Batch greedy maximization of non-submodular functions: Guarantees and applications to experimental design

Jayanth Jagalur-Mohan   JAGALUR@MIT.EDU

Youssef Marzouk   YMARZ@MIT.EDU

Massachusetts Institute of Technology
Cambridge, MA 02139 USA

Abstract

We propose and analyze batch greedy heuristics for cardinality constrained maximization of non-submodular non-decreasing set functions. We consider the standard greedy paradigm, along with its distributed greedy and stochastic greedy variants. Our theoretical guarantees are characterized by the combination of submodularity and supermodularity ratios. We argue how these parameters define tight modular bounds based on incremental gains, and provide a novel reinterpretation of the classical greedy algorithm using the minorize–maximize (MM) principle. Based on that analogy, we propose a new class of methods exploiting any plausible modular bound. In the context of optimal experimental design for linear Bayesian inverse problems, we bound the submodularity and supermodularity ratios when the underlying objective is based on mutual information. We also develop novel modular bounds for the mutual information in this setting, and describe certain connections to polyhedral combinatorics. We discuss how algorithms using these modular bounds relate to established statistical notions such as leverage scores and to more recent efforts such as volume sampling. We demonstrate our theoretical findings on synthetic problems and on a real-world climate monitoring example.

Keywords: Greedy methods, submodularity, non-submodular functions, optimal experimental design, inverse problems, mutual information, uncertainty quantification, Bayesian statistics

1. Introduction

Many design problems in engineering and science are inherently combinatorial. A typical goal in these problems is to select a subset of indices from a larger candidate set, by maximizing or minimizing an objective subject to certain constraints. In resource allocation, for instance, a goal may be to assign tasks to different actors to ensure their expedited and efficient completion. In the climate sciences, a goal might be to create the most informative network of monitoring stations by identifying a set of locations from a larger list of plausible candidate sites.

Problems such as these can be formulated as the optimization of set functions. They have a rich mathematical structure and have been studied extensively (Wolsey and Nemhauser, 1999; Lovász, 2007; Schrijver, 2003; Papadimitriou and Steiglitz, 1998). Often the set functions exhibit properties such as *submodularity* (Definition 1), which allow for efficient

optimization solutions (Fujishige, 2005). The optimization of submodular functions has been a topic of intense research in the past decade owing to these functions' significance for many problems in machine learning. The maximization and minimization of such functions under very general matroid constraints has been a focus of several studies (Vondrak, 2008; Vondrák et al., 2011; Jegelka and Bilmes, 2011; Horel and Singer, 2016; Sviridenko et al., 2017; Khanna et al., 2017b). For a catalog of historical efforts, as well as more recent investigations on these topics, we refer to the monographs by Bach (2013); Krause and Golovin (2014).

While submodularity naturally arises in many problems, there are important problems where the objective lacks this property. Our interest is one such case, where the core task is to maximize a non-submodular but non-decreasing objective. Such objectives are ubiquitous in *optimal experimental design* (OED), a fundamental problem in statistics which involves the specification of all aspects of an experiment. The mathematical foundations of experimental design have a history spanning almost a century (Fisher, 1936; Wald, 1943). Methods developed in this field are relevant to engineering, the social and medical sciences, and econometrics. In broad terms, a typical goal is to seek a design that provides the best "return" (suitably defined) for the least amount of experimental effort. More specifically, we will consider optimal experimental design under cardinality constraints in the Bayesian setting, with an objective that reflects an end goal of parameter inference. For a classical overview of experimental design, we refer the reader to Pukelsheim (2006); Fedorov (2013), and for a perspective on Bayesian optimal experimental design, we highlight the review by Chaloner and Verdinelli (1995).

When maximizing a submodular function under cardinality constraints, the greedy heuristic of successively picking the best candidate performs remarkably well despite its simplicity (Nemhauser et al., 1978; Nemhauser and Wolsey, 1978). The same technique works well in many cases with a non-submodular objective (Das and Kempe, 2011; Bian et al., 2017; Khanna et al., 2017a; Elenberg et al., 2018; Qian and Singer, 2019). Yet the repeated function evaluations necessary to update the incremental gains (Definition 5) at every step may comprise a staggering computational cost if each function evaluation is computationally intensive. It is natural, then, to ask if there are favorable properties of the function that circumvent the need to repeatedly update the incremental gains. To what extent do such properties hold for a given function, and how can we exploit them? And how does the worst case bound suffer if the incremental gains are only periodically updated?

Motivated by these questions, we propose and analyze a variant of the greedy heuristic for non-submodular functions which we refer to as *batch greedy*. We also analyze distributed and stochastic versions of this batch greedy approach. Every step in these algorithms selects a batch of indices, thus reducing the overall number of function evaluations. In many practical problems—for instance Bayesian optimal experimental design—this heuristic still yields excellent solutions, as we shall demonstrate.

The solution index set of the greedy algorithm results from a sequence of locally optimal choices. The choice in every step is determined by the incremental gains associated with each index, or—in the batch greedy setting—by the *sum* of incremental gains corresponding to a fixed-size subset of indices. The sum of incremental gains is simply a modular function (Definition 23) and the individual incremental gains define a subgradient (Definition 27). Analogous to results in the convex analysis for the optimization of continuous functions, the

performance of the greedy algorithm can be tied to properties of the subgradient. The choice of subgradient is not unique for any real valued continuous function, and the same is true for set functions. This observation motivates us to recast sequential greedy algorithms in a more abstract framework using minorize–maximize (MM) optimization principles. Using this framework, we propose several alternative modular bounds and use them to develop novel techniques for linear Bayesian optimal experimental design.

1.1 Outline and summary of key contributions

Much of the manuscript is written with Bayesian optimal experimental design as a motivation. Yet our theoretical results are rather general: applicable to any monotone set function. In describing these results (Section 3 and Section 4.1), we have favored a more abstract notation to help reach a broader readership. These sections can be read more or less independently of the rest of the paper. With the same intention, we disperse much of the relevant literature survey into the sections that follow.

- In Section 2 we introduce some notation and background on set functions, and give an overview of the Bayesian experimental design problem. We formulate the design objectives we study, and comment on some of their set theoretic properties.

- In Section 3 we discuss our primary results. We provide approximation guarantees for the batch greedy heuristic in the context of maximizing monotone but non-submodular objectives. We analyze the standard batch greedy paradigm (Theorem 7), along with distributed batch greedy (Theorem 8) and stochastic batch greedy (Theorem 9) variants. Our results portray the joint expressive power of the submodularity and supermodularity ratios. These two parameters quantify how much a function deviates from submodularity and supermodularity, respectively. We then bound these parameters for the information theoretic objective in linear Bayesian experimental design.

- Viewing the sub/super-modularity ratios as parameters that define tight modular bounds, we argue in Section 4 that the classical greedy heuristic can be viewed as one instance of optimization based on the MM principle. Building on that insight, we put forth a general framework for optimizing monotone set functions using any modular bound (Theorem 12).

- In Section 4.2, for the problem of linear Bayesian experimental design, we explicitly discuss one such MM-based optimization approach exploiting modular bounds on the mutual information objectives. These modular bounds are constructed using standard inequalities but, as we show, they have surprising connections to notions in polyhedral combinatorics.

- In Section 5 we investigate the performance of our algorithms on random instances of structured inverse problems, and on the problem of designing sensor networks for improving climate models. We end with a broader discussion of future directions in Section 6.

- Appendix A contains some background to help with the reading, and all the technical results are collected in Appendix B.

2. Preliminaries and notation

2.1 Set function properties and subset selection

Consider an index set $\mathscr{V} := \{1, \ldots, m\}$, $m \in \mathbb{Z}_{>0}$, and let its power set (i.e., set of all subsets) be denoted by $2^{\mathscr{V}}$. We will refer to $\mathscr{V}$ as the candidate set. Any real-valued set-function $F : 2^{\mathscr{V}} \to \mathbb{R}$ such that $F(\emptyset) = 0$ is submodular (Fujishige, 2005; Bach, 2013) if and only if, for all subsets $\mathscr{A}, \mathscr{B} \subseteq \mathscr{V}$, we have:

$$F(\mathscr{A}) + F(\mathscr{B}) \geq F(\mathscr{A} \cup \mathscr{B}) + F(\mathscr{A} \cap \mathscr{B}). \tag{1}$$

A function is *supermodular* if its negation is submodular, and it is *modular* (see also Definition 23) if it is both supermodular and submodular.

An alternative but equivalent (see Bach, 2013, Proposition 2.2) definition of submodularity highlights the *diminishing returns* property and is often easier to demonstrate in practice:

Definition 1 (Submodular set function defined using first-order differences) *(Fujishige, 2005; Bach, 2013) The set function F is submodular if and only if, for all $\mathscr{A}, \mathscr{B} \subseteq \mathscr{V}$ and $\nu \in \mathscr{V}$ such that $\mathscr{A} \subseteq \mathscr{B}$ and $\nu \notin \mathscr{B}$, we have*

$$F(\mathscr{A} \cup \{\nu\}) - F(\mathscr{A}) \geq F(\mathscr{B} \cup \{\nu\}) - F(\mathscr{B}).$$

In the experimental design problems we investigate, the indices of the candidate set will correspond to individual components of a multivariate random variable $Y \in \mathbb{R}^m$. For the purpose of selecting $k < m$ components we introduce the notion of a selection operator.

Definition 2 (Selection operator) *We refer to $\mathcal{P} \in \mathbb{R}^{m \times k}$, $k < m$, as a selection operator with index set $\mathscr{I}(\mathcal{P}) \equiv \mathscr{I}_{\mathcal{P}} = \{i_1, \ldots, i_k\} \subset \mathscr{V}$, when $\mathcal{P} = [e_{i_1}, \ldots, e_{i_k}]$. The e_{i_j} are distinct canonical unit vectors from the m-dimensional identity matrix I_m.*

Two selection operators that are unique up to permutation of their columns will have identical index sets, and are equivalent for our purposes. We will denote by $\mathcal{S}(k)$ the set of all permutation invariant selection operators $\mathcal{P}$ with $|\mathscr{I}(\mathcal{P})| = k$. It is clear that $|\mathcal{S}(k)| = \binom{m}{k}$. As an abstract operator, $\mathcal{P}$ can be understood as a full-column rank matrix, and an isometry on a k-dimensional subspace of $\mathbb{R}^m$. Applying $\mathcal{P}$ on Y allows us to select k components out of m,

$$Y_{\mathcal{P}} := \mathcal{P}^{\top} Y = [Y_{i_1}, \ldots, Y_{i_k}]^{\top} \in \mathbb{R}^k. \tag{2}$$

2.2 Bayesian inference and optimal experimental design

In Bayesian parameter inference, we seek to characterize the distribution of some parameters of interest, $X \in \mathbb{R}^n$, given a realization of some (related) observations $Y \in \mathbb{R}^m$. More specifically, having endowed X with a prior distribution (whose density we denote by π_X) and knowing the conditional density of the data $\pi_{Y|X}$, we wish to characterize the posterior distribution $\pi_{X|Y=y^*}$ for some realization y^* of Y.[1] The associated Bayesian optimal experimental design problem is to find the subset of observations $Y_{\mathcal{P}}$ that is most informative

---

1. To simplify this exposition, we assume all random variables to have densities with respect to a suitable base measure.

about the parameters of interest X. Solving the inference problem with $Y_{\mathcal{P}}$, or equivalently those observations specified by the index set $\mathscr{I}_{\mathcal{P}}$, will result in a posterior $\pi_{X|Y_{\mathcal{P}}}$ that in general differs from $\pi_{X|Y}$. We refer to the latter as the "full posterior" to indicate it is obtained by conditioning on all candidate observations.

To assess the quality of selected observation subset, we will evaluate the *mutual information* between the inference parameter X and the selected observations $Y_{\mathcal{P}}$. Of course, many other experimental design criteria could be considered and have been employed in the literature. Here we focus on mutual information as it is broadly applicable, and well-founded in Bayesian decision theory as measure of the utility of an experiment (Lindley, 1956) in fully non-Gaussian/nonlinear settings (Chaloner and Verdinelli, 1995).

Definition 3 (Mutual information) *(Cover and Thomas, 2006) Let two random variables X and Y have joint density $\pi_{X,Y}$ and let π_X and π_Y denote the densities of their respective marginals. The mutual information $\mathcal{I}(X;Y)$ is the relative entropy or Kullback–Leibler divergence (Definition 22) between the joint density and the product density $\pi_X \pi_Y$.*

$$\mathcal{I}(X;Y) := D_{KL}(\pi_{X,Y} \| \pi_X \pi_Y) = \mathbb{E}_{\pi_{X,Y}} \log\left(\frac{\pi_{X,Y}}{\pi_X \pi_Y}\right).$$

Bayes' rule allows us to rewrite mutual information as an *expected information gain* from prior to posterior, i.e., $\mathcal{I}(X;Y) = \mathbb{E}_{\pi_Y} D_{KL}(\pi_{X|Y} \| \pi_X)$. This interpretation is quite intuitive: a larger mutual information or expected information gain means that the posterior differs more strongly from the prior, on average.

Formally, we can now state the problem of interest as follows: Given a desired number of observations $k < m$, we seek a selection operator $\mathcal{P}_{\mathrm{opt}} \in \mathcal{S}(k)$ (Definition 2) such that the mutual information between the inference parameter X and the selected observations $Y_{\mathcal{P}_{\mathrm{opt}}} := \mathcal{P}_{\mathrm{opt}}^\top Y$ is maximized:

$$\mathcal{P}_{\mathrm{opt}} = \underset{\mathcal{P} \in \mathcal{S}(k) \subset \mathbb{R}^{m \times k}}{\arg\max} \mathcal{I}\left(X; \mathcal{P}^\top Y\right). \tag{Prob-Max}$$

Alternatively, we could indirectly determine the selection operator by first finding a complementary set of observations that do not significantly inform the inference parameters. While this approach may seem convoluted, its value is easier to appreciate when the goal is to remove a small fraction of observations from the parent set while retaining the bulk. To the best of our knowledge this has not been studied in the context of Bayesian inference, but the underlying *reverse/backward* principle can be found in earlier investigations concerning subset and feature selection (Couvreur and Bresler, 2000; Zhang, 2011), graph cut approximation (Bian et al., 2015), optimization of a certain class of set functions (Il'ev, 2001), and classical experimental design using elementary symmetric polynomials (Mariet and Sra, 2017).

For any selection operator $\mathcal{P} \in \mathcal{S}(k)$, let $\mathcal{P}^c \in \mathcal{S}(m-k)$ denote its complement, meaning their corresponding index sets are such that $\mathscr{I}_{\mathcal{P}} \cup \mathscr{I}_{\mathcal{P}^c} = \mathscr{V}$ and $\mathscr{I}_{\mathcal{P}} \cap \mathscr{I}_{\mathcal{P}^c} = \emptyset$. Our objective now is to minimize the loss of mutual information by discarding an optimal subset of observations $Y_{\mathcal{P}_{\mathrm{opt}}^c} := \mathcal{P}_{\mathrm{opt}}^{c\,\top} Y$. The loss is measured with respect to mutual information as determined by the set of all candidate observations.

$$\mathcal{P}_{\mathrm{opt}}^c = \underset{\mathcal{P}^c \in \mathcal{S}(m-k) \subset \mathbb{R}^{m \times m-k}}{\arg\min} \mathcal{I}(X;Y) - \mathcal{I}\left(X; Y \setminus \mathcal{P}^{c\top} Y\right). \tag{Prob-Min}$$

The optimal index sets pertinent to (`Prob-Max`) and (`Prob-Min`) are identical; hence we do not distinguish between them notationally. Note, however, that the objectives differ in how they vary as functions of $\mathscr{I}_{\mathcal{P}}$. The objective in (`Prob-Max`) is a non-decreasing function with respect to cardinality of $\mathscr{I}_{\mathcal{P}}$, and consequently non-increasing with respect to $\mathscr{I}_{\mathcal{P}^c}$. In contrast, the objective in (`Prob-Min`) has opposite relationships with respect to the same cardinalities. More importantly, the two approaches have philosophical and practical differences.

We show in Proposition 4 that the objective in (`Prob-Max`), is a submodular function when the observations are conditionally independent.

Proposition 4 *Given random variables $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$, let $\mathcal{P} \in \mathbb{R}^{m \times k}$ be a selection operator such that $\mathcal{P}^{\top} Y = [Y_{i_1}, \ldots, Y_{i_k}]^{\top}$, with $\mathscr{I}_{\mathcal{P}} \subset \mathscr{V}$. The mutual information $\mathcal{I}(X; \mathcal{P}^{\top} Y)$, between X and $\mathcal{P}^{\top} Y$ is submodular if $Y_{i_j} | X$ are independent.*

The proof is given in Appendix B.

It is worth emphasizing that the submodularity of the mutual information $\mathcal{I}(X; \mathcal{P}^{\top} Y)$ for conditionally independent observations holds even when the underlying joint distribution is non-Gaussian. As a simple corollary of Proposition 4, the objective in (`Prob-Min`) can be shown to be supermodular under the same assumptions. More precisely, it is the *supermodular dual*, but we will defer more discussion on the topic of duals to Section 4.2.

3. Batch greedy algorithms for maximizing monotone set functions

We now focus solely on cardinality-constrained maximization of monotone set functions. We will not restrict these functions to be submodular. We begin with a summary of various existing greedy heuristics, to contextualize the results that follow.

3.1 Greedy algorithms: a brief history

In the case of cardinality-constrained maximization of non-decreasing submodular functions, the greedy heuristic of successively picking the candidate corresponding to the highest incremental gain (Definition 5) performs well despite its simplicity. It has a constant factor $(1 - 1/e)$ approximation guarantee (Nemhauser et al., 1978), which cannot be improved in general by any other polynomial time algorithm (Nemhauser and Wolsey, 1978). If the function can be shown to have small curvature $c \in [0, 1]$ (Definition 30), then the greedy algorithm possesses a more refined guarantee, $\frac{1}{c}(1 - e^{-c})$, (Conforti and Cornuéjols, 1984). If the function is not submodular, then by incorporating a submodularity ratio $\gamma \in [0, 1]$ (Definition 32), Das and Kempe (2011) showed how the greedy algorithm achieves a $(1 - e^{-\gamma})$ approximation guarantee. Additionally, using the notion of generalized curvature $\alpha \in [0, 1]$ (Definition 31) for non-submodular functions, Bian et al. (2017) proved a more expressive approximation guarantee, $\frac{1}{\alpha}(1 - e^{-\alpha \gamma})$.

Definition 5 (Incremental gain) *We denote the incremental or marginal gain of a set $\mathscr{A} \subset \mathscr{V}$ given a set $\mathscr{B} \subset \mathscr{V}$ as $\rho_{\mathscr{A}}(\mathscr{B}) := F(\mathscr{A} \cup \mathscr{B}) - F(\mathscr{B})$. For $\nu \in \mathscr{V}$, we use the shorthand $\rho_{\nu}(\mathscr{B})$ for $\rho_{\{\nu\}}(\mathscr{B})$.*

Closely related variants of the greedy heuristic can be better choices depending on the context and needs.

- Robertazzi and Schwartz (1989) explored an accelerated version wherein the computed incremental gains are stored and exploited in the successive step, possibly reducing the overall number of function evaluations.

- Mirzasoleiman et al. (2013) proposed a two-stage parallelized version which reduces the number of function evaluations per parallel process. The approximation guarantee for the algorithm, however, in general depends on the size of the candidate set and cardinality constraint. That dependence can only be overcome in special cases.

- Mirzasoleiman et al. (2015) analyzed a randomized version of the greedy heuristic, termed stochastic greedy. This algorithm achieves, in expectation, a $(1 - 1/e - \epsilon)$ approximation guarantee relative to the optimum solution. The number of function evaluations does not depend on the cardinality constraint, but linearly on the size of the candidate set, thus reducing the complexity substantially.

- Liu et al. (2016) analyzed the greedy heuristic wherein the locally optimal decision involved selecting the best possible set of $q > 1$ indices. This necessarily requires evaluating incremental gains associated with all combinatorial possibilities, a potentially severe overhead but one which offers better guarantees. The algorithm was referred to as batch greedy by Liu et al. (2016), where the batch size is the cardinality of the *locally combinatorially optimal* set chosen in each step.

3.2 Batch greedy algorithm and its analysis

The variants of greedy discussed in Section 3.1 were mostly analyzed only in the context of non-decreasing submodular functions. Our approach (Algorithms 1 to 3) can be understood to be yet another distinct variant of the greedy heuristic, but one which we analyze for the more general case of monotone non-submodular objectives. While we label our approach as batch greedy, it is unlike the algorithm of Liu et al. (2016) and in some sense its polar opposite.

In particular, we investigate the greedy strategy of picking multiple candidates in each step but relying solely on the incremental gains associated with individual candidates. This naturally reduces the computational overhead by avoiding combinatorial combinations, but at the expense of inferior approximation guarantees. It is in this sense the exact opposite of the algorithm proposed by Liu et al. (2016), since we are at the other end of the trade-off spectrum.

Algorithm 1 describes the standard batch greedy algorithm. The total cardinality constraint k is necessarily the sum of batch sizes across all steps, $k = \sum_i q_i$. We seek approximation guarantees for Algorithm 1 pertinent to the maximization of any non-decreasing set function. To aid our arguments we introduce the *supermodularity ratio*, which has very recently been used in other contexts too (Tzoumas et al., 2017; Bogunovic et al., 2018; Karaca and Kamgarpour, 2018).

Algorithm 1 Standard batch greedy algorithm

1: Input $F, \mathscr{V}, l, \{q_1, \ldots, q_l\}$
2: Initialize $\mathscr{A} = \emptyset$
3: **for** $i = 1$ to l **do**
4:     Determine $\rho_a(\mathscr{A}) \quad \forall a \in \mathscr{V} \setminus \mathscr{A}$.
5:     Find $\mathscr{Q} \subseteq \mathscr{V} \setminus \mathscr{A}$, $|\mathscr{Q}| = q_i$, comprising the indices with the highest incremental gains.
6:     $\mathscr{A} \leftarrow \mathscr{A} \cup \mathscr{Q}$
7: **end for**
        return Index set $\mathscr{A}$

Definition 6 (Supermodularity ratio) *The supermodularity ratio of a non-negative set function F with respect to a set $\mathscr{V}$ and a parameter $k \geq 1$ is*

$$\eta_{\mathscr{V},k}(F) = \min_{\mathscr{B} \subseteq \mathscr{V}, \mathscr{A}:|\mathscr{A}| \leq k, \mathscr{A} \cap \mathscr{B} = \emptyset} \frac{\rho_{\mathscr{A}}(\mathscr{B})}{\sum_{\nu \in \mathscr{A}} \rho_\nu(\mathscr{B})}.$$

The supermodularity ratio is inspired by and related to the submodularity ratio (Definition 32), originally introduced by Das and Kempe (2011). It is a lower bound on the ratio of the incremental gain associated with any set compared against the sum of incremental gains associated with its elements. In contrast, the reciprocal of submodularity ratio is an upper bound on the same quantity. Bian et al. (2017) defined the submodularity ratio without the cardinality parameter k (Definition 33) by taking a minimum across all possibilities. In the same way, we can define the supermodularity ratio without the cardinality parameter as the largest scalar η such that

$$\frac{\rho_{\mathscr{A}}(\mathscr{B})}{\sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} \rho_\nu(\mathscr{B})} \geq \eta, \qquad \forall \mathscr{A}, \mathscr{B} \subseteq \mathscr{V}.$$

We will refer to both $\eta_{\mathscr{V},k}$ and η as supermodularity ratio, preferring one over the other depending on the context. Informally, the supermodularity ratio quantifies how close a set function is to being supermodular, while the submodularity ratio performs the same task for submodularity. More formally, we can prove that a function F is supermodular *iff* the supermodularity ratio $\eta(F) = 1$ (see proof in Appendix B.1). For all set functions which are not supermodular (but which may or may not be submodular) we have the condition $\eta < 1$ as a direct corollary.

