of Theorem 2. The A-optimality of the Van Trees' bound, i.e.,*

$$f^A(\mathcal{S}) = \operatorname{Tr}\left(\mathbf{I}_x^{-1}\right) - \operatorname{Tr}\left(\mathbf{B}_{\mathcal{S}}\right), \tag{26}$$

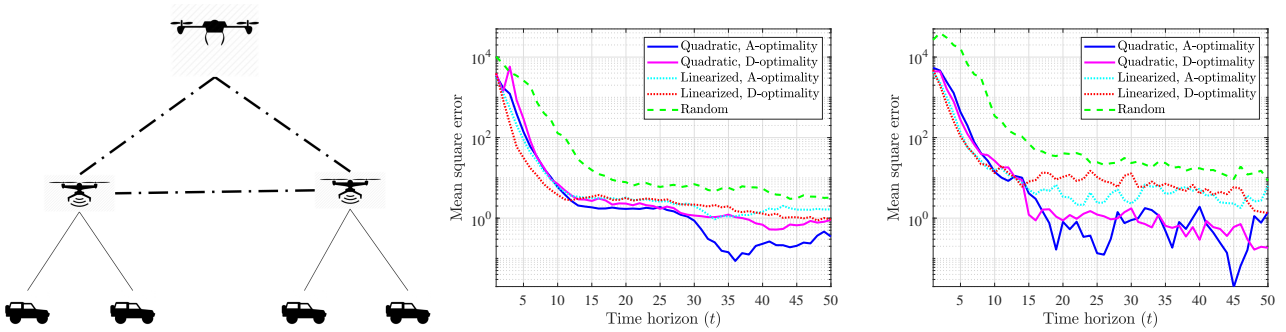
is monotone and weak submodular. Furthermore, if $\mathbf{z}_i = \mathbf{0}$ *and* $\mathbf{X}_i = \mathbf{x}_i\mathbf{x}_i^\top$*, the additive and multiplicative weak-submodularity constants satisfy*

$$c_{f^A} \leq \max_j \gamma_j, \quad \epsilon_{f^A} \leq \max_j \frac{\lambda_{\min}(\mathbf{B}_{[n]})^2}{\lambda_{\max}(\sigma_j^2\mathbf{P} + \mathbf{I}_x^{-1})}(\gamma_j - 1) \tag{27}$$

where

$$\gamma_j = \frac{\lambda_{\max}(\mathbf{I}_x^{-1})^2(\lambda_{\max}(\sigma_j^2\mathbf{P}) + 1)}{\lambda_{\min}(\mathbf{B}_{[n]})^2(\lambda_{\min}(\sigma_j^2\mathbf{P}) + 1)}. \tag{28}$$

The conditions on additive and multiplicative WSCs of A-optimality in Theorem 6 are essentially conditions on observations' SNR. Specifically, γ_j can be interpreted as the normalized SNR of the i^{th} observation. In this regard, according to (27) the behavior of the A-optimality approaches that of a submodular function if the energy of observation with highest SNR is relatively small. Additionally, if observations are fairly uncorrelated, conditions in (27) are met even if the SNRs are large. Furthermore, we note that the simplifying condition $\mathbf{z}_i = \mathbf{0}$ and $\mathbf{X}_i = \mathbf{x}_i\mathbf{x}_i^\top$ is motivated by the phase retrieval application, studied in Section 5.



(a) tracking via a swarm of UAVs (b) tracking under identical noise powers (c) tracking under random noise powers

Figure 1. Comparison of MSEs for random, linearized, and quadratic observation selection schemes in the multi-target tracking application.

5. Applications and Numerical Tests

In this section we test the efficacy of the proposed quadratic observation selection optimality criteria in two applications: multi-object tracking using radar measurements and phase retrieval from magnitude measurements. We limit our experiments to A-optimality and D-optimality for their intuitive interpretation: A-optimality of $\mathbf{B}_S$ for a minimum variance unbiased estimator attaining (18) with equality is equivalent to the mean-square error (MSE), while, in the context of parameter estimation in linear models with independent and homoscedastic noise, D-optimality is equivalent to the maximization of entropy of parameters (Krause et al., 2008b).