Now we state the result corresponding to Algorithm 1.

Theorem 7 *Let F be a non-decreasing function with $F(\emptyset) = 0$. The batch greedy algorithm for maximizing $F(\mathscr{A})$ subject to $|\mathscr{A}| \leq k$ outputs a set $\mathscr{A}$ such that*

$$F(\mathscr{A}) \geq \left(1 - \prod_{i=1}^{l} \left(1 - \frac{q_i \eta_{\mathscr{V},q_i} \gamma_{\mathscr{V},k}}{k}\right)\right) \max_{\mathscr{B} \subset \mathscr{V}, |\mathscr{B}| \leq k} F(\mathscr{B})$$

where $\gamma_{\mathscr{V},k}$ is the submodularity ratio and $\eta_{\mathscr{V},q_i}$ is the supermodularity ratio.

The proof is given in Appendix B.1.

If we chose q indices during each step with the total number of indices $k = ql$, the approximation guarantee in Theorem 7 can be simplified using a standard logarithmic inequality:

$$F(\mathscr{A}) \geq \left(1 - e^{-\eta_{\mathscr{V},q}\gamma_{\mathscr{V},k}}\right) \max_{\mathscr{B} \subset \mathscr{V}, |\mathscr{B}| \leq k} F(\mathscr{B}). \tag{3}$$

Alternatively q can also be the maximum batch size across all steps, $q = \max q_i$, $i = 1, \ldots, l$. The approximation guarantee as given in (3) can be viewed as a straightforward generalization of that by Das and Kempe (2011) pertinent to the variant of greedy with batch selection. Letting the function be submodular, $\gamma_{\mathscr{V},k} \geq 1$, and choosing one index in every step, $q = 1$, reduces the guarantee to the classical result by Nemhauser et al. (1978) since $\eta_{\mathscr{V},1} = 1$ for any set function. Selecting more than one index during each step worsens the guarantee since $\forall k_1, k_2$ with $k_1 \geq k_2 \geq 1$ we have $\eta_{\mathscr{V},k_1} \leq \eta_{\mathscr{V},k_2}$; this claim holds by definition because the constraint set of the latter is contained in the former.

In Figure 1 we visualize the approximation factor $1 - e^{-\eta_{\mathscr{V},q}\gamma_{\mathscr{V},k}}$ in (3) for the range of submodularity and supermodularity ratios. If the function is modular, the approxima-
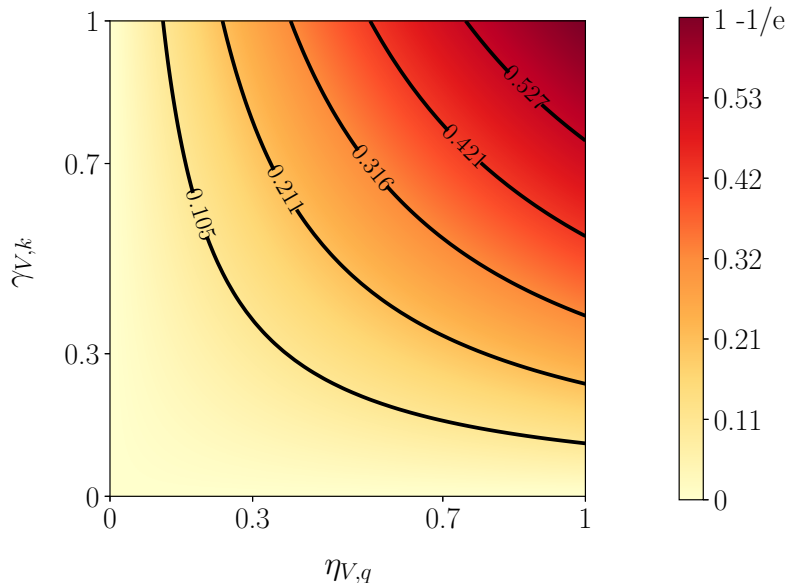

Figure 1: Visualization of the approximation factor $1 - e^{-\eta_{\mathscr{V},q}\gamma_{\mathscr{V},k}}$ in (3).

tion factor $1 - e^{-\eta_{\mathscr{V},q}\gamma_{\mathscr{V},k}}$ reduces to $1 - 1/e$, highlighting a gap in our analysis similar

to that of Das and Kempe (2011). The supermodularity ratio $\eta_{\mathscr{V},q}$ is necessary to characterize the batch greedy heuristic, while the submodularity ratio $\gamma_{\mathscr{V},k}$ characterizes the non-submodularity of the function. Together, as they appear in (3) they are not sufficient to refine the worst case bound accounting for modularity. The notions of curvature (Definition 30) for submodular functions and generalized curvature (Definition 31) for non-submodular functions have been used in non-batch settings to give refined approximation guarantees avoiding such gaps (Conforti and Cornuéjols, 1984; Bian et al., 2017). It should be possible to incorporate curvature in an analysis of the batch greedy heuristic to generalize the result in Theorem 7, but we defer this investigation to the future.

If we consider the result in Theorem 7 directly and simplify it for the case when we choose all the indices in one step, we get a tight bound:

$$F(\mathscr{A}) \geq \eta_{\mathscr{V},k}\gamma_{\mathscr{V},k} \max_{\mathscr{B}\subset\mathscr{V},|\mathscr{B}|\leq k} F(\mathscr{B}). \tag{4}$$

The product $\eta_{\mathscr{V},k}\gamma_{\mathscr{V},k} \leq 1$ by definition. The algorithm returns an optimal index set if $\eta_{\mathscr{V},k}\gamma_{\mathscr{V},k} = 1$. One instance of this scenario is when the function is modular, in which case it is both supermodular and submodular, meaning $\gamma_{\mathscr{V},k} = \eta_{\mathscr{V},k} = 1, \forall k$ and hence $\eta_{\mathscr{V},k}\gamma_{\mathscr{V},k} = 1, \forall k$. Of course, we can always optimize a modular function exactly, and hence this observation is not entirely useful on its own. If we view the product of supermodularity and submodularity ratios as a measure of deviation from modularity, however, then they together prescribe favorable circumstances for a function that is not necessarily modular to be *almost* maximized exactly.

3.3 Distributed batch greedy algorithm and its analysis

The distributed greedy algorithm for submodular maximization was first proposed by Mirzasoleiman et al. (2013). Our distributed *batch* greedy approach is described in Algorithm 2. The parameters $\widehat{l}$ and $\widetilde{l}$ are the number of iterations for which the batch greedy algorithm is run in each round. We define $\widehat{k} := q\widehat{l}$ and $\widetilde{k} := q\widetilde{l}$ as the cumulative number of indices returned at the end round. Similar to the setup in Mirzasoleiman et al. (2013), we allow for index sets larger than the cardinality constraint to be returned, meaning $k \leq \min(\widehat{k},\widetilde{k})$.

Algorithm 2 Distributed batch greedy algorithm

1: Input $F, \mathscr{V}, n_p, \widehat{l}, \widetilde{l}, q$
2: Partition the set $\mathscr{V}$ into n_p sets $\mathscr{V}_1, \ldots, \mathscr{V}_{n_p}$
3: Run the standard batch greedy algorithm, with batch size q, for $\widehat{l}$ iterations within each set $\mathscr{V}_i$ to yield the corresponding solution set $\mathscr{A}^{\text{bg}}_{i,[n_p,\widehat{k}]}$
4: Merge the result sets: $\bigcup_i \mathscr{A}^{\text{bg}}_{i,[n_p,\widehat{k}]} =: \mathscr{M}$
5: Run the standard batch greedy algorithm, with batch size q, for $\widetilde{l}$ iterations on $\mathscr{M}$ to yield the solution set $\mathscr{A}^{\text{d-bg}}_{[n_p,\widetilde{k}]}$
     return Index set $\mathscr{A}^{\text{d-bg}}_{[n_p,\widetilde{k}]}$

In the most general case, the input to Algorithm 2 could include different batch sizes in each round, across parallel processes, and across different iterations. For the purpose

of exposition and analysis, we have favored a simplified version of the algorithm with a fixed/uniform batch size q. Our theoretical analysis, summarized as Theorem 8 below, extends the work of Mirzasoleiman et al. (2013) not only to batch settings but also to non-submodular functions.

Theorem 8 *Let F be a non-decreasing function with $F(\emptyset) = 0$. The distributed batch greedy algorithm for maximizing $F(\mathscr{A})$ subject to $|\mathscr{A}| \leq k = ql$ outputs a set $\mathscr{A}^{d\text{-}bg}_{[n_p, \widehat{k}]}$ such that*

$$F\left(\mathscr{A}^{d\text{-}bg}_{[n_p, \widehat{k}]}\right) \geq \left(1 - e^{-\eta_{\mathscr{V},q}\,\gamma_{\mathscr{V},\widehat{k}}\,(\widetilde{l}/l)}\right)\left(1 - e^{-\eta_{\mathscr{V},q}\,\gamma_{\mathscr{V},k}\,(\widehat{l}/l)}\right)\frac{\gamma_{\mathscr{V},k}}{k}\max_{\mathscr{B} \subset \mathscr{V}, |\mathscr{B}| \leq k} F(\mathscr{B}),$$

where $\gamma_{\mathscr{V},k}$, $\gamma_{\mathscr{V},\widehat{k}}$ are submodularity ratios and $\eta_{\mathscr{V},q_i}$ is the supermodularity ratio.

The proof is given in Appendix B.1.

If F is submodular ($\gamma = 1$), and further if $\widehat{l} = \widetilde{l} = l$, then Theorem 8 can be simplified as follows:

$$F(\mathscr{A}^{d\text{-}bg}) \geq \frac{(1 - e^{-\eta_{\mathscr{V},q}})^2}{\min(n_p, k)}\max_{\mathscr{B} \subset \mathscr{V}, |\mathscr{B}| \leq k} F(\mathscr{B}). \tag{5}$$

Note that the term appearing in the denominator of (5) is $\min(n_p, k)$ and not k, the natural expected term as per Theorem 8. In this case of F being submodular, we rely on Mirzasoleiman et al. (2013, Theorem 4.1) as opposed to Lemma 34 to prove the slightly improved result in (5). The dependence of the distributed solution on $\min(n_p, k)$ is in general unavoidable, but as argued by Mirzasoleiman et al. (2013), the ground set $\mathscr{V}$ and function F can exhibit rich geometrical structure that can be used to prove stronger results.

3.4 Stochastic batch greedy algorithm and its analysis

The stochastic greedy algorithm for submodular functions was first proposed by Mirzasoleiman et al. (2015). Here we present a batch variant of the stochastic greedy algorithm, in Algorithm 3 below.

Algorithm 3 Stochastic batch greedy algorithm

1: Input $F, \mathscr{V}, l, \{q_1, \ldots, q_l\}, s$
2: Initialize $\mathscr{A} = \emptyset$
3: **for** $i = 1$ to l **do**
4:     Construct a randomly sampled set $\mathscr{R}$ by sampling s random elements from $\mathscr{V} \setminus \mathscr{A}$.
5:     Determine $\rho_a(\mathscr{A})$, $\forall a \in \mathscr{R}$.
6:     Find $\mathscr{Q} \subseteq \mathscr{R}$, $|\mathscr{Q}| = q_i$, comprising the indices with the highest incremental gains.
7:     $\mathscr{A} \leftarrow \mathscr{A} \cup \mathscr{Q}$
8: **end for**
      return Index set $\mathscr{A}$

In the most general case, the input to Algorithm 3 includes batch sizes q_i, $i = 1, \ldots, l$, indicating the number of indices selected at each step. We will analyze the algorithm (wlog) for the case of fixed batch size q with the cardinality constraint $k = ql$. Furthermore we will

restrict our analysis to the regime when the cardinality constraint is at most $\mathcal{O}(\sqrt{m})$; this is very much reflective of practical scenarios when stochastic greedy algorithms are warranted. Our theoretical analysis extends the work of Mirzasoleiman et al. (2015) not only to batch settings but also to non-submodular functions; the main result is stated in Theorem 9.

Theorem 9 *Let F be a non-decreasing function with $F(\emptyset) = 0$, and $0 < \epsilon < 1$ be a tolerance parameter. The stochastic batch greedy algorithm for maximizing $F(\mathscr{A})$ with $s = \frac{m}{k} \log \frac{q}{\epsilon}$, $|\mathscr{A}| \leq k \leq \lfloor \sqrt{m/e} - 1/e \rfloor$, and $\frac{m-2k}{2ek^2} \geq \frac{q-1}{\log^2 \frac{q}{\epsilon}}$ outputs a set $\mathscr{A}$ such that*

$$\mathbb{E}[F(\mathscr{A})] \geq \left(1 - e^{-(1-\epsilon)\gamma_{\mathscr{V},k}\eta_{\mathscr{V},q}}\right) \max_{\mathscr{B} \subset \mathscr{V}, |\mathscr{B}| \leq k} F(\mathscr{B}),$$

where $\gamma_{\mathscr{V},k}$ is the submodularity ratio and $\eta_{\mathscr{V},q}$ is the supermodularity ratio.

The proof is given in Appendix B.1.

If the function F is submodular ($\gamma = 1$) and if the batch size is equal to one, then our result reduces to that shown by Mirzasoleiman et al. (2015). It is worth highlighting that the assumptions of Theorem 9—specifically the last inequality relating m, k, q, and ϵ—stem primarily from the complexity of a required probability bound, detailed as Lemma 36. The complexities arise specifically when the batch size is greater than one. The corresponding constraints, however, should be easily satisfied in practice. For instance, suppose that $k \leq \frac{1}{\kappa}\sqrt{m/e}$ for some $\kappa > 1$. Then it is easy to see that $\frac{m-2k}{2ek^2} \geq \frac{q-1}{\log^2 \frac{q}{\epsilon}}$ holds when $\epsilon \leq qe^{-\frac{\sqrt{2(q-1)}}{\kappa}}$. When $q = 1$, this constraint on the tolerance parameter is trivial since we know $0 < \epsilon < 1$. When $q > 1$, a sufficiently large κ will ensure that the condition can be satisfied.

3.5 Complexity of the batch greedy algorithms

Recall that we denote the size of the candidate set $\mathscr{V}$ as m and the desired cardinality as k. The number of function evaluations needed to find the standard greedy solution is $\mathcal{O}(mk)$.

- In the batch heuristic (Algorithm 1) with a uniform batch size q, the number of function evaluations reduces to $\mathcal{O}(\frac{mk}{q})$. Thus a suitably chosen batch size reduces the computational overhead of a large cardinality constraint, but the linear dependence on the size of the candidate set remains the same.

- For the distributed batch greedy heuristic (Algorithm 2), if $\widehat{k} = \widetilde{k} = k$, then the number of function evaluations per parallel process to construct the set $\mathscr{M}$ is $\mathcal{O}(\frac{mk}{qn_p})$. To determine the distributed batch greedy solution using $\mathscr{M}$ we will further need $\mathcal{O}(\frac{k^2 n_p}{q})$ function evaluations.

- For the stochastic batch greedy heuristic (Algorithm 3), the number of function evaluations needed to find the solution is $\widetilde{\mathcal{O}}(\frac{m}{q})$; here $\widetilde{\mathcal{O}}$ absorbs the $\log q$ dependency that arises since the random set $\mathscr{R}$ drawn at each step scales logarithmically with the batch size. The complexity does not depend on the cardinality constraint due to the probabilistic nature of the algorithm.

In all three instances, the batch variant of the greedy heuristic reduces the complexity by a factor $1/q$.

3.6 Theoretical guarantees for linear Bayesian optimal experimental design

The Bayesian linear–Gaussian model arises in numerous practical applications, and is a building block for countless others. It is particularly important in *inverse problems* (Stuart, 2010; Kaipio and Somersalo, 2006), where the observations may depend indirectly on the parameters through the action of a smoothing forward operator. Examples of such problems include computerized tomography and electromagnetic source inversion. Here we discuss theoretical guarantees for the performance of Algorithms 1 to 3 when used in optimal experimental design for these problems. First, we set up the relevant notation and discuss some important features of linear Bayesian experimental design.

3.6.1 Linear Bayesian inverse problem

Without loss of generality, we model the parameters X and observation noise ϵ as zero-mean, normally distributed random variables with covariance matrices Γ_X and $\Gamma_{Y|X}$, respectively. Here Y denotes the observed data. We assume that X and ϵ are independent of each other. The linear forward model that maps parameters to data is represented by $G \in \mathbb{R}^{m \times n}$. Hence

$$Y = GX + \epsilon \tag{6}$$

serves as our statistical model for the data and specifies the likelihood, i.e., $Y|x \sim \mathcal{N}(Gx, \Gamma_{Y|X})$. The data Y thus have a marginal distribution $\mathcal{N}(0, \Gamma_Y)$, where the covariance Γ_Y is

$$\Gamma_Y := G\Gamma_X G^\top + \Gamma_{Y|X}. \tag{7}$$

The linearity of the forward model, along with Gaussianity of the prior and observation noise, allows us to characterize the posterior in closed form: $X|Y \sim \mathcal{N}(\mu_{X|Y}, \Gamma_{X|Y})$. Here $\Gamma_{X|Y}$ is the posterior covariance matrix and $\mu_{X|Y}$ is the posterior mean, which is a function of the actual realization of the data y:

$$\Gamma_{X|Y} := \left(\Gamma_X^{-1} + G^\top \Gamma_{Y|X}^{-1} G \right)^{-1}, \tag{8a}$$

$$\mu_{X|Y}(y) := \Gamma_{X|Y} G^\top \Gamma_{Y|X}^{-1} y. \tag{8b}$$

In (8a), the term $G^\top \Gamma_{Y|X}^{-1} G$ is the Hessian of the negative log-likelihood (i.e., the Fisher information matrix).

The task of optimal experimental design involves selecting a *subset* of observations $Y_\mathcal{P} = \mathcal{P}^\top Y \in \mathbb{R}^k$, corresponding to some selection operator $\mathcal{P} \in \mathbb{R}^{m \times k}$, such that $X|Y_\mathcal{P} \sim \mathcal{N}(\mu_{X|Y_\mathcal{P}}, \Gamma_{X|Y_\mathcal{P}})$ has minimal uncertainty (given constraints on k). We make this goal precise by using the mutual information between X and $Y_\mathcal{P}$, $\mathcal{I}(X; Y_\mathcal{P})$, as our design objective. Analogous to (8) we can write the statistical moments of the posterior $X|Y_\mathcal{P}$ as

$$\Gamma_{X|Y_\mathcal{P}} := \left(\Gamma_X^{-1} + G_\mathcal{P}^\top \Gamma_{Y_\mathcal{P}|X}^{-1} G_\mathcal{P} \right)^{-1}, \tag{9a}$$

$$\mu_{X|Y_\mathcal{P}}(y) := \Gamma_{X|Y_\mathcal{P}} G_\mathcal{P}^\top \Gamma_{Y_\mathcal{P}|X}^{-1} \mathcal{P}^\top y. \tag{9b}$$

In (9) $G_\mathcal{P} := \mathcal{P}^\top G \in \mathbb{R}^{k \times n}$ and $\Gamma_{Y_\mathcal{P}|X} := \mathcal{P}^\top \Gamma_{Y|X} \mathcal{P} : \mathbb{R}^{k \times k}$ is the compression of $\Gamma_{Y|X}$ by $\mathcal{P}$ to $\mathbb{R}^k$. In linear algebraic terms, $\Gamma_{Y_\mathcal{P}|X}$ is a principal submatrix of $\Gamma_{Y|X}$. In a similar

manner, the marginal covariance of $Y_{\mathcal{P}}$ is

$$\Gamma_{Y_{\mathcal{P}}} := G_{\mathcal{P}}\Gamma_X G_{\mathcal{P}}^\top + \Gamma_{Y_{\mathcal{P}}|X} = \mathcal{P}^\top \Gamma_Y \mathcal{P} : \mathbb{R}^{k \times k}. \tag{10}$$

From (9a) we can identify $G_{\mathcal{P}}^\top \Gamma_{Y_{\mathcal{P}}|X}^{-1} G_{\mathcal{P}}$ as the relevant Hessian term when the likelihood is specified using any subset of observations $Y_{\mathcal{P}}$. It is clear that this term is *not* a compression of the full Hessian $G^\top \Gamma_{Y|X}^{-1} G$ unless $\Gamma_{Y|X}$ is diagonal, meaning that the observation errors are uncorrelated.

3.6.2 SPECTRAL PROPERTIES OF THE MUTUAL INFORMATION

Since the inference parameters X and the data Y are jointly Gaussian random variables, the mutual information between them, $\mathcal{I}(X;Y)$, can be written as

$$\mathcal{I}(X;Y) = \mathcal{I}(Y;X) = \frac{1}{2}\log\frac{\det(\Gamma_Y)}{\det(\Gamma_{Y|X})} = \frac{1}{2}\log\frac{\det(\Gamma_X)}{\det(\Gamma_{X|Y})}. \tag{11}$$

The above expression can be easily verified from first principles. From (11), it is clear that we could alternatively express mutual information as a function of the generalized eigenvalues of the definite pencils $(\Gamma_Y, \Gamma_{Y|X})$ or $(\Gamma_X, \Gamma_{X|Y})$ (Definition 19). The first can be viewed as a "data space" pencil, while the second is the corresponding "parameter space" pencil. Since $\Gamma_Y \succeq \Gamma_{Y|X} \succ 0$ and $\Gamma_X \succeq \Gamma_{X|Y} \succ 0$, the eigenvalues of both definite pencils are lower bounded by one. Here the symbol '$\succeq$' denotes the Löwner ordering, or the positive semi-definite ordering of Hermitian matrices (Definition 17). Furthermore, the two generalized eigenvalue problems, while differing in dimension, have identical generalized eigenvalues that are strictly greater than one. We make this fact precise through the following proposition.

Proposition 10 *Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ be jointly Gaussian random variables as defined in Section 3.6.1. Then the following conditions are equivalent:*

1. *The definite pairs $(\Gamma_Y - \Gamma_{Y|X}, \Gamma_{Y|X})$ and $(\Gamma_X - \Gamma_{X|Y}, \Gamma_{X|Y})$ have identical non-trivial generalized eigenvalues, $\sigma_j > 0$.*

2. *The definite pairs $(\Gamma_Y, \Gamma_{Y|X})$ and $(\Gamma_X, \Gamma_{X|Y})$ have identical generalized eigenvalues that are strictly greater than 1, $1 + \sigma_j > 1$.*

The proof is given in Appendix B.2.

Using the notation of Theorem 10, we can now write $\mathcal{I}(X;Y)$ as follows:

$$\mathcal{I}(X;Y) = \frac{1}{2}\sum_j \log\left(1 + \sigma_j\right). \tag{12}$$

Such an expression for mutual information in terms of the generalized eigenvalues σ_j has been highlighted in many works (Alexanderian et al., 2016; Giraldi et al., 2017) that adopt a Bayesian formalism. From a more classical statistics perspective, mutual information can be written using the squared canonical correlation scores (Bach and Jordan, 2002) between the concerned random variables. These scores can be computed using generalized eigenvalue problems that are different from those in Theorem 10, but not surprisingly have a similar dual representation.

If we were to determine $\mathcal{I}(X; Y_{\mathcal{P}})$, the mutual information between X and a subset of observations $Y_{\mathcal{P}}$, then the relevant generalized eigenvalues are those of the definite pairs $(\Gamma_{Y_{\mathcal{P}}}, \Gamma_{Y_{\mathcal{P}}|X})$ or equivalently $(\Gamma_X, \Gamma_{X|Y_{\mathcal{P}}})$. The equivalence holds since Theorem 10 applies to these new definite pairs constructed using the compressed covariance operators. If we denote by $\widehat{\sigma}_j$ the eigenvalues for the case when we use a subset of observations $Y_{\mathcal{P}}$, then we can write $\mathcal{I}(X; Y_{\mathcal{P}})$ in a manner analogous to (11) and (12) as

$$\mathcal{I}(X; Y_{\mathcal{P}}) = \frac{1}{2} \log \frac{\det(\Gamma_{Y_{\mathcal{P}}})}{\det(\Gamma_{Y_{\mathcal{P}}|X})} = \frac{1}{2} \log \frac{\det(\Gamma_X)}{\det(\Gamma_{X|Y_{\mathcal{P}}})} = \frac{1}{2} \sum_j \log\left(1 + \widehat{\sigma}_j\right). \tag{13}$$

The expression for mutual information $\mathcal{I}(X; Y_{\mathcal{P}})$ is quite revealing upon closer scrutiny. First, notice that we can rewrite (13) as the difference of log principal determinant of the data marginal and observation error covariance operators,

$$\mathcal{I}(X; Y_{\mathcal{P}}) = \frac{1}{2}\left(\log \det\left(\mathcal{P}^\top \Gamma_Y \mathcal{P}\right) - \log\det\left(\mathcal{P}^\top \Gamma_{Y|X}\mathcal{P}\right)\right). \tag{14}$$

It is thus the difference between two set functions, each one defined as the log determinant of the principal submatrix of a definite matrix. Here the set comprises the indices of rows/columns of the principal submatrix, with each index corresponding to a unique observation. Each of those set functions can be inferred to be submodular (Gantmacher and Kreĭn, 1960; Kotelyanskiĭ, 1950; Fan, 1967, 1968; Kelmans and Kimelfeld, 1983; Johnson and Barrett, 1985). Mutual information can thus be interpreted as the difference between two submodular functions (and hence is a DS function, following the terminology by Iyer and Bilmes (2012)). This is not entirely surprising since it has been shown that *every* set function can be decomposed as the difference between two submodular functions (see Narasimhan and Bilmes, 2005, Lemma 4), (see Iyer and Bilmes, 2012, Lemma 3.1). This fact is akin to how any continuous function, subject to weak conditions, can be expressed as the sum of a convex and concave part (see Yuille and Rangarajan, 2003, Theorem 1). Such a decomposition is not unique and can be exponentially hard to compute (Iyer and Bilmes, 2012), but in our case it is readily apparent.

The sum of the generalized eigenvalues, $\sum \widehat{\sigma}_j$, is the expected *symmetrized* Kullback–Leiber divergence between the prior π_X and posterior $\pi_{X|Y_{\mathcal{P}}}$; this relationship is shown in Proposition 38. Note that the expected symmetrized Kullback-Leiber divergence $\mathbb{E}_{\pi_{Y_{\mathcal{P}}}}\left[D_{\mathrm{KL}}(\pi_{X|Y_{\mathcal{P}}} \| \pi_X) + D_{\mathrm{KL}}(\pi_X \| \pi_{X|Y_{\mathcal{P}}})\right]$ is nothing but the mutual information $\mathcal{I}(X; Y_{\mathcal{P}})$ plus an additional term. It is worth contrasting this design criterion with that used in traditional Bayesian A-optimal design, where the trace of the posterior covariance operator (i.e., sum of simple eigenvalues) is minimized. In the former case the generalized eigenvalues correspond to the largest reductions in posterior variance *relative* to the prior (see Spantini et al., 2015, Corollary 3.1), while in the latter case the simple eigenvalues represent the largest absolute contributions to the posterior variance, without regard to the prior. On the other hand, Bayesian D-optimal design in the linear–Gaussian case, wherein we minimize the determinant of the posterior covariance operator, is identical to maximizing mutual information; this can be easily inferred from (13).

3.6.3 Bounds for the performance of the batch greedy algorithm

We now provide bounds for the submodularity and supermodularity ratios of the set function objective of (Prob-Max) in the setting of linear Bayesian inverse problems, where the objective is given by (13). These bounds enter our theoretical guarantees for the solution of (Prob-Max) using Algorithm 1.

Proposition 11 *In the linear–Gaussian setting defined in Section 3.6.1, the submodularity ratio γ and supermodularity ratio η pertinent to (Prob-Max) can be both lower bounded by $\frac{\log \zeta_{min}}{\log \zeta_{max}}$, where ζ is any generalized eigenvalue of the definite pair $\left(\Gamma_Y, \Gamma_{Y|X} \right)$.*

The proof is given in Appendix B.1.

Recall from Section 3.6.1 that $\zeta = 1 + \sigma$, where σ is any generalized eigenvalue of the definite pair $\left(\Gamma_Y - \Gamma_{Y|X}, \Gamma_{Y|X} \right)$. We know $\sigma_{\min} \geq 0$, and thus $\zeta_{\min} \geq 1$. If the least eigenvalue is 1, its algebraic multiplicity is the difference between the number of observations and the dimension of the inference parameters (Theorem 10). In such cases we obtain a trivial bound on the submodularity and supermodularity ratios. If the dimension of the inference parameters is greater than the number of observations, as is typically the case in inverse problems (for instance, consider the limit $n \to \infty$ analyzed by Stuart (2010)), we obtain a non-trivial bound on the same parameters. The empirical performance of Algorithm 1 is however impressive in all cases as will be demonstrated. Such gaps between worst-case bounds and practical performance are commonplace in algorithmic analysis. In this particular case the culprit is not necessarily the analysis framework, but a loose bound on the parameters featuring in the approximation guarantee.

3.7 Relationship to other algorithmic approaches

As we have previously remarked, our results Theorems 7 to 9 can be interpreted as the natural generalization of certain existing results (Das and Kempe, 2011; Mirzasoleiman et al., 2013, 2015) to the batch—and in some cases also to the non-submodular—settings. We now contrast our results with specific aspects of these works and other relevant investigations.

- Das and Kempe (2011) describe the so-called "oblivious" algorithm, wherein indices are selected ignoring any possible dependence among them. In the batch greedy algorithm we propose, at each step we also ignore any dependence among the indices that are selected. However we *do* account for dependencies with the indices already selected. When the batch size is equal to the cardinality constraint, i.e., $q = k$, or equivalently if all indices are chosen in one step, i.e., $l = 1$, the batch greedy heuristic is the oblivious algorithm. When we have a batch size greater than one but strictly less than the cardinality constraint, then the batch heuristic can be understood as the interleaving of the oblivious and classical greedy algorithms.

- Our ideas involving batch greedy heuristics are similar in spirit to adaptive sampling algorithms (Balkanski et al., 2018; Balkanski and Singer, 2018b,a; Chekuri and Quanrud, 2019b; Fahrbach et al., 2019; Chekuri and Quanrud, 2019a; Chen et al., 2019; Balkanski et al., 2019), in that both try to add a large set of elements at each step as opposed to growing the solution set incrementally. Qian and Singer (2019) analyze an

adaptive sampling algorithm for the maximization of non-submodular functions. That work introduces the notion of differential submodularity ($\alpha_{\mathrm{DF}} \in [0, 1]$), which captures how tightly incremental gains corresponding to a given non-submodular function can be bounded by incremental gains corresponding to submodular functions. The corresponding approximation guarantee for cardinality constraint maximization was shown to be $(1 - 1/e^{\alpha_{\mathrm{DF}}^2} - \epsilon)$, where $\epsilon > 0$ is a tolerance parameter in the adaptive sampling algorithm. Qian and Singer (2019) show how feature selection for regression and Bayesian A-optimal experimental design are γ^2-differentially submodular, where γ is the submodularity ratio. Using similar strategies, it can be seen that the mutual information-based Bayesian experimental design criterion is also γ^2-differentially submodular. Consequently, the adaptive algorithm by Qian and Singer (2019) corresponds to a $(1 - 1/e^{\gamma^4} - \epsilon)$ performance guarantee. In our case, however, we have a strictly better $(1 - 1/e^{\gamma^2})$ guarantee as described in Theorem 7. The approximation factor $(1 - 1/e^{\gamma^2})$ is obtained by recognizing that the submodularity and supermodularity ratios are lower bounded by the same term (see Proposition 11) and hence we can replace η by γ in (3). It is worth emphasizing that our approximation factor $(1 - 1/e^{\gamma^2})$ is independent of batch size, and has no tolerance parameter.

4. Sequential greedy algorithms based on the MM principle

We now explore algorithms of a slightly different flavor, which are still greedy in a sense but do not necessarily make the locally optimal choice corresponding to incremental gain. They are based on the MM (minorize–maximize or majorize–minimize) principle (Lange, 2016), and involve maximizing (resp. minimizing) at every step a minorizing (resp. majorizing) counterpart function to eventually obtain the $\arg\max$ (resp. $\arg\min$). Every instance of applying the MM principle is prompted by the need to transform a hard optimization problem into a sequence of simpler ones. Here simplicity can refer to one of many desirable attributes, such as reduction in overall computational complexity, linearization or convexification of the problem, easing the handling of complex constraints, and so on. For a broad survey of techniques and algorithms based on the MM principle we refer the readers to the excellent text by Lange (2016).

While most of the work in optimization using the MM principle has been in the setting of continuous functions, in the recent past these have been adapted to the world of set functions (Iyer et al., 2012, 2013). Narasimhan and Bilmes (2005); Iyer and Bilmes (2012), explored these ideas for the most general case of difference between two submodular functions. Concrete hardness results on multiplicative inapproximability of such optimization problems were also shown by Iyer and Bilmes (2012). The MM principle proves handy in these scenarios; it is deployed by first seeking tight modular bounds for one (or both) of the terms in the difference, and subsequently optimizing the resulting submodular (or modular) objective using existing methods. The key property of submodular functions that enables such approaches is that one can define a subgradient (Definition 27) and supergradient (Definition 28) at every point (Fujishige, 2005; Jegelka and Bilmes, 2011), meaning for every subset in the power set. Calculation of those semigradients, however, will require repeated function evaluations of appropriate subsets, and the semidifferential sets (Defini-

tions 27 and 28) that contain them constitute a polyhedral partitioning of $\mathbb{R}^m$ (Iyer and Bilmes, 2015).

Our efforts utilizing the MM principle differ from the approaches summarized above and perhaps to a certain degree from how MM algorithms are generally thought of. The techniques by Iyer et al. (2012, 2013); Iyer and Bilmes (2012) are applicable for fairly general constraints and can be best understood as an iterative framework where one continues until local convergence is achieved. The algorithms we consider, however, are sequential in a 'bonafide' sense, meaning an individual element or a batch of elements are selected at every step and appended to the previously chosen set of elements. Our focus is restricted to cardinality-constrained problems. We begin by providing a reinterpretation of the classical greedy heuristic.

4.1 Reinterpreting the classical greedy heuristic for maximization

The definitions of the submodularity and supermodularity ratios allow us to write the following bounds for the incremental gain $\rho_{\mathscr{A}}(\mathscr{B})$,

$$\eta \sum_{\nu \in \mathscr{A}} \rho_{\nu}(\mathscr{B}) \leq \rho_{\mathscr{A}}(\mathscr{B}) \leq \frac{1}{\gamma} \sum_{\nu \in \mathscr{A}} \rho_{\nu}(\mathscr{B}). \tag{15}$$

If we view $\rho_{\mathscr{A}}(\mathscr{B})$ as a function of $\mathscr{A}$, then what we have above are modular bounds on $\rho_{\mathscr{A}}(\mathscr{B})$ defined using the incremental gains $\rho_{\nu}(\mathscr{B})$, $\nu \in \mathscr{A}$. Alternatively if we view the set $\mathscr{B}$ as one individual entity and the elements of $\mathscr{A}$ as distinct entities, we can write modular bounds on $F(\mathscr{A} \cup \mathscr{B})$ in terms of $F(\mathscr{B})$ and $\rho_{\nu}(\mathscr{B})$, $\nu \in \mathscr{A}$:

$$F(\mathscr{B}) + \eta \sum_{\nu \in \mathscr{A}} \rho_{\nu}(\mathscr{B}) \leq F(\mathscr{A} \cup \mathscr{B}) \leq F(\mathscr{B}) + \frac{1}{\gamma} \sum_{\nu \in \mathscr{A}} \rho_{\nu}(\mathscr{B}). \tag{16}$$

Thus we can view the submodularity and supermodularity ratios as parameters that allow us to define the tightest possible modular bounds of the above form for any $\mathscr{A}, \mathscr{B} \subseteq \mathscr{V}$. Note that the bounds are also tight in the sense that if $\mathscr{A} \setminus \mathscr{B} = \emptyset$ then (16) reduces to a trivial statement. Viewed through the lens of MM algorithms, the modular lower/upper bound serves as the minorizing/majorizing counterpart to $F(\mathscr{A} \cup \mathscr{B})$.

With these perspectives, the greedy heuristic for maximization in Algorithm 1 can be easily paraphrased in the language of MM algorithms. Having selected the set $\mathscr{A}_{j-1}$ after $j-1$ steps, we seek $\mathscr{A}_{j-1 \to j} \subseteq \mathscr{V} \setminus \mathscr{A}_{j-1}$, $|\mathscr{A}_{j-1 \to j}| = q_j$ that maximizes $F(\mathscr{A}_{j-1} \cup \mathscr{A}_{j-1 \to j})$. We achieve this by maximizing the minorizing counterpart: the modular lower bound as in (16) defined using the incremental gains $\rho_{\nu}(\mathscr{A}_{j-1})$, $\nu \in \mathscr{V} \setminus \mathscr{A}_{j-1}$.

Incremental gains associated with individual elements, coupled with the submodularity or supermodularity ratios, are an obvious combination for defining modular bounds for a function. Many problems, however, may have additional attributes that enable other ways of defining modular bounds. The greedy heuristic can be adapted appropriately to utilize these special modular bounds. Such bounds could be easier to compute and perhaps perform better empirically. In Algorithm 4 we outline a greedy procedure built on the ansatz of an abstract modular lower bound $\mathcal{M}_{\uparrow}[\cdot]$ and modular upper bound $\mathcal{M}_{\downarrow}[\cdot]$ for the incremental gain $\rho_{\mathscr{A}}(\mathscr{B})$ associated with any $\mathscr{A}, \mathscr{B} \subseteq \mathscr{V}$:

$$\mathcal{M}_{\uparrow}[\rho_{\mathscr{A}}(\mathscr{B})] \leq \rho_{\mathscr{A}}(\mathscr{B}) \leq \mathcal{M}_{\downarrow}[\rho_{\mathscr{A}}(\mathscr{B})]. \tag{17}$$

The symbols $\downarrow, \uparrow$ are a visual cue to pinpoint whether the bound is from above or below.

Algorithm 4 Greedy algorithm using modular lower bounds

1: Input $F, \mathcal{V}, l, \{q_1, \ldots, q_l\}$
2: Initialize $\mathcal{A} = \emptyset$
3: **for** $i = 1$ to l **do**
4:     Determine $\mathcal{M}_\uparrow \left[\rho_{\mathcal{V} \setminus \mathcal{A}}(\mathcal{A}) \right]$.
5:     Find $\mathcal{Q} \subseteq \mathcal{V} \setminus \mathcal{A}, |\mathcal{Q}| = q_i$ that maximizes $\mathcal{M}_\uparrow \left[\rho_{\mathcal{Q}}(\mathcal{A}) \right]$
6:     $\mathcal{A} \leftarrow \mathcal{A} \cup \mathcal{Q}$
7: **end for**
        return Index set $\mathcal{A}$

Now we give performance bounds for the abstract batch greedy approach of Algorithm 4.

Theorem 12 *Let F be a non-decreasing function. Algorithm 4 for maximizing $F(\mathcal{A})$ subject to $|\mathcal{A}| \leq k$ outputs a set $\mathcal{A}$ such that*

$$F(\mathcal{A}) \geq \left(1 - \prod_{i=1}^{l} \left(1 - \frac{q_i}{k \left(1 + \tau_i \right)} \right) \right) \max_{\mathcal{B} \subset \mathcal{V}, |\mathcal{B}| \leq k} F(\mathcal{B})$$

where $\tau_i > 0$ is a parameter that encapsulates the effectiveness of the modular lower bound $\mathcal{M}_\uparrow [\cdot]$ in locally describing the function.

The proof is given in Appendix B.1.

The analysis of Algorithm 4 is along the same lines as that of Algorithm 1. The effectiveness of the modular lower bound in locally describing the function is baked into the approximation guarantee via the parameter τ_i. If $\mathcal{A}^*$ is the optimal set of the maximization problem, then for the ith step, τ_i is defined as the smallest scalar satisfying

$$\rho_{\mathcal{A}^*} \left(\mathcal{A}_{i-1} \right) - \mathcal{M}_\uparrow \left[\rho_{\mathcal{A}^*} \left(\mathcal{A}_{i-1} \right) \right] \leq \frac{\tau_i k \rho_{\mathcal{A}_i} \left(\mathcal{A}_{i-1} \right)}{q_i}. \tag{18}$$

The dependence of τ_i is strictly on the modular lower bound, but in many cases estimating it is easier when it is defined as the smallest scalar satisfying

$$\mathcal{M}_\downarrow \left[\rho_{\mathcal{A}^*} \left(\mathcal{A}_{i-1} \right) \right] - \mathcal{M}_\uparrow \left[\rho_{\mathcal{A}^*} \left(\mathcal{A}_{i-1} \right) \right] \leq \frac{\tau_i k \mathcal{M}_\uparrow \left[\rho_{\mathcal{A}_i} \left(\mathcal{A}_{i-1} \right) \right]}{q_i}. \tag{19}$$

Notice that τ_i estimated using the definition in (19) is always larger than what is necessary to satisfy (18).

One instance of τ_i which we have previously seen, indirectly, is when the modular bounds are defined using incremental gains and submodularity/supermodularity ratios. In that case, with a little effort it is easy to see that τ_i according to (19) will be $\frac{1 - \gamma_{\mathcal{V}, k} \eta_{\mathcal{V}, k}}{\gamma_{\mathcal{V}, k} \eta_{\mathcal{V}, q_i}}$. But this value is sub-optimal when compared to the τ_i one can implicitly infer from the result in Theorem 7, which is $\frac{1 - \gamma_{\mathcal{V}, k} \eta_{\mathcal{V}, q_i}}{\gamma_{\mathcal{V}, k} \eta_{\mathcal{V}, q_i}}$. The sub-optimality stems from the fact that $\eta_{\mathcal{V}, q_i} \geq \eta_{\mathcal{V}, k}$, $\forall q_i \leq k$. The discrepancy is easily explained since in the proof of Theorem 12 we allow for the modular lower bound $\mathcal{M}_\uparrow [\cdot]$ to be negative. If we assume non-negativity of the modular lower bound, the proof can be accordingly modified to obtain the optimal result.

4.2 Adapting MM techniques for linear Bayesian optimal experimental design

In this section we develop specialized MM-based algorithms for the linear Bayesian optimal experimental design problem formulated in Section 3.6.2.

4.2.1 Modular bounds for mutual information design criteria

In Section 3.6.2 we discussed some key spectral properties of design criteria based on mutual information. Now we derive helpful majorizations and minorizations using tight modular bounds on these information theoretic objectives.

Proposition 13 *Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ be jointly Gaussian random variables as defined in Section 3.6.1. Let $\mathcal{P} \in \mathbb{R}^{m \times k}$ be a selection operator such that $Y_{\mathcal{P}} := \mathcal{P}^\top Y = [Y_{i_1}, \ldots, Y_{i_k}]^\top$. The mutual information between X and $Y_{\mathcal{P}}$ can be bounded on both sides by the following modular functions:*

$$\mathcal{I}(X; Y_{\mathcal{P}}) \geq \frac{1}{2}\text{trace}\left(\mathcal{P}^\top \left(\log\left(\Gamma_Y\right) - \log \text{diag}\left(\Gamma_{Y|X}\right)\right)\mathcal{P}\right),$$

$$\mathcal{I}(X; Y_{\mathcal{P}}) \leq \frac{1}{2}\text{trace}\left(\mathcal{P}^\top \left(\log \text{diag}\left(\Gamma_Y\right) - \log\left(\Gamma_{Y|X}\right)\right)\mathcal{P}\right).$$

The proof is given in Appendix B.3.

Note that the lower bound in the proposition above is always positive when mutual information is submodular (Corollary 14), but in general it need not be. The machinery used to obtain Proposition 13 exploits the linear algebraic structure of the information theoretic objectives. Using the same arsenal of tools we can provide the following alternative bounds:

$$\mathcal{I}(X; Y_{\mathcal{P}}) \geq \frac{1}{2}\text{trace}\left(\mathcal{P}^\top \left(\log\left(\Gamma_Y\right) - \log\left(\Gamma_{Y|X}\right) + I_m \log\left(\varrho_1\right)\right)\mathcal{P}\right), \tag{20a}$$

$$\mathcal{I}(X; Y_{\mathcal{P}}) \leq \frac{1}{2}\text{trace}\left(\mathcal{P}^\top \left(\log\left(\Gamma_Y\right) - I_m \log\left(\varrho_2\right) - \log\left(\Gamma_{Y|X}\right)\right)\mathcal{P}\right). \tag{20b}$$

We include arguments for these bounds (20) in the proof of Proposition 13 as well. The constants $\varrho_1, \varrho_2 \in (0, 1]$ depend on the range of the spectrum of covariance operators $\Gamma_{Y|X}, \Gamma_Y$ respectively. If $\text{eig}(\Gamma_{Y|X}) \in [\alpha_s, \alpha_l] \subset \mathbb{R}_{>0}$ and $\text{eig}(\Gamma_Y) \in [\beta_s, \beta_l] \subset \mathbb{R}_{>0}$, then $\varrho_1 := \frac{4\alpha_s \alpha_l}{(\alpha_s + \alpha_l)^2}$ and $\varrho_2 := \frac{4\beta_s \beta_l}{(\beta_s + \beta_l)^2}$. Note that lower bound in (20) is not guaranteed to be positive even when mutual information is submodular, and the bounds in (20) are tighter when the cardinality of $\mathscr{I}(\mathcal{P})$ is smaller.

To understand the modular bounds in Proposition 13 and their significance, it is helpful to discuss two simple corollaries. We begin with the more restricted of the two, which is valid only when the observations are conditionally independent.

Corollary 14 *For random variables $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ as defined in Proposition 13, if $Y_{i_k}|X$ are additionally independent, then the following bounds hold for any selection operator $\mathcal{P} \in \mathbb{R}^{m \times k}$:*

$$\mathcal{I}(X; Y) - \mathcal{I}(X; Y \setminus Y_{\mathcal{P}}) \leq \frac{1}{2}\text{trace}\left(\mathcal{P}^\top \left(\log\left(\Gamma_Y\right) - \log\left(\Gamma_{Y|X}\right)\right)\mathcal{P}\right) \leq \mathcal{I}(X; Y_{\mathcal{P}}).$$

The proof is given in Appendix B.3.

Recall that $\mathcal{I}(X;Y_{\mathcal{P}})$ is the objective corresponding to (Prob-Max), and $\mathcal{I}(X;Y) - \mathcal{I}(X;Y \setminus Y_{\mathcal{P}})$ is the objective corresponding to (Prob-Min). If $\mathcal{P}$ is square with its corresponding index set $\mathscr{I}_{\mathcal{P}} = \mathscr{V}$, then the corollary statement reduces to a simple identity. In all other circumstances the inequalities are strict.

We now highlight a few properties encapsulated by Corollary 14 which relate to concepts in polyhedral combinatorics.

- We know the mutual information $\mathcal{I}(X;Y_{\mathcal{P}})$ is submodular when observations are conditionally independent (Proposition 4). The term $\mathcal{I}(X;Y) - \mathcal{I}(X;Y \setminus Y_{\mathcal{P}})$ is the supermodular dual (Definition 29) of $\mathcal{I}(X;Y_{\mathcal{P}})$, and both the functions have the same base polytope (Definition 24).