5.1. Constrained multi-target tracking

We study a multi-object tracking application where a control unit surveys an area via a swarm of unmanned aerial vehicles (UAVs) (Figure 1(a)). The UAVs are equipped with GPS and radar systems, and can communicate with each other over locally established communication channels. Due to limitations on the rate of communication between the *swarm leaders* and the control unit and in order to reduce delays in tracking due to intensive computation, only a subset of the gathered measurements is communicated to the control unit. In order to track the locations of the objects in the environment, the control unit employs extended Kalman filter (EKF) to process the received measurements. Therefore, the goal of the swarm leaders is to perform *observation selection and information gathering* and select a subset of range and angular measurements such that (i) the communication constraint is satisfied, and (ii) the mean-square error of the EKF estimate of the objects' locations is minimized. Let $\mathbf{u}_i^t$ and $\mathbf{s}_j^t$ denote the location of the i^{th} UAV and the j^{th} object, respectively.[2] The range measurements of the radar system follow a quadratic model:

$$r_{ij} = \frac{1}{2}\|\mathbf{u}_i^t - \mathbf{s}_j^t\|_2^2 + \nu_{ij}. \tag{29}$$

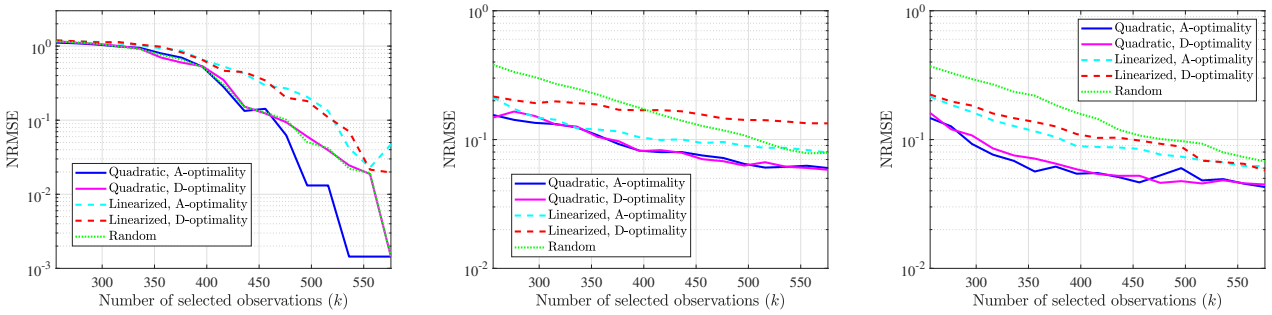
[2]For simplicity, objects and UAVs' altitudes are fixed.

A comparison with (19) reveals that, in this model, (relative to each pair of UAV-target) $\mathbf{X}_i$'s are all equal to the identity matrix and $\mathbf{z}_i = \mathbf{u}_i^t - \hat{\mathbf{s}}_j^{t-1}$ where $\hat{\mathbf{s}}_j^{t-1}$ is the prior estimates of objects' locations (after centering $\mathbf{s}_j^t$ around its approximate expectation $\mathbb{E}[\mathbf{s}_j^t] \approx \hat{\mathbf{s}}_j^{t-1}$). Therefore, we can employ the proposed quadratic observation selection framework directly. The angular measurements on the other hand have the nonlinear form

$$\alpha_{ij} = \arctan\frac{u_i^t(1) - s_j^t(1)}{u_i^t(2) - s_j^t(2)} + \eta_{ij}. \tag{30}$$

To select the angular measurements, we follow the locally optimal approach of (Flaherty et al., 2006) and linearize (30) around the prior estimates of objects' locations $\hat{\mathbf{s}}_j^{t-1}$.

We implement a Monte Carlo simulation with 50 independent instances where 10 moving objects are initially uniformly distributed in a 5×10 area. At each time instance, the objects move in a random direction with a constant velocity set to 0.2. The swarm consists of 10 UAVs, equidistantly spread over the area, that move according to a periodic *parallel-path* search pattern (Vincent and Rubin, 2004). The initial phases of the UAVs' motions are uniformly distributed to provide a better coverage of the area. The UAVs can acquire range and angular measurements of the objects that are within the maximum radar detection range. The maximum radar detection range is set such that, at each time step, the UAVs together collect approximately 130-170 range and angular measurements. The communication bandwidth constraints limit the number of measurements transmitted to the control unit to 10% of the gathered measurements. We consider two noise models: in the first scenario the noise terms $\nu_{ij}, i = 1, \ldots, 10, j = 1, \ldots, 10$ are i.i.d. Gaussian noises with $\sigma_{ij} = 0.01$ while in the second scenario we logarithmically space the interval $(0.001, 0.01)$ to generate 10 points and select σ_{ij} for each observation uniformly at random from one of these 10 numbers. We use A-optimality and D-optimality as the selection criteria and assess the performance of different schemes using the MSE of the EKF estimates of objects' locations.



(a) Gaussian with identical noise powers (b) DFT with identical noise powers (c) DFT with random noise powers

Figure 2. Comparison of NRMSEs for random, linearized, and quadratic observation selection schemes for the phase retrieval.

Figure 1(b) and Figure 1(c) illustrate the results for the two noise models. At the beginning of tracking, all schemes have relatively high error. However, since the observations selected by the proposed schemes are chosen according to the exact range model, as time passes the MSE of the proposed scheme becomes significantly lower than those of locally optimal and random selection methods (especially under the A-optimality criterion). Figure 1 also depicts that the MSE of the estimates formed from the observations selected by the proposed quadratic observation selection scheme using A-optimality is lower than the MSE achieved by selecting the observation via D-optimality. This reduction is due to the fact that if the estimator (here the EKF) is a minimum variance unbiased estimator attaining (18) with equality, the A-optimality scalarization of the Van Trees' bound becomes equivalent to the MSE, the performance measure shown in Figure 1. Therefore, intuitively one expects to achieve lower MSE using the A-optimality scalarization of the Van Trees' bound, which is the case in this experiment.

5.2. Phase retrieval from magnitude measurements

Phase retrieval is the task of predicting a possibly complex unknown variable from its magnitude measurements. Specifically, we are given measurements (Candes et al., 2015a;b)

$$y_i = g_i(\boldsymbol{\theta}) = \frac{1}{2}|\mathbf{a}_i^*\boldsymbol{\theta}|^2 + \nu_i, \tag{31}$$

where $\mathbf{a}_i \in \mathbb{C}^n$ is the i^{th} measurement vectors, and $*$ denotes the conjugate transpose operator. The model in (31) is an instance of the quadratic model (19) as it can be written as $y_i = \frac{1}{2}\boldsymbol{\theta}^*(\mathbf{a}_i\mathbf{a}_i^*)\boldsymbol{\theta} + \nu_i$. However, since $g_i(\boldsymbol{\theta})$ is not holomorphic (i.e. complex differentiable) in general, the results of Theorem 2 cannot be applied directly. Therefore, we approximate the Van Trees' bound by using the partial Wirtinger derivative of $g_i(\mathbf{x})$. That is, we treat $\boldsymbol{\theta}$ as a real-valued variable and compute the gradient of $g_i(\boldsymbol{\theta})$ to obtain

$$\begin{aligned}\hat{\mathbf{B}}_{\mathcal{S}} &= \left(\sum_{i\in\mathcal{S}} \frac{1}{\sigma_i^2}\mathbf{a}_i\mathbf{a}_i^*\mathbf{P}\mathbf{a}_i\mathbf{a}_i^* + \mathbf{I}_x\right)^{-1} \\ &= \left(\sum_{i\in\mathcal{S}} \frac{1}{\tilde{\sigma}_i^2}\mathbf{a}_i\mathbf{a}_i^* + \mathbf{I}_x\right)^{-1},\end{aligned} \tag{32}$$