- The diagonal of the matrix appearing in the middle term deserves special attention. We denote the diagonal of $\frac{1}{2}\left(\log\left(\Gamma_Y\right) - \log\left(\Gamma_{Y|X}\right)\right)$ by the vector $d \in \mathbb{R}^m$. Note that the covariance $\Gamma_{Y|X}$ in this context is diagonal since we are presently discussing the case of uncorrelated observation error. The vector d can be inferred to be a point in the base polytope (Definition 24) of $\mathcal{I}(X;Y_{\mathcal{P}})$ or equivalently of $\mathcal{I}(X;Y) - \mathcal{I}(X;Y \setminus Y_{\mathcal{P}})$. As a result, it is naturally a subgradient (Definition 27) of $\mathcal{I}(X;Y_{\mathcal{P}})$ and a supergradient of $\mathcal{I}(X;Y) - \mathcal{I}(X;Y \setminus Y_{\mathcal{P}})$ defined at $\mathscr{V}$ or $\emptyset$. The individual components of d are all positive, which must be the case since the base polytope lies in the positive orthant for all non-decreasing submodular functions.

- Most interestingly, d is not necessarily one of the extreme points of the base polytope but rather can lie in its interior. Note that there are at most $m!$ extreme points and these are at least in theory determinable using the 'greedy algorithm' (see Edmonds, 1971, Rado–Edmonds theorem). Typically subgradients defined using extremal points are used to drive the optimization steps (Iyer et al., 2012, 2013; Iyer and Bilmes, 2012). Using a non-extremal point to drive optimization, as we do here, is a novel approach, albeit specific to the setting of linear–Gaussian Bayesian problems.

We now consider a generalization of Corollary 14 pertinent to the case of correlated observation errors. Recall that in this case the mutual information $\mathcal{I}(X;Y_{\mathcal{P}})$ is neither submodular nor supermodular, but simply a non-decreasing set function.

Corollary 15 *For random variables $X \in \mathbb{R}^n, Y \in \mathbb{R}^m$ as defined in Proposition 13, the following bounds hold for any selection operator $\mathcal{P} \in \mathbb{R}^{m \times k}$:*

$$\mathcal{I}(X;Y) - \mathcal{I}(X;Y \setminus Y_{\mathcal{P}}) \leq \frac{1}{2}\text{trace}\left(\mathcal{P}^{\top}\left(\log\left(\Gamma_Y\right) - \log\left(\Gamma_{Y|X}\right)\right)\mathcal{P}\right) + \mathfrak{D}_{\mathcal{P}^c}, \qquad (21)$$

$$\Longleftrightarrow \mathcal{I}(X;Y_{\mathcal{P}}) \geq \frac{1}{2}\text{trace}\left(\mathcal{P}^{\top}\left(\log\left(\Gamma_Y\right) - \log\left(\Gamma_{Y|X}\right)\right)\mathcal{P}\right) - \mathfrak{D}_{\mathcal{P}}, \qquad (22)$$

where $\mathfrak{D}_{\mathcal{P}} := \frac{1}{2}\text{trace}\left(\mathcal{P}^{\top}\left(\log\text{diag}\left(\Gamma_{Y|X}\right) - \log\left(\Gamma_{Y|X}\right)\right)\mathcal{P}\right)$.

The proof is given in Appendix B.3.

The term $\mathfrak{D}_{\mathcal{P}}$ is non-negative and strictly increases with increasing cardinality of the index set $\mathscr{I}(\mathcal{P})$. When the observation errors are uncorrelated, meaning that the covariance

21

operator $\Gamma_{Y|X}$ is diagonal, $\mathfrak{D}_{\mathcal{P}}$ evaluates to zero and we recover the statement in Corollary 14 by combining (21) and (22). Adding (21) and (22) we obtain,

$$\mathcal{I}(X;Y) - \mathcal{I}(X;Y \setminus Y_{\mathcal{P}}) \leq \mathcal{I}(X;Y_{\mathcal{P}}) + \frac{1}{2}\text{trace}\left(\log \text{diag}\left(\Gamma_{Y|X}\right) - \log\left(\Gamma_{Y|X}\right)\right). \qquad (23)$$

Loosely speaking, $\mathfrak{D} := \frac{1}{2}\text{trace}\left(\log \text{diag}\left(\Gamma_{Y|X}\right) - \log\left(\Gamma_{Y|X}\right)\right)$ reflects the deviation from submodularity of the mutual information $\mathcal{I}(X;Y_{\mathcal{P}})$: the smaller its magnitude, the more $\mathcal{I}(X;Y_{\mathcal{P}})$ behaves as a submodular function. If $\mathfrak{D}_{\mathcal{P}^c}$ and $\mathfrak{D}_{\mathcal{P}}$ in Corollary 15 are replaced by $\mathfrak{D}$, the inequalities are akin to affine modular bounds on the set functions $\mathcal{I}(X;Y) - \mathcal{I}(X;Y \setminus Y_{\mathcal{P}})$ and $\mathcal{I}(X;Y_{\mathcal{P}})$. The pair $(d, \mathfrak{D})$, where the vector $d \in \mathbb{R}^m$ is the diagonal of $\frac{1}{2}\left(\log\left(\Gamma_Y\right) - \log\left(\Gamma_{Y|X}\right)\right)$, characterizes the affine modular bound in Corollary 15. Such a pair in general defines generalized lower and upper polyhedra (Definition 26) for arbitrary set functions that need not be sub/super-modular. Iyer and Bilmes (2015) use those theoretical constructs to study polyhedral aspects of submodular functions.

4.2.2 MM ALGORITHMS FOR MAXIMIZING INFORMATION GAIN

For the problem of maximizing information gain (`Prob-Max`), wherein the objective is non-decreasing, we have previously discussed the performance of the standard batch greedy algorithm (Algorithm 1) in Section 3.6. We now consider Algorithm 4, implemented with the modular lower bounds for incremental gain given in Proposition 13. We will refer to this algorithm as `MMGreedy` in the comparative discussion of Section 5. Note that incremental gain in the present context corresponds to information gain, which in turn can be written as conditional mutual information. The inequalities in Proposition 13 can be suitably adapted by replacing the covariance operators with the appropriate conditional covariances. We describe this process below.

Suppose our goal is to select $k < m$ observations altogether, and we have already selected $k_1 < k$. Let the indices of the selected k_1 observations correspond to the selection operator $\mathcal{P}_1$, $\mathscr{I}_{\mathcal{P}_1} \subset \mathscr{V}$. We now seek the remaining $k_2 = k - k_1$ observations whose corresponding index selection operator $\mathcal{P}_2$, with $\mathscr{I}_{\mathcal{P}_2} \subset \mathscr{V} \setminus \mathscr{I}_{\mathcal{P}_1}$, is $\arg\max_{\mathcal{P},|\mathscr{I}_{\mathcal{P}}|=k_2} \mathcal{I}(X;Y_{\mathcal{P}}|Y_{\mathcal{P}_1})$. In the proposed MM framework, this is achieved by maximizing the minorizing modular lower bound,

$$\frac{1}{2}\text{trace}\left(\widehat{\mathcal{P}}^\top \left(\log\left(\Gamma_{Y_{\mathcal{P}_1^c}|Y_{\mathcal{P}_1}}\right) - \log \text{diag}\left(\Gamma_{Y_{\mathcal{P}_1^c}|X,Y_{\mathcal{P}_1}}\right)\right)\widehat{\mathcal{P}}\right) \leq \mathcal{I}(X;Y_{\mathcal{P}}|Y_{\mathcal{P}_1}). \qquad (24)$$

In (24), $\mathcal{P}_1^c$ is the complement of $\mathcal{P}_1$ with $\mathscr{I}_{\mathcal{P}_1^c} = \mathscr{V} \setminus \mathscr{I}_{\mathcal{P}_1}$. The selection operator $\widehat{\mathcal{P}}$ in (24) has the same index set as $\mathcal{P}$ and is simply its counterpart when selecting from the reduced dimension random variable $Y_{\mathcal{P}_1^c} \in \mathbb{R}^{m-k_1}$. The approximation guarantee in this case follows from Theorem 12. Bounding the parameter τ_i (see Theorem 12) that encapsulates the effectiveness of the modular lower bound is a non-trivial exercise which we will not pursue. Our empirical results (see Section 5), however, demonstrate that the performance of the algorithm is comparable to that of standard batch greedy (Algorithm 1) and in some instances slightly better.

4.2.3 MM algorithms for minimizing information loss

Our motivation to independently study minimizing information loss is triggered by a combination of factors. The asymmetry that exists between incrementally choosing good observations—as opposed to discarding bad observations—is perhaps the most intriguing reason. In the context of linear Bayesian inverse problems, Corollary 14 encapsulates our motivation perfectly. Observe how the same modular bound proves useful in selecting the best observation and discarding the worst. Yet each procedure carried out sequentially, by updating the modular bound, yields different answers in general. While Corollary 14 is restricted to the submodular case with conditionally independent observations, we have the generalization in Corollary 15 with slightly different modular bounds.

(`Prob-Min`) is the formal representation of the problem of minimizing information loss. Exploiting the previously discussed modular bounds for mutual information, we propose the following approach for its solution. Suppose our goal is to discard $k < m$ observations, and we have already removed $k_1 < k$. Let the indices of the discarded k_1 observations correspond to the selection operator $\mathcal{P}_1$, $\mathscr{I}_{\mathcal{P}_1} \subset \mathscr{V}$. We now seek $k_2 := k - k_1$ observations whose corresponding index selection operator $\mathcal{P}_2$, with $\mathscr{I}_{\mathcal{P}_2} \subset \mathscr{V} \setminus \mathscr{I}_{\mathcal{P}_1}$, is the solution of the following problem:

$$\mathcal{P}_2 = \underset{\mathcal{P},|\mathscr{I}_{\mathcal{P}}|=k_2}{\arg\min} \ \mathcal{I}(X;Y) - \mathcal{I}(X;Y \setminus Y_{\mathcal{P}_1}, Y_{\mathcal{P}}) = \underset{\mathcal{P},|\mathscr{I}_{\mathcal{P}}|=k_2}{\arg\min} \ \mathcal{I}(X;Y|Y_{\mathcal{P}_1}) - \mathcal{I}(X;Y \setminus Y_{\mathcal{P}_1}, Y_{\mathcal{P}}).$$

$$(25)$$

We solve (25) using Algorithm 5 by minimizing the majorizing modular upper bound, as given in Proposition 16. We refer to this algorithm as `MMReverseGreedy`, alluding to the fact that we remove indices sequentially from the candidate set until we reach the desired cardinality. While it is unclear if this approach can be supplemented with an approximation guarantee, numerical results indicate excellent empirical performance.

Algorithm 5 Sequential greedy algorithm for minimizing information loss

---

1: Input $G, \Gamma_X, \Gamma_{Y|X}, \mathscr{V}, l, \{q_1, \ldots, q_l\}$
2: Initialize $\mathscr{I}(\mathcal{P}_r) = \emptyset$
3: **for** $i = 1$ to l **do**
4:   Determine $\mathcal{M}^{\downarrow}[\mathcal{I}(X;Y|Y_{\mathcal{P}_r}) - \mathcal{I}(X;Y \setminus Y_{\mathcal{P}_r}, Y_{\mathcal{P}})]$ as per Proposition 16.
5:   Find $\mathcal{P}$ with $\mathscr{I}_{\mathcal{P}} \subseteq \mathscr{V} \setminus \mathscr{I}_{\mathcal{P}_r}$ and $|\mathscr{I}_{\mathcal{P}}| = q_i$ that minimizes $\mathcal{M}^{\downarrow}[\mathcal{I}(X;Y|Y_{\mathcal{P}_r}) - \mathcal{I}(X;Y \setminus Y_{\mathcal{P}_r}, Y_{\mathcal{P}})]$
6:   $\mathscr{I}_{\mathcal{P}_r} \leftarrow \mathscr{I}_{\mathcal{P}_r} \cup \mathscr{I}_{\mathcal{P}}$
7: **end for**
    return Index set $\mathscr{I}_{\mathcal{P}_r}$

Proposition 16 *For random variables $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ as defined in Proposition 13, and selection operators $\mathcal{P} \in \mathbb{R}^{m \times k}$ and $\mathcal{P}_1 \in \mathbb{R}^{m \times k_1}$ such that $k_1 < k$, the following bound*

holds:

$$\mathcal{I}(X;Y|Y_{\mathcal{P}_1}) - \mathcal{I}(X;Y \setminus Y_{\mathcal{P}_1}, Y_{\mathcal{P}}) \le \mathrm{trace}\left(\widehat{\mathcal{P}}^{\top}\left(\log\left(\Gamma_{Y|Y_{\mathcal{P}_1}}\right) - \log\left(\Gamma_{Y|X,Y_{\mathcal{P}_1}}\right)\right)\widehat{\mathcal{P}}\right)$$
$$+ \mathrm{trace}\left(\widehat{\mathcal{P}}^{c\top}\left(\log\mathrm{diag}\left(\Gamma_{Y|X,Y_{\mathcal{P}_1}}\right) - \log\left(\Gamma_{Y|X,Y_{\mathcal{P}_1}}\right)\right)\widehat{\mathcal{P}}^{c}\right)$$
$$+ \mathrm{trace}\left(\widehat{\mathcal{P}}^{c\top}\left(\log\mathrm{diag}\left(\Gamma_{Y\setminus Y_{\mathcal{P}_1}|X}\right) - \log\left(\Gamma_{Y|X,Y_{\mathcal{P}_1}}\right)\right)\widehat{\mathcal{P}}^{c}\right).$$

Here $\widehat{\mathcal{P}}$ is the counterpart of $\mathcal{P}$ when selecting from the reduced dimension random variable $Y_{\mathcal{P}_1^c} \in \mathbb{R}^{m-k_1}$.

The proof is given in Appendix B.3.

The two trace terms involving $\widehat{\mathcal{P}}^c$ are non-negative and identically zero if the observations are conditionally independent. In that scenario, the proposition statement reduces to a slightly modified version of Corollary 14.

4.2.4 NUMERICAL AND COMPUTATIONAL ISSUES CONCERNING MM ALGORITHMS

Our aim here is to provide some technical information on how to compute the modular bounds, and the associated computational costs. With regard to computational complexity, our focus here is not the number of function calls to the value oracle model, but complexity in terms of number of floating point operations (FLOPs). This is a more pertinent comparative metric since the optimization is performed using a minorizing/majorizing surrogate, not the actual objective itself. We will use the symbol $\mathscr{O}(\cdot)$ when referring to FLOP count, rather than $\mathcal{O}(\cdot)$, which we previously used to specify the number of function evaluations.

For the algorithms outlined in Sections 4.2.2 and 4.2.3, one critical task is estimating the diagonal of the matrix logarithm of a definite matrix. In each iteration, the matrix of interest is obtained using a Schur complement operation. More precisely, it is the Schur complement corresponding to one of the principal submatrices of the (larger) matrix in the previous iteration. The computational overhead of the Schur complement is moderate compared to that of determining the diagonal entries of the matrix logarithm. The cumulative cost of the latter task across all the iterations is what determines the complexity of the algorithm.

If the objective is submodular, and when the batch size is one, it is easy to see that we only require the relative ordering of the diagonal entries as opposed to absolute accuracy of each. However there is no direct way to obtain relative ordering and, surprisingly, nor is there a direct efficient way to estimate the diagonal of a matrix logarithm. At first glance, it is tempting to consider estimating each entry of the diagonal as the bilinear form $e_i^{\top}\log(A)e_i$, but such efforts (Golub and Meurant, 2009; Bai et al., 1996; Golub and Meurant, 1994) are quickly realized to be inefficient in this case. Avoiding computation of the matrix logarithm in all but special cases in not an option. We refer to the readers to Higham (2008, Chap. 11) for a comprehensive survey on computing matrix logarithms.

In our case, since we seek the logarithm of a definite matrix, we can use its eigenvalue decomposition. Recall that for a definite matrix $A \in \mathbb{R}^{m\times m}$ with eigendecomposition $A = U\Sigma U^*$, its matrix logarithm is simply $\log A = U\log\Sigma U^*$. Note how any favorable decay in the spectrum of the matrix A may not necessarily aid the estimation of its matrix logarithm, since eigenmodes corresponding to the smallest eigenvalues could also dominate

the estimate. In special cases, though, randomized methods that provide inexpensive low rank approximations of a matrix could still be useful.

In general, however, the leading order complexity for the algorithms in Sections 4.2.2 and 4.2.3 is $\mathscr{O}(\frac{m^3 k}{q})$. Here m is the size of the candidate set, k is the desired cardinality, and q is the batch size. The dominating cost is that of determining the sub/super–gradient using the modular bounds of Proposition 13. Asymptotically, we do not improve on the complexity of batch greedy based on incremental gains (Algorithm 1); the dominating cost there is that of evaluating log determinants.

4.2.5 Contextualizing with other approaches

In various simplified settings, the sequential methods outlined in Sections 4.2.2 and 4.2.3 for the linear Bayesian optimal experimental design problem share similarities to existing techniques. Consider the case of independent and identically distributed observation errors; in this case the objective in (11) is submodular and defined entirely by the data marginal π_Y. (Prob-Max) is now equivalent to finding the mode of a fixed-size determinantal point process (DPP) (Kulesza and Taskar, 2011, 2012), with the kernel of the DPP being the marginal covariance of the data, Γ_Y.

In the MM approach of Section 4.2.2, we define a tight modular lower bound on the objective using the trace of the logarithm of the data marginal covariance, or its appropriate conditional counterpart. The individual diagonal entries of the matrix logarithm are akin to *weighted* leverage scores associated with each row/column index; here the weights are simply the logarithms of the eigenvalues. Historically, the notion of leverage scores was introduced in the context of linear regression for outlier detection, and to assess the amount of influence exerted by one observation regardless of its actual value (Chatterjee and Hadi, 1986; Hoaglin and Welsch, 1978). In the recent past such scores have been useful in performing subset selection (Boutsidis et al., 2009) and in linear regression (Drineas et al., 2011), both of which directly relate to optimal experimental design.

Sampling based on leverage scores versus DPPs. Volume sampling in DPPs and sampling based on traditional statistical leverage scores are clearly different. But in precisely what way? Contrasting their differences and understanding the nuances is a useful exercise. Consider a DPP defined through an L-ensemble specified by $\mathcal{L} \succeq 0$; often the semi-definite matrix $\mathcal{L}$ is also termed the kernel matrix. One popular interpretation is obtained by expressing $\mathcal{L}$ as a Gram matrix, $\mathcal{B}^\top \mathcal{B}$, where the columns of $\mathcal{B}$ are feature vectors representing items in the candidate set. Furthermore, if each column B_i is written as the product of a quality term $q_i \in \mathbb{R}_{>0}$ and a vector of normalized diversity features $\phi_i \in \mathbb{R}^d$, $\|\phi_i\| = 1$, then we have $\mathcal{L}_{ij} = q_i \phi_i^\top \phi_j q_j$. Now suppose that $\mathcal{U}\Lambda\mathcal{U}^\top$ is the eigendecomposition of $\mathcal{L}$. Then it is easy to see that $q_i^2 = \mathcal{U}\Lambda\mathcal{U}^\top_{[i,i]}$. The diagonal entries in $\mathcal{U}\Lambda\mathcal{U}^\top$ are strictly speaking not statistical leverage scores, since we are weighing the contribution of each eigenvector by its corresponding eigenvalue. But note that when using randomized numerical linear algebra methods, the practice is always to give more importance to higher eigenmodes. In the context of this comparative discussion, it is evident that sampling based on leverage scores does not account for diversity among the candidates, but accounts only for the quality of each one.

5. Numerical results

In this section we evaluate the performance of the batch greedy algorithms when applied to linear Bayesian optimal experimental design. Our focus here to is to demonstrate how batch size affects performance and to evaluate the impact of different modular bounds; hence we only consider the standard greedy algorithm and the MM algorithms. The performances of distributed and stochastic variants of the greedy heuristic, compared to the standard greedy paradigm, have been showcased previously by Mirzasoleiman et al. (2013, 2015). In the discussion and figures that follow, we label the approach of maximizing information gain using Algorithm 1 as `StdGreedy`; the approach of maximizing information gain using the minorizing surrogate (Algorithm 4 and Section 4.2.2) as `MMGreedy`; and analogously, the approach of minimizing information loss using the majorizing surrogate (Algorithm 5 and Section 4.2.3) as `MMReverseGreedy`.

5.1 An inverse problem with structured random operators

We consider a Bayesian inference problem in the setting of Section 3.6.1, with a randomly generated linear forward operator G. The dimension of the parameters X is set to $n = 20$, while cardinality of the candidate set of observations Y is fixed at $m = 100$. Realizations of the forward operator are constructed by independently generating random singular values and left/right singular vectors. The spectrum of G, while random, has a prescribed exponential rate of decay reflective of many real world inverse problems (Spantini et al., 2015). The random left and right singular vectors are obtained by taking a modified Gram-Schmidt QR factorization of a random normal matrix of the appropriate dimension. We specify the prior and observation error covariances, Γ_X and $\Gamma_{Y|X}$, using squared exponential kernels with correlation lengths 0.105 and 0.021, respectively, on a unit domain. We draw 1000 random instances of the forward operator G and solve the experimental design problem of maximizing mutual information $\mathcal{I}(X, \mathcal{P}^\top Y)$ in each case. The spectra of the relevant operators are shown in Figure 2. If a spectrum corresponds to a random operator, we plot the median and indicate the spread.

In Figure 3 we illustrate the influence of batch size for each algorithm. On the horizontal axis we indicate the number of observations included, and on the vertical axis we plot the amount of mutual information captured at that cardinality relative to the maximum amount possible (i.e., using all the candidate observations Y). We consider seven different batch sizes, corresponding to $q \in \{1\%, 10\%, 20\%, 30\%, 40\%, 50\%, 100\%\}$. (Here we have expressed the batch size q as a fraction of the size of the candidate set.) For each algorithm and for each choice of batch size, Figure 3 shows the median performance across the random instances of the forward model. Figure 4 illustrates the variability in performance across these random problem instances; here, in addition to the median mutual information, we show the 0.1/0.9 quantiles and range of mutual information obtained at each cardinality. Kinks in the continuous lines of Figures 3–4 reflect the start of each new batch calculation. Improved gains at smaller batch sizes are consistent with our theoretical claims in Theorems 7 and 12. The numerical results also corroborate our remark in Section 3.6, that the trivial approximation guarantee for $m > n$ is merely an artifact of loose bounds on the sub/super-modularity ratios. Interestingly, while we do not have a theoretical analysis of
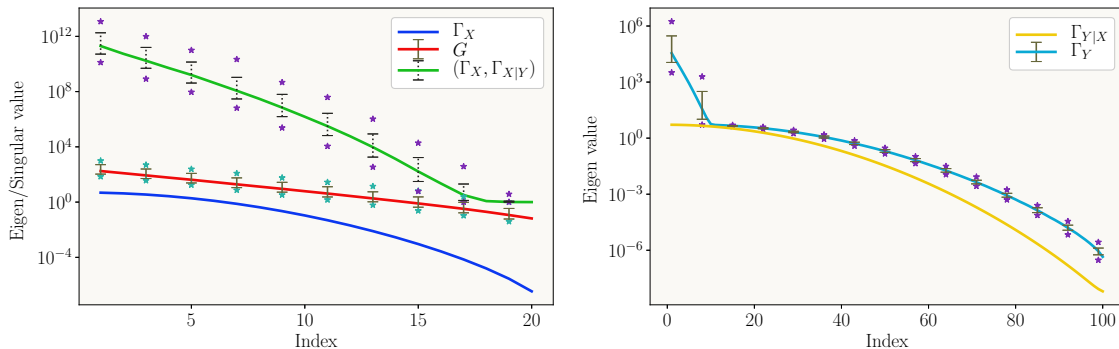
Figure 2: Spectrum of the relevant operators of the inverse problem with correlated observation error. The solid line is the median across 1000 random instances of the forward model. The whiskers capture the interquantile range (10% to 90%), and the $\star$ symbols mark the maximum and minimum eigen/singular value. The prior and observation error covariances are not random.

the `MMReverseGreedy` algorithm, the effect of batch size on its performance is similar to that of the other two heuristics.