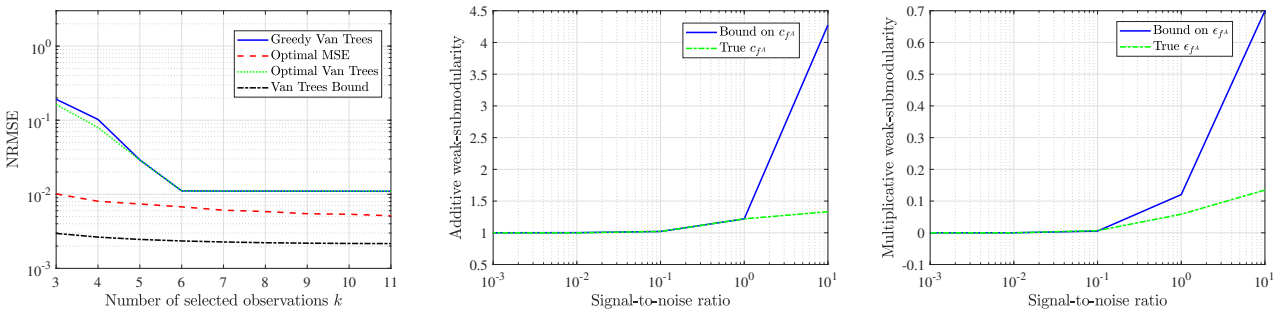
where $\tilde{\sigma}_i^2 = \sigma_i^2/(\mathbf{a}_i^*\mathbf{P}\mathbf{a}_i)$. The expression for $\hat{\mathbf{B}}_{\mathcal{S}}$ in (32) reveals an interesting *noise adjustment* effect. Because of the specific structure of the quadratic model in phase retrieval, i.e., $\mathbf{X}_i$'s are of rank 1, the (approximate) Van Trees' bound resembles the structure of the moment matrix in linear models except that the noise powers are now observation-specific and are related to $\boldsymbol{\theta}$ through the covariance matrix $\mathbf{P}$.

In order to investigate the performance of the proposed observation selection framework, we consider the task of predicting a complex signal $\boldsymbol{\theta} \in \mathbb{C}^n$ with $n = 128$ by selecting a subset of size k from $m = 1280$ observations. The complex signal $\boldsymbol{\theta}$ is distributed as standard circularly-symmetric complex Gaussian, and we vary k from 256 to 576. We consider two standard scenarios where for each we average the results over 50 independent instances, and use the Wirtinger flow algorithm (Candes et al., 2015b) as the oracle estimator.

5.2.1. Complex Gaussian measurements

First we examine the case in which all the elements in the measurement vectors $\mathbf{a}_i$ are independent random variables following a standard complex Gaussian distribution. Thus every element is regarded as a random variable in which the real part and the imaginary part are drawn from the standard Gaussian distribution $\mathcal{N}(0, 1)$ independently. We also assume that the noise powers are $\sigma_i = 0.001$ for all measurements.

Figure 2 (a) illustrates the normalized root MSE (NRMSE) results. The subset selected to achieve the A-optimality of $\hat{\mathbf{B}}_{\mathcal{S}}$ provides the lowest estimation error. This observation is



(a) MSE vs Van Trees' bound (b) additive WSC (c) multiplicative WSC

Figure 3. Evaluation of theoretical results in Section 4 for a small-scale phase retrieval task.

again supported by the relation of A-optimality and MSE for minimum variance unbiased estimators satisfying (18) with equality. The performance of subset selected to achieve the D-optimality of $\hat{\mathbf{B}}_{\mathcal{S}}$ and that of random subset selection are virtually the same, which is also expected since D-optimality is intuitively achieved by selecting observations according to the entropy gains. Since $\mathbf{a}_i$'s are i.i.d. and the noise powers are equal for all measurements, the entropy gain is similar for all measurements. Hence, D-optimality acts relatively similarly to random subset selection in this scenario. Locally optimal observation selection schemes achieve the lowest performance, partly because of the sensitivity of these schemes to the initial estimate used in linearization.

5.2.2. Discrete Fourier measurements

In traditional phase retrieval problems, the measurements are the magnitude of the Fourier transform of the signal. Hence, next we consider the $\mathbf{a}_i$'s to be rows of an $m \times m$ DFT matrix restricted to the first n columns. Figure 2 (b) shows the NRMSE results, where, as we see, the proposed quadratic observation selection framework achieves the lowest error. Note that, in contrast to the case of complex Gaussian measurement vectors, the entropy gain of the measurements is expected to be different under this scenario. Therefore, D-optimality of $\hat{\mathbf{B}}_{\mathcal{S}}$ results in a superior performance compared to the random observation selection scheme. Furthermore, since selecting more observations reduces the estimation error, as k increases, the performance gap between different schemes decreases.

Finally, we consider a case where the noise powers are not identical for all measurements. Rather, we logarithmically space the interval $(0.001, 0.01)$ to generate 10 points and select σ_i for each observation uniformly at randomly from one of these 10 numbers. Figure 2 (c) depicts the NRMSE results, where we again observe the superiority of the proposed optimality criteria to select a subset of observation with the lowest estimation error.

5.3. Evaluation of theoretical results

To numerically study the connection between Van Trees' bound and MSE, we consider a phase retrieval problem with $m = 12$ Gaussian observations where we find the optimal subsets via exhaustive search. Figure 3 (a) shows comparison of NRMSE of greedy maximization of A-optimality of Van Trees' bound, maximization of A-optimality of Van Trees' bound via exhaustive search, and maximization of MSE via exhaustive search. We also plot the theoretical Van Trees' bound (see (20)) evaluated at the set found by the approach of maximization of MSE via exhaustive search. As we see, our approach performs similarly to the exhaustive Van Trees' bound optimization which shows near optimality of the greedy selection scheme. Also, as we select more observations all three selection schemes approach the theoretical Van Trees' bound, which is supported by the asymptotic exactness of Van Trees' bound.

Figure 3(b) and Figure 3(c) show the true values of additive and multiplicative WSCs found by exhaustive search as well as our established bounds for A-optimality of Van Trees' bound in Theorem 6. As we see, with smaller SNRs, the gap between the two is negligible however for larger SNRs the established bounds are relatively weak.

6. Conclusion

We studied the task of observation selection and information gathering for quadratic measurement models. Since the moment matrix of the optimal estimator is unknown due to nonlinearity of the observation model, typical optimality criteria no longer possess connections to MSE, the desired performance metric. To address this challenge, we derived new criteria by relying on the Van Trees' inequality and proved that they are monotone (weak) submodular set functions under certain conditions on the unknown parameters and the features of the model. Following these results, we developed an efficient greedy observation selection algorithm with theoretical bounds on its achievable utility.

Acknowledgements

This work was supported in part by NSF grant 1809327, DARPA grant D19AP00004, and ONR grants N00014-18-1-2829 and N00014-19-1-2054. We thank Alexandros G. Dimakis and Ethan R. Elenberg for their constructive discussion and also thank the reviewers for their valuable feedback.

References

Bellman, R. (1997). *Introduction to matrix analysis*. SIAM.

Bhaskara, A., Rostamizadeh, A., Altschuler, J., Zadimoghaddam, M., Fu, T., and Mirrokni, V. (2016). Greedy column subset selection: New bounds and distributed algorithms. In *International Conference on Machine Learning (ICML)*. Omnipress.

Bian, A. A., Buhmann, J. M., Krause, A., and Tschiatschek, S. (2017). Guarantees for greedy maximization of non-submodular functions with applications. In *International Conference on Machine Learning (ICML)*, pages 498–507. Omnipress.

Candes, E. J., Eldar, Y. C., Strohmer, T., and Voroninski, V. (2015a). Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251.