In Figure 5, we compare the performance of the greedy algorithms, for a range of batch sizes, to that of a *random* selection of indices. To obtain the latter results, we draw 1000 random index selections for each random forward model. In general, the greedy algorithms vastly outperform random selection. The median performance of even the one-shot ($q = 100\%$) greedy approaches is better than the median performance of random selection. This difference is greatest for the `MMGreedy` and `MMReverseGreedy` algorithms, which show significantly better performance at large batch sizes than `StdGreedy`. As shown in the right two panels of Figure 5, the median performance of these two greedy heuristics is better than even the 90% quantile of performance of random selection.
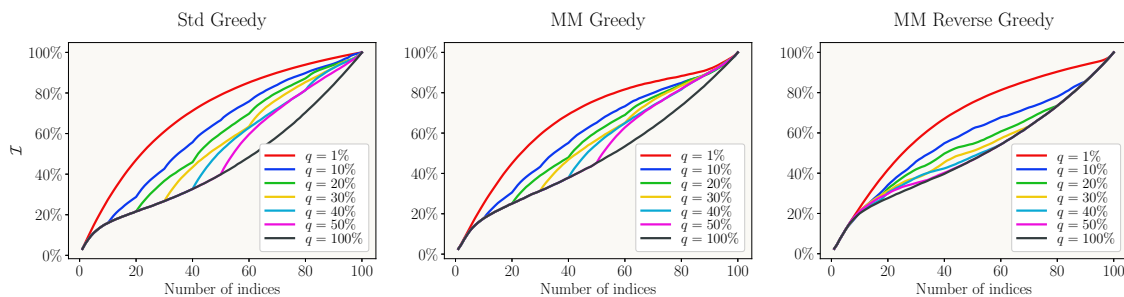
Figure 3: Performance of each greedy heuristic for different batch sizes, ranging from single index selection ($q = 1\%$) to a one-shot approach ($q = 100\%$). The solid line is the median across 1000 random instances of the forward model.

Fixing the batch size, we compare the three greedy algorithms' performance more directly in Figure 6. We consider the cases $q \in \{1\%, 20\%, 40\%, 100\%\}$. Once again, we only
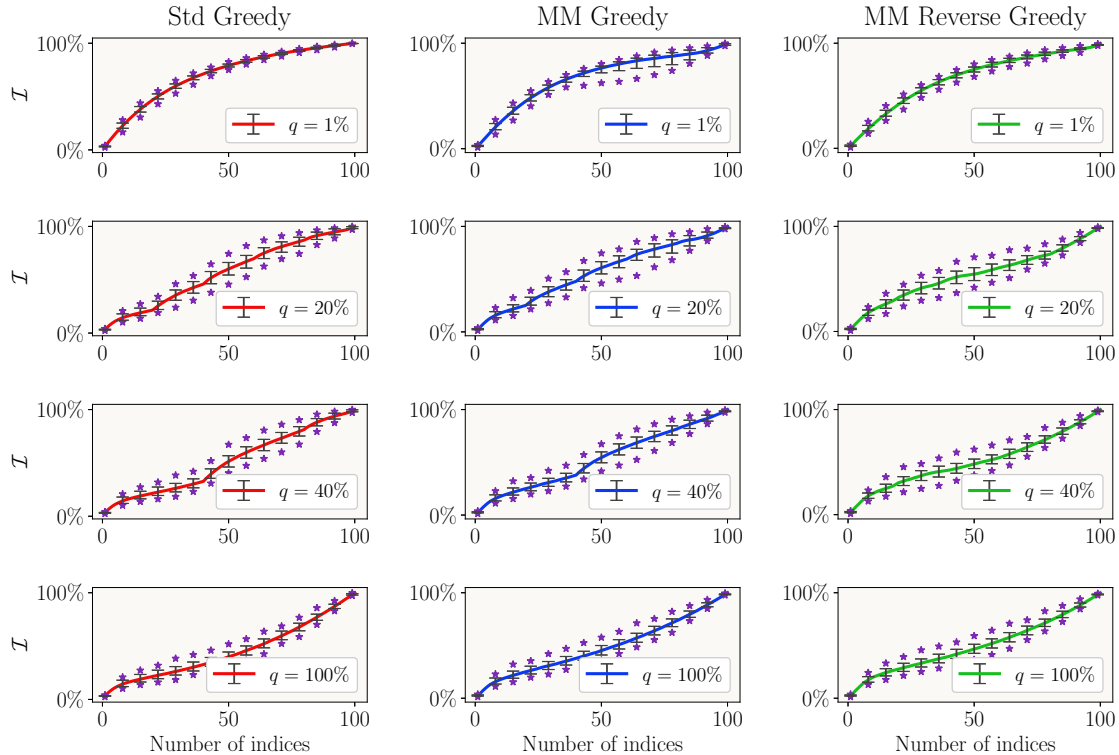
Figure 4: Performance of the greedy heuristics for different batch sizes across all 1000 random instances of the forward model. The solid line is the median; the whiskers bound the 10% to 90% interquantile range, and the ⋆ marks the maximum and minimum mutual information captured.

plot the median performance across 1000 random instances of the forward model to retain visual clarity. When choosing indices one at a time ($q = 1\%$) the standard greedy heuristic does marginally better than both MM greedy heuristics, but at larger batch sizes, both MM approaches provide better gains. This distinction is more readily apparent in Figure 7, where we plot the difference between the relative amount of information captured by `MMGreedy` and `MMReverseGreedy` in comparison to the standard greedy heuristic. Here we also indicate the spread due to the forward model being random. The differences in the performances of the three heuristics are stark at relatively lower cardinalities, and diminish at higher cardinality since information saturates. The upside to using the `MMGreedy` heuristic in comparison to the standard greedy heuristic is much greater than the downside, as indicated by the interquantile range and the maximum/minimum mutual information captured.

The results we have discussed thus far are for the case of correlated but structured observation error covariance. Smaller batch sizes clearly yield more information gain, and should be preferred unless computational demands dictate otherwise. But many realistic problems have error terms with a less structured correlation (see Section 5.2), or independent and identically distributed (i.i.d.) observation errors. In the latter case, the experimental
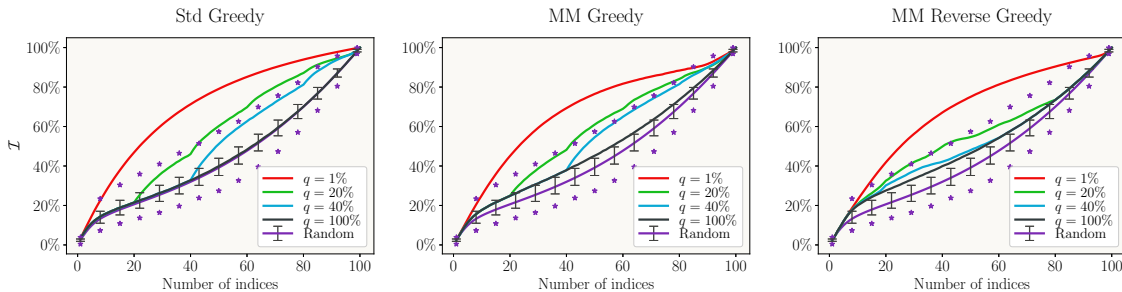
28

Figure 5: Performance of each greedy heuristic, for different batch sizes q, compared to a random selection of indices. The solid line corresponding to a given batch size is the median across 1000 random instances of the forward model. The solid line labeled "random" is the median performance across 10^6 cases: 1000 random instances of the forward model tensorized with 1000 random selections of indices. For the random selection results, the whiskers bound the 10% to 90% interquantile range and the $\star$ symbols mark the maximum and minimum mutual information captured. Note that the random selection results are the same in each of the three panels above; only the batch greedy algorithms are different.
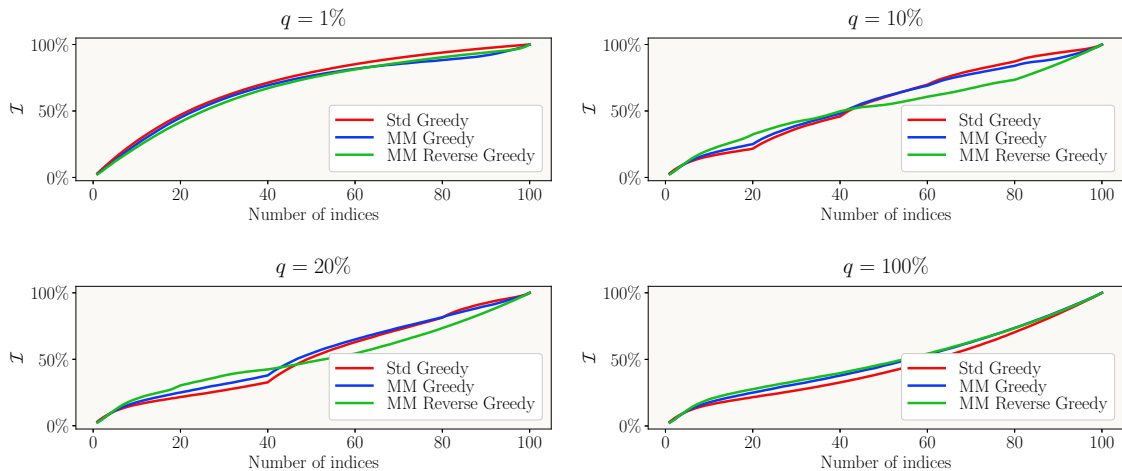
Figure 6: A comparative study of the greedy heuristics for four different batch sizes. The solid line is the median across 1000 random instances of the forward model.

design objective is also submodular. In these scenarios, the use of a smaller batch size may be unwarranted, as our numerical experiments in Section 5.2 indicate. More generally, if the observation error covariance is not strongly correlated, meaning the observations are nearly conditionally independent, the advantage of a smaller batch size is diminished.

We have also numerically investigated the performance of the algorithms when the dimension of the parameters is greater than the cardinality of the candidate set of observations, i.e., $n > m$. Recall that in such cases the results in Proposition 11 provide non-trivial bounds on the sub/super–modularity ratios. The StdGreedy and MMGreedy
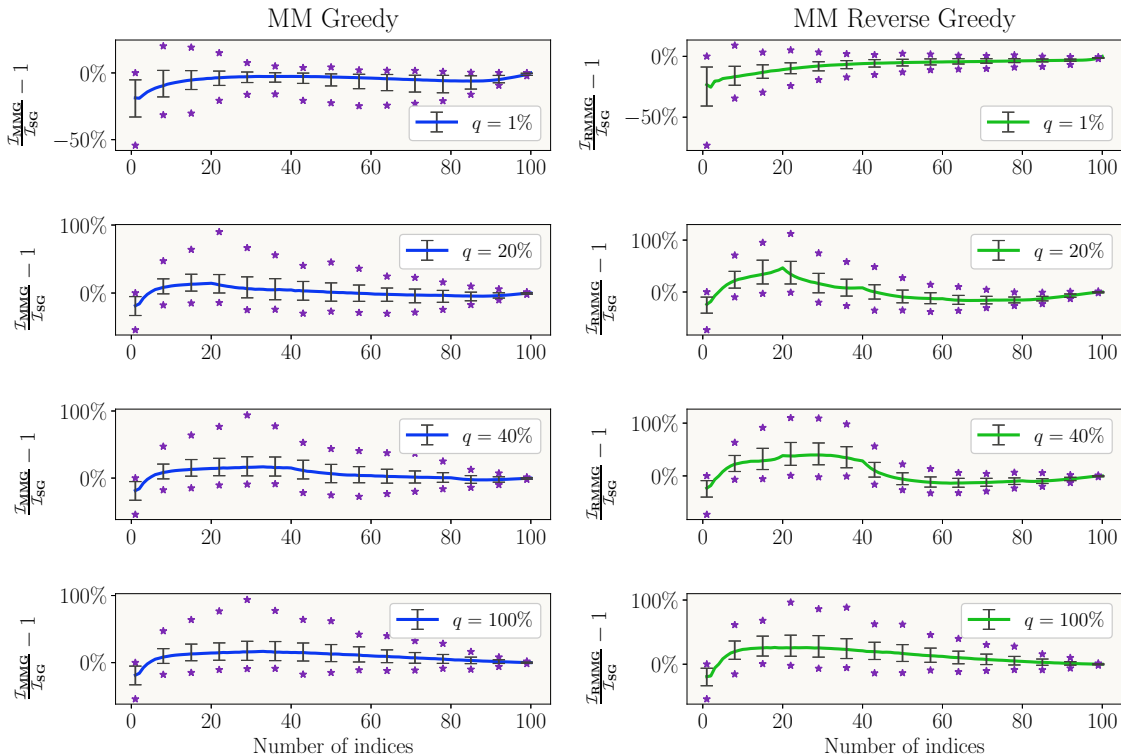
Figure 7: MM-based batch greedy approaches compared against the standard batch greedy heuristic for different batch sizes, across all 1000 random instances of the forward model. The solid line is the median; the whiskers capture the 10% to 90% interquantile range, and the $\star$ marks the maximum and minimum mutual information captured.

heuristics perform similarly to the case discussed in this section. The performance of the `MMReverseGreedy` heuristic, however, appears insensitive to the batch size.

5.2 Optimal sensor placement to improve climate models

We now consider a problem of optimizing sensor networks for climate models (which we refer to as the "SNCM" problem). Given a desired cardinality, our goal is to select field observation sites that will yield the most informative data for parameter inference. In the present application, the parameters are uncertain inputs to the land-surface component of a climate model, while the field data correspond to certain observable outputs of the same model. Our example is based on the land-surface component of the Energy Exascale Earth System Model (E3SM), an ongoing effort led by the US Department of Energy. E3SM comprises multiple model components, each with its own set of uncertain input parameters. These individual components can be coupled, and together they simulate the earth's atmosphere, ocean, land surface, and sea ice (Hurrell et al., 2013). Uncertainty in any one component can contribute to a large spread in the overall model predictions. This

affects our understanding of severe climate events, their timing, and our ability to cope with the consequences.

The *simplified* E3SM land model (Ricciuto, 2020), henceforth referred to as sELM, is a "land model" derived from E3SM that simulates carbon cycle processes relevant to the earth system in an efficient way. This allows for large regional ensemble simulations that would otherwise be infeasible using the complete land model. The simulation region of our focus is the eastern part of the north American continent located between the latitudes 28.25°N–48.25°N and longitudes 66.25°W–96.25°W. Please see Figures 8 and 9b for a depiction of the region. The simulation resolution is 0.5° in each direction, which corresponds to a grid of 41×61 points. Only 1642 of those grid points fall within the land area, however, and at those locations we have access to the sELM outputs. In the version of sELM we consider, there are 47 input parameters; these parameters have no spatial variability and have the same prior distribution at every location. Drawing realizations of these parameters yields a simulation ensemble with 2000 samples. The code for sELM is publicly available, and more details about the E3SM land models can be found in the works by Lu and Ricciuto (2019); Ricciuto et al. (2018).

To set up the optimal experimental design problem, we focus exclusively on one output of the sELM, the gross primary production (GPP). GPP can be understood to be a proxy for the amount of carbon flux attributable to the natural vegetation at that location. GPP is a function of the sELM input parameters and relevant meteorological quantities such as temperature. In the version of sELM we consider, the GPP is output as monthly averages for thirty years starting from the year 1980. In Figure 8 we plot the monthly GPP averaged across the 2000 samples of the parameter ensemble and across thirty years of output history. The trends in the plot reflect expected seasonal variations, with more activity in the tropical southern regions. If we treat the GPP output at any grid point as a random variable, then its variance is affected by the uncertainty of the sELM input parameters and the temporal variation of meteorological quantities. In Figure 9b we plot the variance of the GPP as output from the sELM.

Isolating the contribution of the model parameter uncertainty to the variance of the GPP output can be accomplished in a number of ways. We could perform a multivariate regression with covariance estimation that accounts for spatial correlation among the GPP output variables (Rothman et al., 2010), or alternatively use factor model approaches that are popular in econometrics (Fan et al., 2008, 2011). However we adopt a more straightforward technique, since our primary goal is to set up a design problem suitable for a comparative study of the batch greedy algorithms we have proposed. We simply estimate the linear relationship between the GPP output, Y, and sELM model parameters, X, using an empirical estimate of the cross-covariance $\Gamma_{Y,X}$. Incorporating this result into the setting of Section 3.6.1 assumes that the parameters have a normal prior and that the error term is independent of the parameters. In Figure 9a we plot the spectrum of the linear forward operator G obtained in such a manner, along with the prior covariance Γ_X. Observe that we have reduced the dimension of the parameters to 10 by retaining only those that have a prior variance larger than $\mathcal{O}(1)$. The number of candidate observations, corresponding to the dimension of Y, is $m = 1642$. The generalized eigenvalues of the matrix pencil $(\Gamma_X, \Gamma_{X|Y})$, also shown in Figure 9a, suggest that the data are only marginally informative about the parameters.
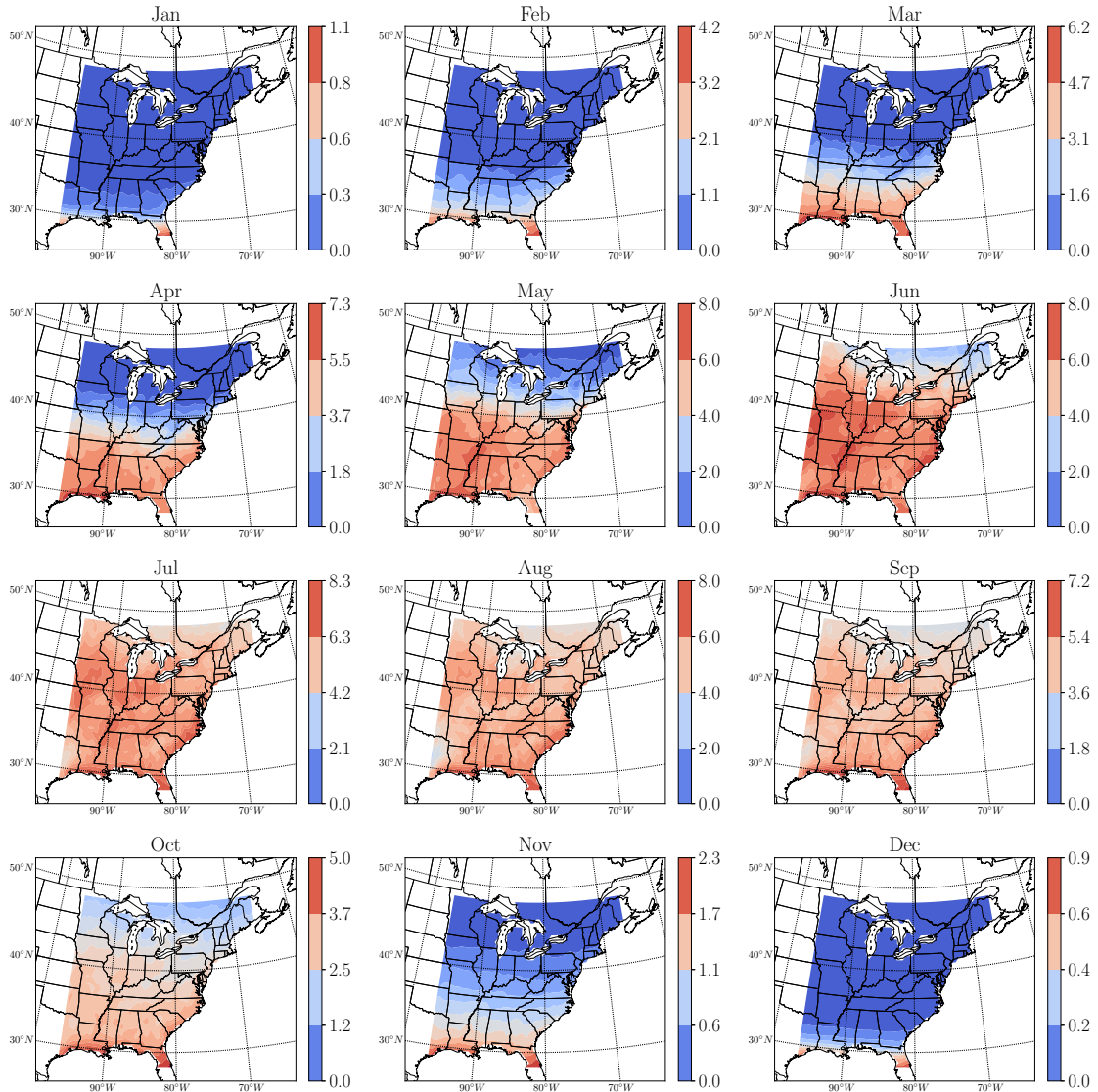
Figure 8: Mean GPP for each month averaged across both the parameter ensemble and temporally.

Using the derived operators, we study the performance of the previously proposed batch greedy algorithms. From Figure 10, it is evident that decreasing the batch size does not reward us with any significant gains; furthermore, all the greedy heuristics have similar performance as indicated in Figure 11. The difference between the relative amount of information captured by the MM greedy heuristics in comparison to the standard greedy heuristic is shown in Figure 12. The MM greedy heuristics have better gains at lower cardinality numbers (except for extremely low cardinality), but these differences return to zero once information saturates. In Figures 10 to 12, we have not shown the case of $q = 1\%$
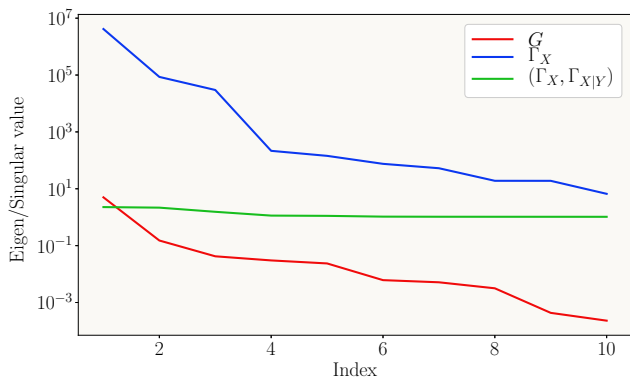
(a) Spectrum of the linearized forward model, Gaussian prior, and the matrix pencil comprising of prior and full posterior.
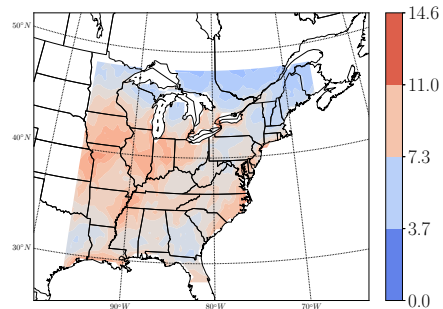
(b) Variance of the GPP marginal visualized on the United States map.