Candes, E. J., Li, X., and Soltanolkotabi, M. (2015b). Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, pages 273–304.

Chamon, L. and Ribeiro, A. (2017). Approximate supermodularity bounds for experimental design. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5409–5418.

Das, A. and Kempe, D. (2011). Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *International Conference on Machine Learning (ICML)*, pages 1057–1064. Omnipress.

Davidian, M. (2017). *Nonlinear models for repeated measurement data*. Routledge.

Elenberg, E. R., Khanna, R., Dimakis, A. G., Negahban, S., et al. (2018). Restricted strong convexity implies weak submodularity. *The Annals of Statistics*, 46(6B):3539–3568.

Fienup, J. R. (1982). Phase retrieval algorithms: A comparison. *Applied optics*, 21(15):2758–2769.

Flaherty, P., Arkin, A., and Jordan, M. I. (2006). Robust design of biological experiments. In *Advances in Neural Information Processing Systems (NIPS)*, pages 363–370.

Hashemi, A., Ghasemi, M., Vikalo, H., and Topcu, U. (2018). A randomized greedy algorithm for near-optimal sensor scheduling in large-scale sensor networks. In *American Control Conference (ACC)*, pages 1027–1032. IEEE.

Hightower, J. and Borriello, G. (2001). Location systems for ubiquitous computing. *Computer*, 34(8):57–66.

Horel, T. and Singer, Y. (2016). Maximization of approximately submodular functions. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3045–3053.

Joshi, S. and Boyd, S. (2009). Sensor selection via convex optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462.

Kay, S. M. (2013). *Fundamentals of statistical signal processing: Practical algorithm development*, volume 3. Pearson Education.

Krause, A. and Golovin, D. (2014). Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems*, pages 71–104. Cambridge University Press.

Krause, A. and Guestrin, C. (2007). Near-optimal observation selection using submodular functions. In *AAAI Conference on Artificial Intelligence*, volume 7, pages 1650–1654.

Krause, A., McMahan, H. B., Guestrin, C., and Gupta, A. (2008a). Robust submodular observation selection. *Journal of Machine Learning Research*, 9(Dec):2761–2801.

Krause, A., Singh, A., and Guestrin, C. (2008b). Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(Feb):235–284.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (Dec. 1978). An analysis of approximations for maximizing submodular set functions I. *Mathematical Programming*, 14(1):265–294.

Sebastiani, P. and Wynn, H. P. (2000). Maximum entropy sampling and optimal bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):145–157.

Shamaiah, M., Banerjee, S., and Vikalo, H. (2010). Greedy sensor selection: Leveraging submodularity. In *Conference on Decision and Control (CDC)*, pages 2572–2577. IEEE.

Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J., and Segev, M. (2015). Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109.

Skolnik, M. I. (1970). Radar handbook.

Summers, T. H., Cortesi, F. L., and Lygeros, J. (2016). On submodularity and controllability in complex dynamical networks. *IEEE Trans. Control of Network Systems*, 3(1):91–101.

Van Trees, H. L. (2004). *Detection, estimation, and modulation theory, part I: Detection, estimation, and linear modulation theory*. John Wiley & Sons.

Vincent, P. and Rubin, I. (2004). A framework and analysis for cooperative search using UAV swarms. In *Symposium on Applied Computing*, pages 79–86. ACM.

Wang, D., Fisher III, J., and Liu, Q. (2016). Efficient observation selection in probabilistic graphical models using Bayesian lower bounds. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 755–764. AUAI Press.

Wang, Y., Yu, A. W., and Singh, A. (2017). On computationally tractable selection of experiments in measurement-constrained regression models. *The Journal of Machine Learning Research*, 18(1):5238–5278.

Williamson, D. P. and Shmoys, D. B. (2011). *The design of approximation algorithms*. Cambridge university press.

Zhang, H. and Vorobeychik, Y. (2016). Submodular optimization with routing constraints. In *AAAI Conference on Artificial Intelligence*.