Figure 9: Linear operators of the SNCM problem.

since it requires substantial computing time, but given the trends we expect its performance will not be any better than that of batch size $q = 10\%$.
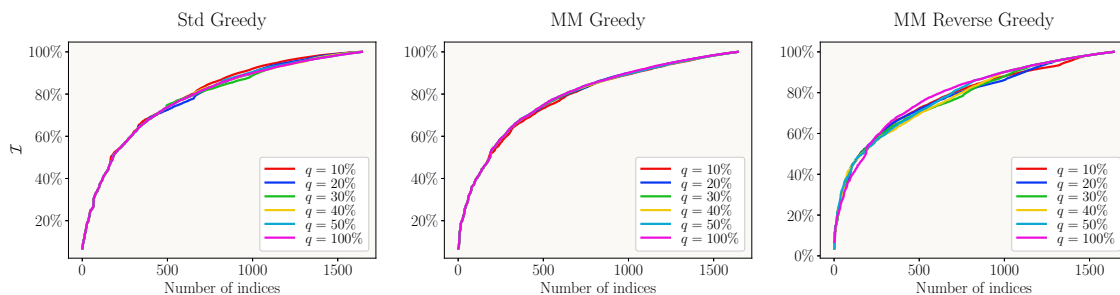

Figure 10: Performance of each greedy heuristic for different batch sizes for the SNCM problem. The batch sizes range from $q = 10\%$ to the one-shot approach with $q = 100\%$.

In Figure 13 we have marked the first 10 locations given by each of the heuristics for the case when the batch size is 10% of the total cardinality. The fact that most of these locations are close to the coastal boundaries—and predominantly the southern coast—is intriguing. This phenomenon is simply a result of the sELM data. To comprehend it better, we have visualized in Figure 14 the rankings of locations at the start of the standard greedy and MM greedy heuristics (i.e., the "one-shot" rankings). These rankings follow from either the incremental gain associated with each location (StdGreedy) or the initial evaluation of the modular lower bound (MMGreedy). While each heuristic provides a different set of locations as the solution of the design problem, the collective information gained from them may not differ significantly. Such a behavior is not unique to the SNCM problem, and is common in many scenarios where combinatorial choices have to be made.
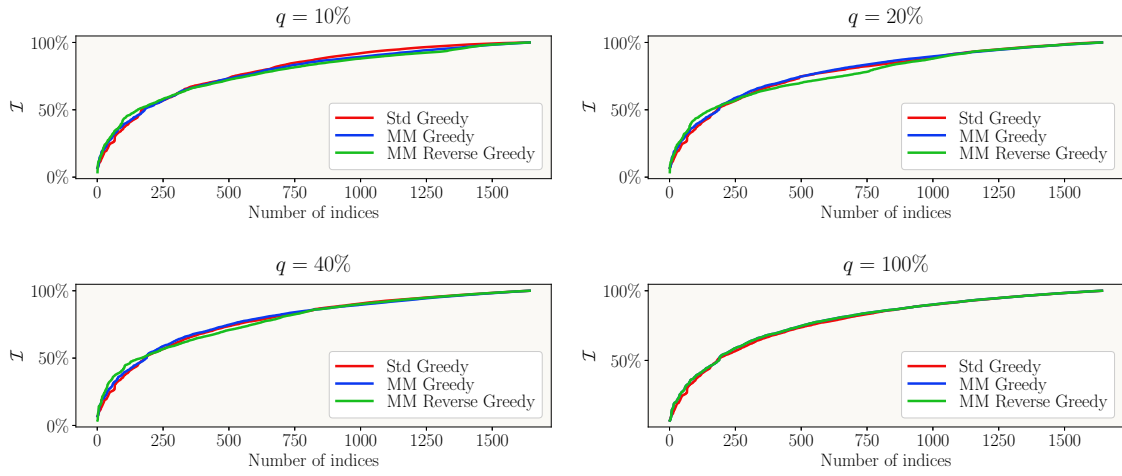
Figure 11: A comparative study of the greedy heuristics for different batch sizes for the SNCM problem.
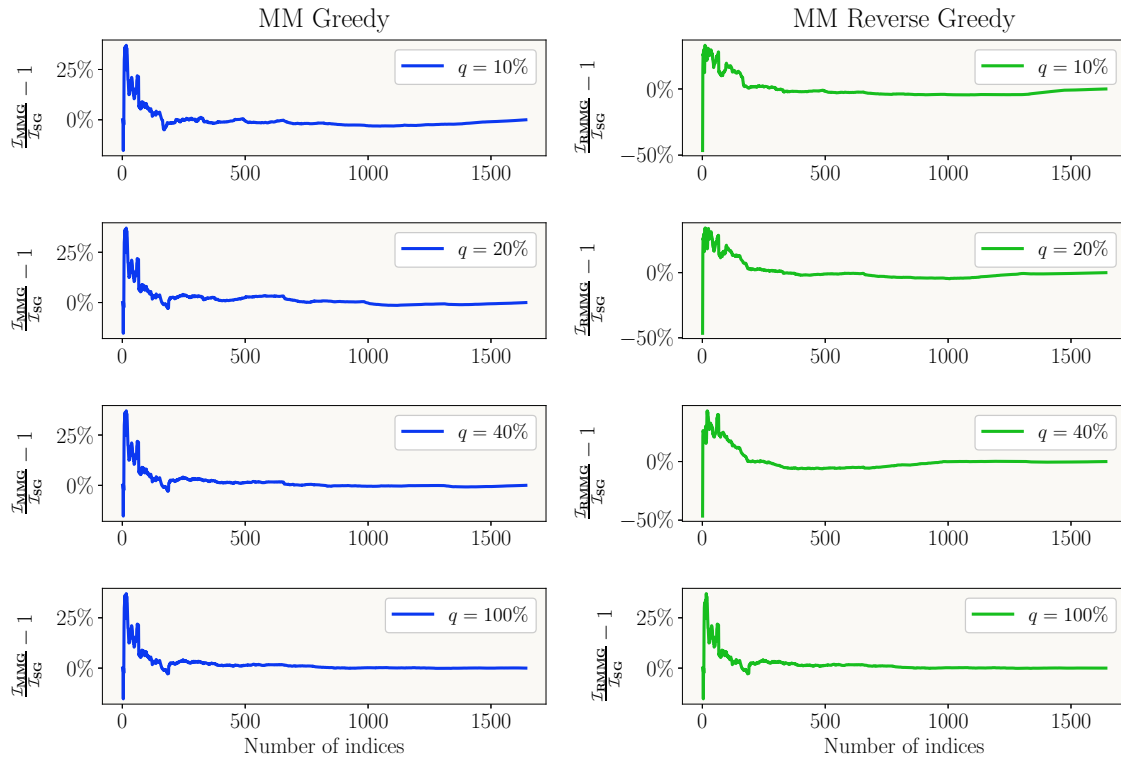

Figure 12: MM-based batch greedy approaches compared against the standard batch greedy heuristic for different batch sizes for the SNCM problem.
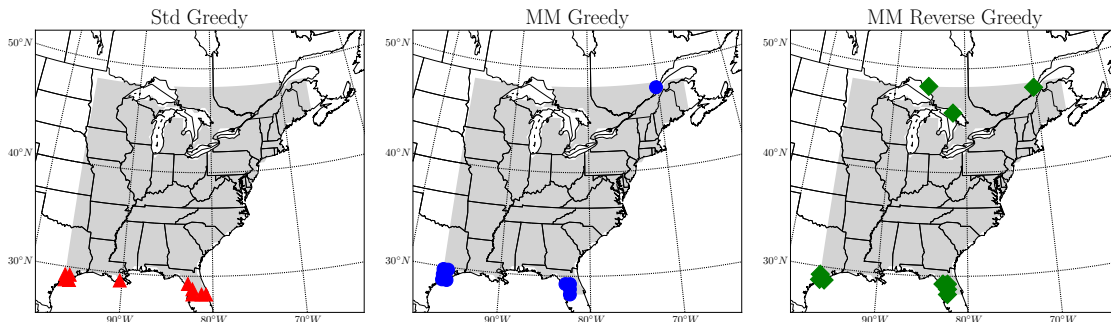
Figure 13: The first ten locations obtained by each greedy heuristic when the batch size q is set to 10% of the full cardinality.
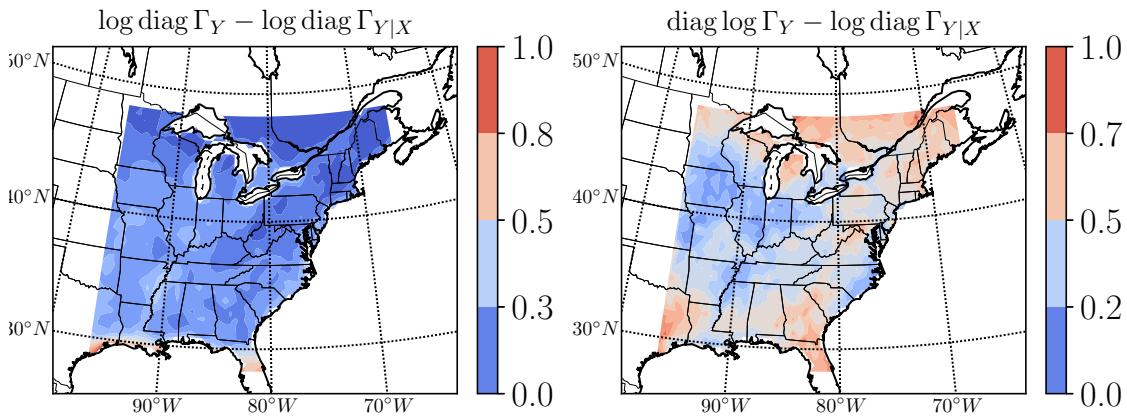

Figure 14: (i) Incremental gains of each candidate location, $\log\operatorname{diag}\Gamma_Y - \log\operatorname{diag}\Gamma_{Y|X}$, scaled to the interval $[0,1]$. (ii) The modular lower bound associated with the one-shot MM greedy approach, $\operatorname{diag}\log\Gamma_Y - \log\operatorname{diag}\Gamma_{Y|X}$, scaled to the interval $[0,1]$.

6. Discussion

This paper has investigated batch greedy heuristics for maximizing monotone non-submodular functions under cardinality constraints. We analyzed batch versions of the standard greedy paradigm, and of its distributed and stochastic variants. Our theoretical guarantees for the resulting batch greedy algorithms are characterized by a combination of submodularity and supermodularity ratios. In the context of linear Bayesian optimal experimental design, we bound these parameters for the mutual information design criterion. Reinterpreting the classical greedy heuristic in the language of MM algorithms, we also argue how any good modular bound—not necessarily one based on incremental gains—can instantiate a related

greedy technique. Based on those insights, we propose several novel modular bounds and algorithms for optimizing information theoretic design criteria in the context of Bayesian experimental design. Now we discuss some further context and potential extensions.

Our result as expressed in (3) does not incorporate curvature of the function. The classical notion of curvature (Conforti and Cornuéjols, 1984) measures how close a submodular set function is to being modular, while the notion of generalized curvature (Bian et al., 2017) measures how close a set function is to being supermodular. Bian et al. (2017) showed the submodularity ratio and the generalized curvature together quantify how close a set function is to being modular. These parameters provide approximation guarantees that refine the worst case bounds depending on the instance of the function. In our case, the product of supermodularity and submodularity ratios characterizes the modularity of a function, but as it appears in Theorem 7, this product does not always refine the worst case bound as tightly as desired. Incorporating curvature into our arguments should result in a more expressive approximation guarantee, and the technical path could be similar to that of Conforti and Cornuéjols (1984); Bian et al. (2017).

The batch size q implicitly appears in the result (3) through the supermodularity ratio. It can be understood to be either the uniform batch size or more generally the maximum batch size. One's computational budget should ultimately dictate the batch size; we know that a smaller maximum batch size will tend to yield better performance. An adaptive strategy to change the batch size across different steps could thus aid the efficient utilization of computational resources. Such a strategy will not influence the worst case bound, but in many practical problems should improve empirical performance. Several factors should impact any adaptive strategy: the number of indices already selected, the number that remains to be selected, and the contrasts between the incremental gains corresponding to each of the remaining indices. It would make sense to measure the latter contrasts relative to the function value evaluated on the set of already chosen indices.

The supermodularity ratio that is essential to characterize the batch greedy heuristic can be a useful theoretical construct in several other settings. Il'ev (2001) analyzed the greedy descent algorithm for minimizing any non-increasing supermodular function. The approximation guarantee was given using the *steepness* of the function, which is a counterpart to curvature for supermodular functions. Using the supermodularity ratio, it should be possible to analyze the case of minimizing any non-increasing set function. Such an effort would mirror the work of Bian et al. (2017) but in the context of minimizing non-increasing functions.

The modular bounds we derived for the information theoretic objectives in Section 4.2 have intriguing connections to concepts in polyhedral combinatorics. When the design objective is a submodular function, the subgradient vector that defines the modular bound is in general a non-extremal point in the base polytope associated with the function. Typically, algorithms rely on enumerating coordinates of the vertices of the polytope and iterate towards a locally optimal solution (Iyer et al., 2012, 2013); our approach is thus somewhat unconventional. The existence of such a subgradient was established using the operator concave inequality (Theorem 43), and was possible since our objective involves an operator monotone function, $\log(\cdot)$, acting on a Hermitian operator (covariance of the data marginal). Hermitian operators that characterize the volume/diversity of subsets arise in other situations too—for instance, as kernels of determinantal point processes. In all such cases it is

unclear if more nuanced and fundamental links exist between classical results in functional analysis and notions in combinatorial optimization. It would be interesting to understand these connections more fully, building on what we have been able to show and exploit.

Dedication

Dedicated to Thara Jagalur, a magnificent spirit who brought tremendous strength to her parents and showed them the transcendental ways of love. She is a shining star and will reverberate in the hearts of all those that knew her.

Acknowledgments

Appendix A. Additional background and definitions

A.1 Definite matrices and generalized eigenvalue problems

Definition 17 (Löwner ordering or the positive semi-definite ordering) *For any two Hermitian matrices A and B, we write $A \succeq B$ if and only if $A - B$ is positive semi-definite. The positive semi-definite condition can be used to define a partial ordering on all Hermitian matrices.*

Definition 18 (Matrix pencil) *(Stewart and Sun, 1990) Given matrices A and B, a matrix pencil is a family of matrices $A - \lambda B$, parametrized by a complex number λ.*

Definition 19 (Definite pencil and definite generalized eigenvalue problem) *(Stewart and Sun, 1990) The pair of Hermitian matrices (A, B) is a definite pencil if*

$$c(A, B) = \inf_{\|x\|=1} \{x^*(A + \sqrt{-1}B)x\} > 0,$$

where $c(A, B)$ is called the Crawford number (Crawford, 1976) of the definite pencil $A - \lambda B$. The generalized eigenvalue problem $A\Upsilon = B\Upsilon\Sigma$ is definite if (A, B) is definite.

Throughout the manuscript we refer to the generalized eigenvalue problem involving matrices A and B through the pair (A, B). We use σ_k to denote any non-trivial generalized eigenvalue of (A, B) and Σ to denote the diagonal matrix of eigenvalues.

A.2 Notions from statistical information theory

Definition 20 (Differential entropy) *(Cover and Thomas, 2006) The entropy of a continuous random variable X with density π_X is defined as*

$$\mathcal{H}(X) := -\mathbb{E}_{\pi_X} \log \pi_X(x).$$

Definition 21 (Condtional differential entropy) *(Cover and Thomas, 2006) For two continuous random variables X and Y, let $\pi_{X,Y}$ denote their joint density and let $\pi_{X|Y}$ denote the density of the conditional distribution of $X|Y$. Then the conditional entropy $\mathcal{H}(X|Y)$ is defined as*

$$\mathcal{H}(X|Y) := -\mathbb{E}_{\pi_{X,Y}} \log \pi_{X|Y}(x|y).$$

Definition 22 (Relative entropy / Kullback–Leibler divergence) *(Cover and Thomas, 2006) The relative entropy or Kullback–Leibler divergence between two distributions with densities π and $\widehat{\pi}$ is defined as*

$$D_{KL}(\pi\|\widehat{\pi}) := \mathbb{E}_\pi \log \frac{\pi(x)}{\widehat{\pi}(x)}.$$

A.3 Submodularity and related concepts

Definition 23 (Modular set function) *(Fujishige, 2005; Bach, 2013) A set function $F: 2^{\mathcal{V}} \to \mathbb{R}$ is modular (i.e., both submodular and supermodular) if and only if there exists $s \in \mathbb{R}^{|\mathcal{V}|}$ such that $F(\mathscr{A}) = \sum_{k \in \mathscr{A}} s_k$.*

In the literature concerning submodular functions, it is common to refer to any vector $s \in \mathbb{R}^{|\mathcal{V}|}$ as the modular set function defined as $s(\mathscr{A}) = \sum_{k \in \mathscr{A}} s_k$. This practice is particularly useful when discussing submodular and base polyhedra, or subgradients.

Definition 24 (Submodular and base polyhedra) *(Fujishige, 2005; Bach, 2013) Let F be a submodular function such that $F(\emptyset) = 0$. The submodular polyhedron $P(F)$ and the base polyhedron $B(F)$ are defined as:*

$$P(F) := \{s \in \mathbb{R}^m : \forall \mathscr{A} \subset \mathcal{V}, s(\mathscr{A}) \le F(\mathscr{A})\}, \text{ and}$$
$$B(F) := \{s \in \mathbb{R}^m : s(\mathcal{V}) = F(\mathcal{V}), \forall \mathscr{A} \subset \mathcal{V}, s(\mathscr{A}) \le F(\mathscr{A})\},$$
$$= P(F) \cap \{s(\mathcal{V}) = F(\mathcal{V})\}.$$

Remark 25 *Analogous to the submodular polyhedron, one can also define the supermodular polyhedron for supermodular functions with the inequalities in Definition 24 being accordingly reversed.*

Definition 26 (Generalized lower and upper polyhedra) *(Iyer and Bilmes, 2015) Let F be any set function, not necessarily sub- or super-modular. The generalized lower polyhedron is defined as:*

$$P_l^{gen}(F) := \{(s, \mathfrak{S}) : s \in \mathbb{R}^m, \mathfrak{S} \in \mathbb{R}, \forall \mathscr{A} \subset \mathcal{V}, s(\mathscr{A}) + \mathfrak{S} \le F(\mathscr{A})\}$$

By reversing the inequality above, we can define the generalized upper polyhedron as:

$$P_u^{gen}(F) := \{(s, \mathfrak{S}) : s \in \mathbb{R}^m, \mathfrak{S} \in \mathbb{R}, \forall \mathscr{A} \subset \mathcal{V}, s(\mathscr{A}) + \mathfrak{S} \ge F(\mathscr{A})\}$$

Definition 27 (Subgradients and subdifferentials of submodular functions) *(Fujishige, 2005) Consider a submodular function $F : \mathscr{D} \to \mathbb{R}$ on a distributive lattice $\mathscr{D} \subseteq 2^{\mathscr{V}}$, with $\emptyset, \mathscr{V} \in \mathscr{D}$. For $\alpha \in \mathbb{R}^{\mathscr{V}}$ and $\mathscr{A} \in \mathscr{D}$, if*

$$\alpha(\mathscr{B}) - \alpha(\mathscr{A}) \leq F(\mathscr{B}) - F(\mathscr{A})$$

holds for each $\mathscr{B} \in \mathscr{D}$, then we call α a subgradient of F at $\mathscr{A}$. We denote by $\partial_F(\mathscr{A})$ the set of all the subgradients of F at $\mathscr{A}$ and call $\partial_F(\mathscr{A})$ the subdifferential of F at $\mathscr{A}$.

Definition 28 (Supergradients and superdifferentials of submodular functions) *(Jegelka and Bilmes, 2011) Consider a submodular function $F : \mathscr{D} \to \mathbb{R}$ on a distributive lattice $\mathscr{D} \subseteq 2^{\mathscr{V}}$, with $\emptyset, \mathscr{V} \in \mathscr{D}$. For $\alpha \in \mathbb{R}^{\mathscr{V}}$ and $\mathscr{A} \in \mathscr{D}$, if*

$$\alpha(\mathscr{B}) - \alpha(\mathscr{A}) \geq F(\mathscr{B}) - F(\mathscr{A})$$

holds for each $\mathscr{B} \in \mathscr{D}$, then we call α a supergradient of F at $\mathscr{A}$. We denote by $\partial^F(\mathscr{A})$ the set of all the supergradients of F at $\mathscr{A}$ and call $\partial^F(\mathscr{A})$ the superdifferential of F at $\mathscr{A}$.

Definition 29 (Supermodular dual) *(Fujishige, 2005) For any submodular function $F : 2^{\mathscr{V}} \to \mathbb{R}$, the function $F^{\#}(\mathscr{X}) := F(\mathscr{V}) - F(\mathscr{V} \setminus \mathscr{X})$, $\mathscr{X} \subseteq \mathscr{V}$ is referred to as its supermodular dual, with the properties $(F^{\#})^{\#} = F$, and $B(F) = B(F^{\#})$. Here B is the base polytope associated with each function (Definition 24).*

Definition 30 (Total curvature of a non-decreasing submodular function) *(Conforti and Cornuéjols, 1984; Sviridenko et al., 2017) The total curvature of a non-decreasing submodular function is the scalar $c \in [0, 1]$ s.t.,*

$$c := \max_{\nu \in \mathscr{V}} \frac{\rho_\nu(\emptyset) - \rho_\nu(\mathscr{V} \setminus \{\nu\})}{\rho_\nu(\emptyset)} = 1 - \min_{\nu \in \mathscr{V}} \frac{\rho_\nu(\mathscr{V} \setminus \{\nu\})}{\rho_\nu(\emptyset)}.$$

Definition 31 (Generalized curvature) *(Bian et al., 2017) The generalized curvature of a non-negative set function F is the smallest scalar α s.t.,*

$$\rho_\nu(\mathscr{A} \setminus \{\nu\} \cup \mathscr{B}) \geq (1 - \alpha)\rho_\nu(\mathscr{A} \setminus \{\nu\}), \qquad \forall \mathscr{A}, \mathscr{B} \subseteq \mathscr{V}, \nu \in \mathscr{A} \setminus \mathscr{B}.$$

Definition 32 (Submodularity ratio as per Das and Kempe (2011)) *The submodularity ratio of a non-negative set function with respect to a set $\mathscr{V}$ and a parameter $k \geq 1$ is*

$$\gamma_{\mathscr{V},k}(F) := \min_{\mathscr{B} \subseteq \mathscr{V}, \mathscr{A}:|\mathscr{A}| \leq k, \mathscr{A} \cap \mathscr{B} = \emptyset} \frac{\sum_{\nu \in \mathscr{A}} \rho_\nu(\mathscr{B})}{\rho_{\mathscr{A}}(\mathscr{B})}$$

Definition 33 (Submodularity ratio as per Bian et al. (2017)) *The submodularity ratio of a non-negative set function F is the largest scalar γ s.t.,*

$$\sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} \rho_\nu(\mathscr{B}) \geq \gamma \rho_{\mathscr{A}}(\mathscr{B}), \qquad \forall \mathscr{A}, \mathscr{B} \subseteq \mathscr{V}.$$

Appendix B. Technical results

Proof of Proposition 4 If $Y_{i_j}|X$ are independent then the conditional entropy of $\mathcal{P}^\top Y|X$ can be written as $\mathcal{H}(\mathcal{P}^\top Y|X) = \sum_{j=1}^k \mathcal{H}(Y_{i_j}|X)$. This allows us to write the mutual information $\mathcal{I}(X; \mathcal{P}^\top Y)$ as follows:

$$\mathcal{I}(X; \mathcal{P}^\top Y) = \mathcal{H}(\mathcal{P}^\top Y) - \mathcal{H}(\mathcal{P}^\top Y|X) = \mathcal{H}(\mathcal{P}^\top Y) - \sum_{j=1}^k \mathcal{H}(Y_{i_j}|X). \qquad (26)$$

Consider selection operators $\mathcal{P}_1 \in \mathbb{R}^{m \times k_1}$ and $\mathcal{P}_2 \in \mathbb{R}^{m \times k_2}$ such that $\mathscr{I}(\mathcal{P}_1) \subseteq \mathscr{I}(\mathcal{P}_2)$. Consider a canonical row e_j that selects Y_{i_j}, with $\mathrm{supp}(e_j) \notin \mathscr{I}(\mathcal{P}_2)$. We define new selection operators $\widehat{\mathcal{P}}_1 := [\mathcal{P}_1, e_j] \in \mathbb{R}^{m \times (k_1+1)}$ and $\widehat{\mathcal{P}}_2 := [\mathcal{P}_2, e_j] \in \mathbb{R}^{m \times (k_2+1)}$ by appending e_j to $\mathcal{P}_1$ and $\mathcal{P}_2$ respectively. The incremental change in mutual information from incorporating Y_{i_j} can be determined for each case. Let us consider $\mathcal{I}(X; \widehat{\mathcal{P}}_2^\top Y) - \mathcal{I}(X; \mathcal{P}_2^\top Y)$; the expression can be adapted for $\mathcal{I}(X; \widehat{\mathcal{P}}_1^\top Y) - \mathcal{I}(X; \mathcal{P}_1^\top Y)$ accordingly.

$$\mathcal{I}(X; \widehat{\mathcal{P}}_2^\top Y) - \mathcal{I}(X; \mathcal{P}_2^\top Y) = \mathcal{H}(\widehat{\mathcal{P}}_2^\top Y) - \mathcal{H}(\mathcal{P}_2^\top Y) - \mathcal{H}(Y_{i_j}|X), \qquad (27a)$$

$$= \mathcal{H}(Y_{i_j}|\mathcal{P}_2^\top Y) - \mathcal{H}(Y_{i_j}|X). \qquad (27b)$$

To obtain (27b) we recognize from the chain rule for entropy that $\mathcal{H}(Y_{i_j}|\mathcal{P}_2^\top Y) = \mathcal{H}(\widehat{\mathcal{P}}_2^\top Y) - \mathcal{H}(\mathcal{P}_2^\top Y)$. Since conditioning cannot increase entropy we can assert,

$$\mathcal{H}(Y_{i_j}|\mathcal{P}_1^\top Y) \geq \mathcal{H}(Y_{i_j}|\mathcal{P}_2^\top Y), \qquad (28)$$

$$\Rightarrow \mathcal{I}(X; \widehat{\mathcal{P}}_1^\top Y) - \mathcal{I}(X; \mathcal{P}_1^\top Y) \geq \mathcal{I}(X; \widehat{\mathcal{P}}_2^\top Y) - \mathcal{I}(X; \mathcal{P}_2^\top Y), \qquad (29)$$

which concludes the proof. ∎

B.1 Proofs concerning batch greedy algorithms

Below we provide a formal proof demonstrating that any function F is supermodular if and only if the supermodularity ratio $\eta = 1$. Recall that the supermodularity ratio without the cardinality parameter is defined as the largest scalar such that

$$\frac{\rho_{\mathscr{A}}(\mathscr{B})}{\sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} \rho_\nu(\mathscr{B})} \geq \eta, \qquad \forall \mathscr{A}, \mathscr{B} \subseteq \mathscr{V}.$$

Proof Assuming $\eta = 1$, we can claim the following inequalities for any set $\mathscr{B} \subseteq \mathscr{V}$, and $\{\nu_1, \nu_2\} \in \mathscr{V} \setminus \mathscr{B}$

$$F(\{\nu_1, \nu_2\} \cup \mathscr{B}) - F(\mathscr{B}) \geq \sum_{i=1}^2 \left(F(\{\nu_i\} \cup \mathscr{B}) - F(\mathscr{B}) \right), \qquad (30)$$

$$\Rightarrow F(\{\nu_1, \nu_2\} \cup \mathscr{B}) - F(\{\nu_1\} \cup \mathscr{B}) \geq F(\{\nu_2\} \cup \mathscr{B}) - F(\mathscr{B}). \qquad (31)$$

This demonstrates supermodularity (see Bach, 2013, Proposition 2.3) having assumed $\eta = 1$. To complete the proof, we prove the proposition statement the other way around by

recursively exploiting the supermodularity property. For any sets $\mathscr{A}, \mathscr{B} \subseteq \mathscr{V}$ with $|\mathscr{A}| \leq k$, $\mathscr{A} \cap \mathscr{B} = \emptyset$ consider $\nu_k \in \mathscr{A}$, we claim because of supermodularity

$$F(\mathscr{A} \cup \mathscr{B}) - F(\mathscr{B}) \geq F(\mathscr{A} \setminus \{\nu_k\} \cup \mathscr{B}) - F(\mathscr{B}) + \underbrace{F(\{\nu_k\} \cup \mathscr{B}) - F(\mathscr{B})}_{\rho_{\nu_k}(\mathscr{B})}. \tag{32}$$

Repeating the same argument but on the term $F(\mathscr{A} \setminus \{\nu_k\} \cup \mathscr{B}) - F(\mathscr{B})$ we have,

$$F(\mathscr{A} \cup \mathscr{B}) - F(\mathscr{B}) \geq F(\mathscr{A} \setminus \{\nu_k, \nu_{k-1}\} \cup \mathscr{B}) - F(\mathscr{B}) + \rho_{\nu_k}(\mathscr{B}) + \rho_{\nu_{k-1}}(\mathscr{B}). \tag{33}$$

Continuing the process gives us the inequality

$$F(\mathscr{A} \cup \mathscr{B}) - F(\mathscr{B}) \geq \sum_{\nu \in \mathscr{A}} \left(F(\{\nu\} \cup \mathscr{B}) - F(\mathscr{B}) \right). \tag{34}$$

∎

B.1.1 STANDARD BATCH GREEDY ALGORITHM

Proof of Theorem 7 Our arguments closely follow those of Nemhauser et al. (1978); Wolsey (1982) with suitable modifications. Let $\mathscr{A}^*$ be a maximizer of F with k elements. Let $\{a_{i_1}, \ldots, a_{i_{q_j}}\}$ be the q_j elements selected during the j–th step of the greedy algorithm. If $\mathscr{A}_{j-1}$ is the set of elements after $j - 1$ steps, then $\mathscr{A}_j = \mathscr{A}_{j-1} \cup \{a_{i_1}, \ldots, a_{i_{q_j}}\}$. If there are l steps altogether, we naturally have $q_1 + q_2 + \cdots + q_l = k$. For a given $j \in \{1, \ldots, l\}$, we denote by $\{b_1, \ldots, b_m\}$ the elements of $\mathscr{A}^* \setminus \mathscr{A}_{j-1}$ (we must have $k \geq m \geq q_j + q_{j+1} + \cdots + q_l$). We then have

$$F(\mathscr{A}^*) \overset{a}{\leq} F(\mathscr{A}^* \cup \mathscr{A}_{j-1}), \tag{35a}$$

$$\overset{b}{\leq} F(\mathscr{A}_{j-1}) + \frac{1}{\gamma_{\mathscr{V},m}} \sum_{i=1}^{m} \left(F(\mathscr{A}_{j-1} \cup \{b_i\}) - F(\mathscr{A}_{j-1}) \right) \tag{35b}$$

$$\overset{c}{\leq} F(\mathscr{A}_{j-1}) + \frac{m}{q_j \gamma_{\mathscr{V},m}} \sum_{t=1}^{q_j} \left(F(\mathscr{A}_{j-1} \cup \{a_{i_t}\}) - F(\mathscr{A}_{j-1}) \right) \tag{35c}$$

$$\overset{d}{\leq} F(\mathscr{A}_{j-1}) + \frac{m}{q_j \eta_{\mathscr{V},q_j} \gamma_{\mathscr{V},m}} \left(F(\mathscr{A}_j) - F(\mathscr{A}_{j-1}) \right) \tag{35d}$$

$$\overset{e}{\leq} F(\mathscr{A}_{j-1}) + \frac{k}{q_j \eta_{\mathscr{V},q_j} \gamma_{\mathscr{V},k}} \left(F(\mathscr{A}_j) - F(\mathscr{A}_{j-1}) \right). \tag{35e}$$

a) holds because F is non-decreasing,

b) holds by definition of the submodularity ratio (Definition 32),

c) holds because $q_j \leq m$, and the q_j elements in the set $\{a_{i_1}, \ldots, a_{i_{q_j}}\}$ are (by virtue of the batch greedy algorithm) those that correspond to the highest incremental gains at step j:

$$\frac{\sum_{i=1}^{m} \left(F(\mathscr{A}_{j-1} \cup \{b_i\}) - F(\mathscr{A}_{j-1}) \right)}{m} \leq \frac{\sum_{t=1}^{q_j} \left(F(\mathscr{A}_{j-1} \cup \{a_{i_t}\}) - F(\mathscr{A}_{j-1}) \right)}{q_j}$$

d) holds by definition of $\eta_{\mathscr{V},q_j}$ as the supermodularity ratio (Definition 6),

e) holds because $m \leq k$, and $\gamma_{\mathscr{V},m} \geq \gamma_{\mathscr{V},k}$.

A simple manipulation of (35e) allows us to write the following,

$$F(\mathscr{A}^*) - F(\mathscr{A}_j) \leq \left(1 - \frac{q_j \eta_{\mathscr{V},q_j} \gamma_{\mathscr{V},k}}{k}\right)\left(F(\mathscr{A}^*) - F(\mathscr{A}_{j-1})\right). \tag{36}$$

This leads to $F(\mathscr{A}^*) - F(\mathscr{A}_l) \leq \prod_{j=1}^{l}\left(1 - \frac{q_j \eta_{\mathscr{V},q_j} \gamma_{\mathscr{V},k}}{k}\right) F(\mathscr{A}^*)$, and the theorem statement follows. ∎

B.1.2 DISTRIBUTED BATCH GREEDY ALGORITHM

Our approach to prove Theorem 8 adapts the strategies of Mirzasoleiman et al. (2013) to the batch and nonsubmodular setting. Similar to the exposition there, we first investigate some approximation properties of the intractable but communication efficient counterpart to Algorithm 2. Here one first distributes the ground set $\mathscr{V}$ to n_p processes. Each process then finds the optimal solution $A^{\mathrm{opt}}_{i,[n_p,k]}$, i.e., a set of cardinality at most k, that maximizes the value of F in each partition. These solutions are then merged $\bigcup_i A^{\mathrm{opt}}_{i,[n_p,k]}$, and the optimal subset $A^{\mathrm{d\text{-}opt}}_{[n_p,k]}$ of cardinality k is found in the combined set. The following lemma relates $A^{\mathrm{d\text{-}opt}}_{[n_p,k]}$ to the combinatorial optimal solution $A^{\mathrm{opt}}_{[k]} = \arg\max_{\mathscr{B}\subset\mathscr{V}, |\mathscr{B}|\leq k} F(\mathscr{B})$.

Lemma 34 *Let F be a non-decreasing function with $F(\emptyset) = 0$, and let $k > 0$. Then*

$$F(A^{d\text{-}opt}_{[n_p,k]}) \geq \frac{\gamma_{\mathscr{V},k}}{k} F(A^{opt}_{[k]})$$

Proof of Lemma 34 Let $A^{\mathrm{opt}}_{[k]} = \{\nu_1, \ldots, \nu_k\}$. Using the notion of submodularity ratio, we have $\sum_{\nu \in A^{\mathrm{opt}}_{[k]}} \rho_\nu(\emptyset) \geq \gamma_{\mathscr{V},k} \rho_{A^{\mathrm{opt}}_{[k]}}(\emptyset)$, $\Rightarrow F(A^{\mathrm{opt}}_{[k]}) \leq \frac{1}{\gamma_{\mathscr{V},k}} \sum_{\nu \in A^{\mathrm{opt}}_{[k]}} F(\nu)$. If $\nu^* = \arg\max_{\nu \in A^{\mathrm{opt}}_{[k]}} F(\nu)$, then $F(A^{\mathrm{opt}}_{[k]}) \leq \frac{k}{\gamma_{\mathscr{V},k}} F(\nu^*)$. Suppose $\nu^* \in \mathscr{V}_j$, the j–th partition of the ground set; then we know that $F(\nu^*) \leq F(A^{\mathrm{opt}}_{j,[k]})$. The lemma statement now follows since $F(A^{\mathrm{opt}}_{j,[n_p,k]}) \leq \max_j F(A^{\mathrm{opt}}_{j,[n_p,k]}) \leq F(A^{\mathrm{d\text{-}opt}}_{[n_p,k]})$. ∎

Apart from Lemma 34 we also need a slight generalization of Theorem 7 and (3) to prove Theorem 8. The result below is straightforward to derive and we omit the proof.

Lemma 35 *Let F be a non-decreasing function with $F(\emptyset) = 0$, and let $\mathscr{A}^{bg}_{[\widehat{k}]}$ be the index set of cardinality $\widehat{k} = q\widehat{l}$ returned by the batch greedy algorithm maximizing $F(\mathscr{A})$ subject to $|\mathscr{A}| \leq k = ql$. Then,*

$$F(\mathscr{A}^{bg}_{[\widehat{k}]}) \geq \left(1 - e^{-\frac{\widehat{l}}{l}\eta_{\mathscr{V},q}\gamma_{\mathscr{V},k}}\right) F(A^{opt}_{[k]}).$$

Here $\gamma_{\mathscr{V},k}$ is the submodularity ratio and $\eta_{\mathscr{V},q}$ is the supermodularity ratio.

Proof of Theorem 8 Let $\widehat{\mathscr{A}} = \arg\max_{\mathscr{B}\subset\mathscr{M},|\mathscr{B}|\leq\widehat{k}} F(\mathscr{B})$, where $\mathscr{M} := \bigcup_i \mathscr{A}^{\mathrm{bg}}_{i,[n_p,\widehat{k}]}$. We know that $F(\widehat{\mathscr{A}}) \geq \max_i F(\mathscr{A}^{\mathrm{bg}}_{i,[n_p,\widehat{k}]})$. From Lemma 35, we have

$$F(\mathscr{A}^{\mathrm{bg}}_{i,[n_p,\widehat{k}]}) \geq \left(1 - e^{-\eta_{\mathscr{V},q}\gamma_{\mathscr{V},k}\widehat{l}/l}\right) F(A^{\mathrm{opt}}_{i,[n_p,k]}), \tag{37}$$

$$\Rightarrow F(\widehat{\mathscr{A}}) \geq \left(1 - e^{-\eta_{\mathscr{V},q}\gamma_{\mathscr{V},k}\widehat{l}/l}\right) \max_i F(A^{\mathrm{opt}}_{i,[n_p,k]}). \tag{38}$$

From Lemma 34, we know that $\max_i F(A^{\mathrm{opt}}_{i,[n_p,k]}) \geq \frac{\gamma_{\mathscr{V},k}}{k} F(A^{\mathrm{opt}}_{[k]})$; therefore

$$F(\widehat{\mathscr{A}}) \geq \left(1 - e^{-\eta_{\mathscr{V},q}\gamma_{\mathscr{V},k}\widehat{l}/l}\right) \frac{\gamma_{\mathscr{V},k}}{k} F(A^{\mathrm{opt}}_{[k]}). \tag{39}$$

Using Lemma 35 we can relate $F(\widehat{\mathscr{A}})$ and $F(\mathscr{A}^{\mathrm{d\text{-}bg}})$ as follows:

$$F(\mathscr{A}^{\mathrm{d\text{-}bg}}_{[n_p,\widetilde{k}]}) \geq \left(1 - e^{-\eta_{\mathscr{V},q}\gamma_{\mathscr{V},\widetilde{k}}\widetilde{l}/l}\right) F(\widehat{\mathscr{A}}). \tag{40}$$

Combining (39) and (40), we obtain the theorem statement. ∎

B.1.3 STOCHASTIC BATCH GREEDY ALGORITHM

We first provide two lemmas that will prove useful in establishing Theorem 9. The first lemma is a bound on the probability of a certain event involving the random set drawn at each step, and the second lemma is a bound on the expected incremental gain at each step.

Lemma 36 *Given solution $\mathscr{A}_i$ after i–steps of the* `StochasticBatchGreedy` *algorithm, if $k \leq \lfloor\sqrt{m/e} - 1/e\rfloor$ and $\frac{m-2k}{2ek^2} \geq \frac{q-1}{\log^2\frac{q}{\epsilon}}$, then for any set $\mathscr{R}$ of size $\frac{m}{k}\log\frac{q}{\epsilon}$ sampled uniformly at random from $\mathscr{V}\setminus\mathscr{A}_i$, we have $\mathbb{P}(|\mathscr{R}\cap\mathscr{A}^*\setminus\mathscr{A}_i| \geq q) \geq (1-\epsilon)\frac{|\mathscr{A}^*\setminus\mathscr{A}_i|}{k}$.*

Proof of Lemma 36 We assume that all random quantities are conditioned on A_i. The set $\mathscr{R}$ consists of s elements from $\mathscr{V}\setminus\mathscr{A}_i$ sampled uniformly at random (w.l.o.g. with repetition). Let $\widetilde{k} := |\mathscr{A}^*\setminus\mathscr{A}_i|$, $\widetilde{m} := |\mathscr{V}\setminus\mathscr{A}_i|$, and $p := \frac{|\mathscr{A}^*\setminus\mathscr{A}_i|}{|\mathscr{V}\setminus\mathscr{A}_i|} = \frac{\widetilde{k}}{\widetilde{m}}$ be the probability representing the event that any element in the random set $\mathscr{R}$ is from $\mathscr{A}^*\setminus\mathscr{A}_i$. We now estimate the probability of $|\mathscr{R}\cap\mathscr{A}^*\setminus\mathscr{A}_i| \geq q$ by first deriving an upper bound on the complementary

event $|\mathscr{R} \cap \mathscr{A}^* \setminus \mathscr{A}_i| \leq q - 1$.

$$\mathbb{P}(|\mathscr{R} \cap \mathscr{A}^* \setminus \mathscr{A}_i| \leq q - 1) \overset{a}{=} \sum_{i=0}^{q-1} \binom{\widetilde{k}}{i}(1-p)^{s-i}p^i, \tag{41a}$$

$$\overset{b}{\leq} \sum_{i=0}^{q-1} \left(\frac{e\widetilde{k}}{i}\right)^i (1-p)^{s-i}p^i, \tag{41b}$$

$$\overset{c}{\leq} (1-p)^s \sum_{i=0}^{q-1} \left(\frac{e\widetilde{k}p}{1-p}\right)^i, \tag{41c}$$

$$\overset{d}{\leq} (1-p)^s \left(1 + (q-1)\left(\frac{e\widetilde{k}^2}{m-2k}\right)\right), \tag{41d}$$

$$\overset{e}{\leq} \exp(-s\frac{\widetilde{k}}{m})\psi(\widetilde{k}). \tag{41e}$$

b) holds by upper bounding the binomial coefficient $\binom{\widetilde{k}}{i} \leq \left(\frac{e\widetilde{k}}{i}\right)^i$.

c) holds because $\left(\frac{1}{i}\right)^i \leq 1 \quad \forall i \in \mathbb{Z}_{\geq 0}$.

d) holds because of the additional assumption we impose on the cardinality constraint $k \leq \lfloor \sqrt{m/e} - 1/e \rfloor$. Using the facts that $m - k + q \leq \widetilde{m}$ and $\widetilde{k} \leq k$ we can claim,

$$\frac{e\widetilde{k}p}{1-p} = \frac{e\widetilde{k}^2}{\widetilde{m} - \widetilde{k}} \leq \frac{e\widetilde{k}^2}{m+q-2k} \leq \frac{e\widetilde{k}^2}{m-2k} \leq \frac{ek^2}{m-2k} \leq 1 \Leftarrow k \leq \lfloor \sqrt{m/e} - 1/e \rfloor$$

e) holds because,

$$(1-p)^s = (1 - \widetilde{k}/\widetilde{m})^s \leq \exp(-s\frac{\widetilde{k}}{\widetilde{m}}) \leq \exp(-s\frac{\widetilde{k}}{m}) \Leftarrow \widetilde{m} \leq m$$

We have defined $\psi(\widetilde{k}) \equiv \psi(\widetilde{k}; q, m, k) := 1 + (q-1)\left(\frac{e\widetilde{k}^2}{m-2k}\right)$.

We now derive a lower bound for the probability of $|\mathscr{R} \cap \mathscr{A}^* \setminus \mathscr{A}_i| \geq q$.

$$\mathbb{P}(|\mathscr{R} \cap \mathscr{A}^* \setminus \mathscr{A}_i| \geq q) \overset{a}{=} 1 - \mathbb{P}(|\mathscr{R} \cap \mathscr{A}^* \setminus \mathscr{A}_i| \leq q - 1), \tag{42a}$$

$$\overset{b}{\geq} 1 - \exp(-s\frac{\widetilde{k}}{m})\psi(\widetilde{k}), \tag{42b}$$

$$\overset{c}{\geq} \left(1 - \exp(-s\frac{k}{m})\psi(k)\right)\frac{\widetilde{k}}{k}, \tag{42c}$$

$$\overset{d}{\geq} (1-\epsilon)\frac{\widetilde{k}}{k}. \tag{42d}$$

b) follows from (41e),

c) For $s \geq m\sqrt{\frac{2e(q-1)}{m-2k}}$, the function $1-\exp(-s\frac{x}{m})\psi(x)$ is concave in the interval $x \in [0, k]$.

It can be easily verified that $\frac{\partial^2 \exp(-s\frac{x}{m})\psi(x)}{\partial^2 x} > 0$ if $s \geq m\sqrt{\frac{2e(q-1)}{m-2k}}$.

$$\Rightarrow \frac{1 - \exp(-s\frac{x}{m})\psi(x)}{x} \geq \frac{1 - \exp(-s\frac{k}{m})\psi(k)}{k},$$

$$\Leftrightarrow 1 - \exp(-s\frac{x}{m})\psi(x) \geq \frac{x}{k}\left(1 - \exp(-s\frac{k}{m})\psi(k)\right).$$

d) We choose the set $\mathscr{R}$ s.t. $|\mathscr{R}| =: s = \frac{m}{k}\log\frac{q}{\epsilon}$, where $0 < \epsilon < 1$. If the batch size q and tolerance ϵ are chosen s.t. $\frac{m-2k}{2ek^2} \geq \frac{q-1}{\log^2\frac{q}{\epsilon}}$, then $\frac{m}{k}\log\frac{q}{\epsilon} \geq m\sqrt{\frac{2e(q-1)}{m-2k}}$, ensuring concavity of the desired function in the previous step. For the cardinality constraint regime $k \leq \lfloor\sqrt{m/e}-1/e\rfloor$, this condition on batch size and tolerance is easily satisfied. Since $\psi(\widetilde{k}) \in [1, q]$, we now have

$$1 - \exp(-s\frac{\widetilde{k}}{m})\psi(\widetilde{k}) \geq \frac{\widetilde{k}}{k}\left(1 - \exp(-s\frac{k}{m})\psi(k)\right) \geq \frac{\widetilde{k}}{k}(1 - \epsilon). \tag{43}$$

∎

Lemma 37 *Given solution $\mathscr{A}_i$ after i–steps of the* `StochasticBatchGreedy` *algorithm, if the assumptions of Lemma 36 hold, then the expected gain $\mathbb{E}[F(\mathscr{A}_{i+1}) - F(\mathscr{A}_i)]$ in step $i+1$ is at least $(1 - \epsilon)\frac{\gamma_{\mathscr{V},k}\eta_{\mathscr{V},q}q}{k}\mathbb{E}[\rho_{\mathscr{A}^*\setminus\mathscr{A}_i}(\mathscr{A}_i)]$.*

Proof of Lemma 37 We assume that all random quantities are conditioned on A_i. We bound $F(\mathscr{A}_{i+1}) - F(\mathscr{A}_i)$ from below. Since $\mathscr{Q}_{i+1} := \mathscr{A}_{i+1} \setminus \mathscr{A}_i$ is chosen by the stochastic batch greedy rule from $\mathscr{R}$, and if $|\mathscr{R} \cap \mathscr{A}^* \setminus \mathscr{A}_i| \geq q$, then $\sum_{a\in\mathscr{Q}_{i+1}} \rho_a(\mathscr{A}_i)$ is at least as large as $\sum_{a\in\mathscr{Q}_r} \rho_a(\mathscr{A}_i)$ in expectation. Here the set $\mathscr{Q}_r$ with $|\mathscr{Q}_r| = q$ is chosen uniformly at random $\mathscr{R} \cap \mathscr{A}^* \setminus \mathscr{A}_i$. The set $\mathscr{R}$ is equally likely to contain any each element of $\mathscr{A}^* \setminus \mathscr{A}_i$. Hence a uniform random subset $\mathscr{Q}_r \subseteq \mathscr{R} \cap \mathscr{A}^* \setminus \mathscr{A}_i$ is actually a uniformly random subset of $\mathscr{A}^* \setminus \mathscr{A}_i$. As a result we obtain

$$\mathbb{E}[\sum_{a\in\mathscr{Q}_{i+1}} \rho_a(\mathscr{A}_i)] \geq \mathbb{P}(|\mathscr{R} \cap \mathscr{A}^* \setminus \mathscr{A}_i| \geq q)\frac{\sum_{\mathscr{Q}_r\subseteq\mathscr{A}^*\setminus\mathscr{A}_i}\sum_{a\in\mathscr{Q}_r} \rho_a(\mathscr{A}_i)}{\binom{|\mathscr{A}^*\setminus\mathscr{A}_i|}{q}}, \tag{44a}$$

$$\geq (1 - \epsilon)\frac{|\mathscr{A}^* \setminus \mathscr{A}_i|}{k}\frac{\binom{|\mathscr{A}^*\setminus\mathscr{A}_i|}{q}\frac{q}{|\mathscr{A}^*\setminus\mathscr{A}_i|}\sum_{a\in\mathscr{A}^*\setminus\mathscr{A}_i} \rho_a(\mathscr{A}_i)}{\binom{|\mathscr{A}^*\setminus\mathscr{A}_i|}{q}}, \tag{44b}$$

$$\geq (1 - \epsilon)\frac{q}{k}\sum_{a\in\mathscr{A}^*\setminus\mathscr{A}_i} \rho_a(\mathscr{A}_i). \tag{44c}$$

45

(44b) holds on account of Lemma 36. Using the notion of submodularity and supermodularity ratios we can claim,

$$\mathbb{E}[\rho_{\mathscr{Q}_{i+1}}(\mathscr{A}_i)] \geq \eta_{\mathscr{V},q}\mathbb{E}[\sum_{a \in \mathscr{Q}_{i+1}} \rho_a(\mathscr{A}_i)], \tag{45}$$

$$\sum_{a \in \mathscr{A}^*\backslash\mathscr{A}_i} \rho_a(\mathscr{A}_i) \geq \gamma_{\mathscr{V},k}\rho_{\mathscr{A}^*\backslash\mathscr{A}_i}(\mathscr{A}_i). \tag{46}$$

Combining (44c), (45) and (46) we have

$$\mathbb{E}[\rho_{\mathscr{Q}_{i+1}}(\mathscr{A}_i)] \geq (1-\epsilon)\frac{\gamma_{\mathscr{V},k}\eta_{\mathscr{V},q}q}{k}\rho_{\mathscr{A}^*\backslash\mathscr{A}_i}(\mathscr{A}_i). \tag{47}$$

By taking expectation over all possible realizations of $\mathscr{A}_i$ we obtain the lemma. ∎

Proof of Theorem 9 Let $\mathscr{A}^*$ be a maximizer of F with k elements, i.e., $\mathscr{A}^* := \arg\max_{\mathscr{B}\subset\mathscr{V},|\mathscr{B}|\leq k} F(\mathscr{B})$. By recursively using Lemma 37 for $i = 1,\ldots,l$, where $l = k/q$, and knowing that $F(\emptyset) = 0$, we obtain:

$$\mathbb{E}[F(\mathscr{A}_l)] \geq F(\mathscr{A}^*) - \left(1 - (1-\epsilon)\frac{\gamma_{\mathscr{V},k}\eta_{\mathscr{V},q}}{l}\right)^l (F(\mathscr{A}^*) - F(\emptyset)), \tag{48}$$

$$\geq \left(1 - e^{-(1-\epsilon)\gamma_{\mathscr{V},k}\eta_{\mathscr{V},q}}\right) F(\mathscr{A}^*). \tag{49}$$

∎

B.1.4 MM BATCH GREEDY ALGORITHM

Proof of Theorem 12 The arguments are similar to the proof of Theorem 7. Let $\mathscr{A}^*$ be a maximizer of F with k elements. Let $\{a_{i_1},\ldots,a_{i_{q_j}}\}$ be the q_j elements selected during the j–th step of the greedy algorithm. If $\mathscr{A}_{j-1}$ is the set of elements after $j-1$ steps, then $\mathscr{A}_j = \mathscr{A}_{j-1} \cup \{a_{i_1},\ldots,a_{i_{q_j}}\}$. If there are l–steps all together, we naturally have $q_1 + q_2 + \cdots + q_l = k$. For a given $j \in \{1,\ldots,l\}$, we denote by $\{b_1,\ldots,b_m\}$ the elements of $\mathscr{A}^* \backslash \mathscr{A}_{j-1}$ (we must have $k \geq m \geq q_j + q_{j+1} + \cdots + q_l$). We then have,

$$F(\mathscr{A}^*) \overset{a}{\leq} F(\mathscr{A}^* \cup \mathscr{A}_{j-1}), \tag{50a}$$

$$\overset{b}{=} F(\mathscr{A}_{j-1}) + \rho_{\mathscr{A}^*}(\mathscr{A}_{j-1}), \tag{50b}$$

$$\overset{c}{\leq} F(\mathscr{A}_{j-1}) + \delta_j + \mathcal{M}_\uparrow[\rho_{\mathscr{A}^*}(\mathscr{A}_{j-1})], \tag{50c}$$

$$\overset{d}{\leq} F(\mathscr{A}_{j-1}) + \delta_j + \frac{k}{q_j}\mathcal{M}_\uparrow[\rho_{\mathscr{A}_j}(\mathscr{A}_{j-1})], \tag{50d}$$

$$\overset{e}{\leq} F(\mathscr{A}_{j-1}) + \delta_j + \frac{k}{q_j}\rho_{\mathscr{A}_j}(\mathscr{A}_{j-1}), \tag{50e}$$

$$\overset{f}{\leq} F(\mathscr{A}_{j-1}) + \frac{k(1+\tau_j)}{q_j}\rho_{\mathscr{A}_j}(\mathscr{A}_{j-1}). \tag{50f}$$

a) holds because F is non-decreasing,

b) is an identity since $F(\mathscr{A}^* \cup \mathscr{A}_{j-1}) - F(\mathscr{A}_{j-1}) =: \rho_{\mathscr{A}^*}(\mathscr{A}_{j-1})$ is the incremental gain,

c) holds by defining δ_j as the smallest scalar satisfying,

$$\rho_{\mathscr{A}^*}(\mathscr{A}_{j-1}) \leq \delta_j + \mathcal{M}_\uparrow\left[\rho_{\mathscr{A}^*}(\mathscr{A}_{j-1})\right], \tag{51}$$

d) holds because $q_j \leq m \leq k$, and the q_j elements in the set $\{a : a \in \mathscr{A}_j \setminus \mathscr{A}_{j-1}\}$ maximize $\mathcal{M}_\uparrow[\rho_a(\mathscr{A}_{j-1})]$, and hence

$$\frac{\mathcal{M}_\uparrow\left[\rho_{\mathscr{A}^*}(\mathscr{A}_{j-1})\right]}{m} \leq \frac{\mathcal{M}_\uparrow\left[\rho_{\mathscr{A}_j}(\mathscr{A}_{j-1})\right]}{q_j},$$

e) holds since $\mathcal{M}_\uparrow\left[\rho_{\mathscr{A}_j}(\mathscr{A}_{j-1})\right] \leq \rho_{\mathscr{A}_j}(\mathscr{A}_{j-1})$,

f) holds by defining τ_j as the smallest scalar satisfying

$$\delta_j \leq \frac{\tau_j k \rho_{\mathscr{A}_j}(\mathscr{A}_{j-1})}{q_j}. \tag{52}$$

δ_j quantifies the slackness of the modular lower bound, while τ_j does so relative to the incremental gain at the current step. A simple manipulation of (50f) allows us to write the following:

$$F(\mathscr{A}^*) - F(\mathscr{A}_j) \leq \left(1 - \frac{q_j}{k(1+\tau_j)}\right)(F(\mathscr{A}^*) - F(\mathscr{A}_{j-1})). \tag{53}$$

This leads to $F(\mathscr{A}^*) - F(\mathscr{A}_l) \leq \prod_{j=1}^{l}\left(1 - \frac{q_j}{k(1+\tau_j)}\right)F(\mathscr{A}^*)$, and the theorem statement follows. ∎

B.2 Proofs concerning linear Bayesian experimental design criterion

Proof of Theorem 10 Let (σ_j, U_j) be any generalized eigenvalue-eigenvector pair of the definite pair $(\Gamma_Y - \Gamma_{Y|X}, \Gamma_{Y|X})$. Since $\Gamma_Y - \Gamma_{Y|X} \succeq 0$ and $\Gamma_{Y|X} \succ 0$, we can claim $\sigma_j \geq 0$. It is evident that $(1 + \sigma_j, U_j)$ is the corresponding generalized eigenvalue-eigenvector pair of the definite pair $(\Gamma_Y, \Gamma_{Y|X})$. To establish item 1 and item 2, we need to show that σ_j and $1 + \sigma_j$ are the corresponding generalized eigenvalues of the definite pair $(\Gamma_X - \Gamma_{X|Y}, \Gamma_{X|Y})$ and $(\Gamma_X, \Gamma_{X|Y})$, respectively. Item 1 ⇔ item 2 then holds automatically.

Using the specified linear statistical model (6), we can deduce $\Gamma_Y - \Gamma_{Y|X} = G\Gamma_X G^\top$. For the generalized eigenvalue-eigenvector pair (σ_j, U_j), we have the relationship

$$G\Gamma_X G^\top U_j = \Gamma_{Y|X} U_j \sigma_j. \tag{54}$$

Multiplying (54) by $G^\top \Gamma_{Y|X}^{-1}$, we have

$$G^\top \Gamma_{Y|X}^{-1} G\Gamma_X G^\top U_j = G^\top U_j \sigma_j. \tag{55}$$

Define $\widetilde{V}_j := \frac{1}{\alpha_j} \Gamma_X G^\top U_j$, where α_j is some scaling parameter to obtain the desired orthogonality, and rewrite (55) as

$$G^\top \Gamma_{Y|X}^{-1} G \widetilde{V}_j = \Gamma_X^{-1} \widetilde{V}_j \sigma_j. \tag{56}$$

From (56), we claim $(\sigma_j, \widetilde{V}_j)$ is a generalized eigenvalue-eigenvector pair of $(G^\top \Gamma_{Y|X}^{-1} G, \Gamma_X^{-1})$. Consequently, $(1+\sigma_j, \widetilde{V}_j)$ is the corresponding pair of $(\Gamma_X^{-1} + G^\top \Gamma_{Y|X}^{-1} G, \Gamma_X^{-1}) =: (\Gamma_{X|Y}^{-1}, \Gamma_X^{-1})$. By definition the definite pairs $(\Gamma_{X|Y}^{-1}, \Gamma_X^{-1})$ and $(\Gamma_X, \Gamma_{X|Y})$ have the same spectrum. Further, if $(1 + \sigma_j, \widetilde{V}_j)$ is a generalized eigenvalue-eigenvector pair of $(\Gamma_{X|Y}^{-1}, \Gamma_X^{-1})$, then $(1 + \sigma_j, \Gamma_X^{-1} \widetilde{V}_j)$ is the corresponding pair of $(\Gamma_X, \Gamma_{X|Y})$. This completes the proof for item 2. We could have alternatively defined $\Gamma_{X|Y}^{-1} \widetilde{V}_j$ as the generalized eigenvector without any bearing on the spectrum. The choice we make is motivated by the orthogonality of eigenvectors we desire.

Let us define $V_j := \Gamma_X^{-1} \widetilde{V}_j$. Then item 1 follows by recognizing that (σ_j, V_j) is the corresponding generalized eigenvalue-eigenvector pair of $(\Gamma_X - \Gamma_{X|Y}, \Gamma_{X|Y})$. ∎

Proposition 38 *Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ be jointly Gaussian random variables as defined in Section 3.6.1. Let $\mathcal{P} \in \mathbb{R}^{m \times k}$ be a selection operator such that $Y_\mathcal{P} := \mathcal{P}^\top Y = [Y_{i_1}, \ldots, Y_{i_k}]^\top$. The expected symmetrized Kullback–Leiber divergence, $\mathbb{E}_{\pi_{Y_\mathcal{P}}} \left[D_{KL}^{sym}(\pi_{X|Y_\mathcal{P}}, \pi_X) \right]$, between the prior π_X and the posterior $\pi_{X|Y_\mathcal{P}}$ is equal to the sum of the generalized eigenvalues $\widehat{\sigma}_j$ of the definite pair $(\Gamma_{Y_\mathcal{P}} - \Gamma_{Y_\mathcal{P}|X}, \Gamma_{Y_\mathcal{P}|X})$ or equivalently $(\Gamma_X - \Gamma_{X|Y_\mathcal{P}}, \Gamma_{X|Y_\mathcal{P}})$, i.e.,*

$$\mathbb{E}_{\pi_{Y_\mathcal{P}}} \left[D_{KL}^{sym}(\pi_{X|Y_\mathcal{P}}, \pi_X) \right] := \mathbb{E}_{\pi_{Y_\mathcal{P}}} \left[D_{KL}(\pi_{X|Y_\mathcal{P}} \| \pi_X) + D_{KL}(\pi_X \| \pi_{X|Y_\mathcal{P}}) \right] = \sum_j \widehat{\sigma}_j.$$

Proof of Proposition 38 Since the prior and posterior are normal distributions, we can write the expected symmetrized KL divergence analytically,

$$\mathbb{E}_{\pi_{Y_\mathcal{P}}} \left[D_{KL}^{sym}(\pi_{X|Y_\mathcal{P}}, \pi_X) \right] = \frac{1}{2} \Big(\mathrm{trace}\left(\Gamma_X^{-1} \Gamma_{X|Y_\mathcal{P}} \right) + \mathbb{E}\left[\mu_{X|Y_\mathcal{P}}^\top \Gamma_X^{-1} \mu_{X|Y_\mathcal{P}} \right]$$
$$+ \mathrm{trace}\left(\Gamma_{X|Y_\mathcal{P}}^{-1} \Gamma_X \right) + \mathbb{E}\left[\mu_{X|Y_\mathcal{P}}^\top \Gamma_{X|Y_\mathcal{P}}^{-1} \mu_{X|Y_\mathcal{P}} \right] - 2n \Big). \tag{57}$$

Above we have assumed without any loss of generality that the prior has zero mean. The posterior mean $\mu_{X|Y_\mathcal{P}} := \Gamma_{X|Y_\mathcal{P}} G_\mathcal{P}^\top \Gamma_{Y_\mathcal{P}|X}^{-1} Y_\mathcal{P}$ is a function of the actual realization of data, unlike the covariance operator $\Gamma_{X|Y_\mathcal{P}}$ which is independent of data.

Simplifying the above expression is easier when we evaluate $\mathbb{E}_{\pi_{Y_\mathcal{P}}} \left[\mu_{X|Y_\mathcal{P}} \mu_{X|Y_\mathcal{P}}^\top \right]$.

$$\mathbb{E}_{\pi_{Y_\mathcal{P}}} \left[\mu_{X|Y_\mathcal{P}} \mu_{X|Y_\mathcal{P}}^\top \right] = \Gamma_{X|Y_\mathcal{P}} G_\mathcal{P}^\top \Gamma_{Y_\mathcal{P}|X}^{-1} \mathbb{E}\left[Y_\mathcal{P} Y_\mathcal{P}^\top \right] \Gamma_{Y_\mathcal{P}|X}^{-1} G_\mathcal{P} \Gamma_{X|Y_\mathcal{P}}, \tag{58a}$$

$$\overset{(i)}{=} \Gamma_{X|Y_\mathcal{P}} G_\mathcal{P}^\top \Gamma_{Y_\mathcal{P}|X}^{-1} \left(\Gamma_{Y_\mathcal{P}|X} + G_\mathcal{P} \Gamma_X G_\mathcal{P}^\top \right) \Gamma_{Y_\mathcal{P}|X}^{-1} G_\mathcal{P} \Gamma_{X|Y_\mathcal{P}}, \tag{58b}$$

$$= \Gamma_{X|Y_\mathcal{P}} \left(\Gamma_X^{-1} + G_\mathcal{P}^\top \Gamma_{Y_\mathcal{P}|X}^{-1} G_\mathcal{P} \right) \Gamma_X G_\mathcal{P}^\top \Gamma_{Y_\mathcal{P}|X}^{-1} G_\mathcal{P} \Gamma_{X|Y_\mathcal{P}}, \tag{58c}$$

$$\overset{(ii)}{=} \Gamma_{X|Y_\mathcal{P}} \Gamma_{X|Y_\mathcal{P}}^{-1} \Gamma_X \left(\Gamma_{X|Y_\mathcal{P}}^{-1} - \Gamma_X \right) \Gamma_{X|Y_\mathcal{P}}, \tag{58d}$$

$$= \Gamma_X - \Gamma_{X|Y_\mathcal{P}}. \tag{58e}$$

48

In the above set of equations, (i) holds since $\Gamma_{Y_\mathcal{P}|X} + G_\mathcal{P}\Gamma_X G_\mathcal{P}^\top = \mathbb{E}_{\pi_{Y_\mathcal{P}}}\left[Y_\mathcal{P} Y_\mathcal{P}^\top\right] =: \Gamma_{Y_\mathcal{P}}$ is the marginal of the data, and (ii) holds since $\Gamma_X^{-1} + G_\mathcal{P}^\top \Gamma_{Y_\mathcal{P}|X}^{-1} G_\mathcal{P} =: \Gamma_{X|Y_\mathcal{P}}^{-1}$ is the posterior precision.

Using the cyclic property of trace and linearity of the expectation operator it is now easy to see that,

$$\mathbb{E}_{\pi_{Y_\mathcal{P}}}\left[\mu_{X|Y_\mathcal{P}}^\top \Gamma_X^{-1} \mu_{X|Y_\mathcal{P}}\right] = \text{trace}\left(\Gamma_X^{-1}\left(\Gamma_X - \Gamma_{X|Y_\mathcal{P}}\right)\right) = n - \text{trace}\left(\Gamma_X^{-1}\Gamma_{X|Y_\mathcal{P}}\right), \qquad (59)$$

$$\mathbb{E}_{\pi_{Y_\mathcal{P}}}\left[\mu_{X|Y_\mathcal{P}}^\top \Gamma_{X|Y_\mathcal{P}}^{-1} \mu_{X|Y_\mathcal{P}}\right] = \text{trace}\left(\Gamma_{X|Y_\mathcal{P}}^{-1}\left(\Gamma_X - \Gamma_{X|Y_\mathcal{P}}\right)\right) = \text{trace}\left(\Gamma_{X|Y_\mathcal{P}}^{-1}\Gamma_X\right) - n. \quad (60)$$

Using (59) and (60) in (57) we have,

$$\mathbb{E}_{\pi_{Y_\mathcal{P}}}\left[D_{\text{KL}}^{sym}(\pi_{X|Y_\mathcal{P}}, \pi_X)\right] = \text{trace}\left(\Gamma_{X|Y_\mathcal{P}}^{-1}\Gamma_X\right) - n. \qquad (61)$$

The proof is completed by recognizing that $\text{trace}\left(\Gamma_{X|Y_\mathcal{P}}^{-1}\Gamma_X\right)$ is the sum of generalized eigenvalues of the definite pair $(\Gamma_X, \Gamma_{X|Y_\mathcal{P}})$, which we know from Theorem 10 is $n + \sum_j \widehat{\sigma}_j$, where $\widehat{\sigma}_j$ are the generalized eigenvalues of the definite pair $(\Gamma_X - \Gamma_{X|Y_\mathcal{P}}, \Gamma_{X|Y_\mathcal{P}})$ or equivalently $(\Gamma_{Y_\mathcal{P}} - \Gamma_{Y_\mathcal{P}|X}, \Gamma_{Y_\mathcal{P}|X})$. ∎

B.2.1 BOUNDS ON SUB/SUP–MODULARITY RATIOS

We adapt an eigenvalue interlacing result for definite pairs which will be useful in proving Proposition 11.

Theorem 39 (Cauchy Interlacing Theorem) *(Horn and Johnson, 2012; Bhatia, 1997), (see also Kressner et al., 2014, Theorem 2.3), (see also Kovač-Striko and Veselić, 1995, Theorem 2.1) Let $A, B \in \mathbb{R}^{n \times n}$ with $A \succeq 0, B \succ 0$, and let $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_{\widehat{n}} > 0$, $\widehat{n} \leq n$, be the eigenvalues of (A, B). For any $Z \in \mathbb{R}^{n \times p}$, with $p \leq n$ and full column-rank, let $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_{\widehat{p}} > 0$, $\widehat{p} \leq p$, be the eigenvalues of $(Z^\top A Z, Z^\top B Z)$. Then:*

$$\gamma_{n-p+k} \leq \mu_k \leq \gamma_k, \qquad k = 1, \ldots, \widehat{p}.$$

If additionally $A \succ 0$, then $\widehat{n} = n$ and $\widehat{p} = p$.

Proof of Proposition 11 Bian et al. (2017) provide a bound on the submodularity ratio for the Bayesian A-optimality design criterion. Our arguments use that proof as a template, but are modified to account for the mutual information design criterion. In addition we seek a bound for the supermodularity ratio, which intriguingly is the same.

In order to bound the *submodularity* ratio we need to lower bound,

$$\frac{\sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} \rho_\nu(\mathscr{B})}{\rho_\mathscr{A}(\mathscr{B})} = \frac{\sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} F(\nu \cup \mathscr{B}) - F(\mathscr{B})}{F(\mathscr{A} \cup \mathscr{B}) - F(\mathscr{B})}, \qquad \forall \mathscr{A}, \mathscr{B} \subseteq \mathscr{V}.$$

Similarly, in order to bound the *supermodularity* ratio we need to lower bound,

$$\frac{\rho_\mathscr{A}(\mathscr{B})}{\sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} \rho_\nu(\mathscr{B})} = \frac{F(\mathscr{A} \cup \mathscr{B}) - F(\mathscr{B})}{\sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} F(\nu \cup \mathscr{B}) - F(\mathscr{B})}, \qquad \forall \mathscr{A}, \mathscr{B} \subseteq \mathscr{V}.$$

It is clear that we can achieve the above tasks by finding lower and upper bounds for $F(\mathscr{A} \cup \mathscr{B}) - F(\mathscr{B})$ and $\sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} F(\nu \cup \mathscr{B}) - F(\mathscr{B})$ respectively.

First we introduce some notation to aid the proof. For any index set $\mathscr{A}$ with corresponding selection operator $\mathcal{P}_{\mathscr{A}}$, let $\zeta_j^{\mathscr{A}}$ represent the eigenvalue of the compressed pencil $\left(\mathcal{P}_{\mathscr{A}}^{\top} \Gamma_Y \mathcal{P}_{\mathscr{A}}, \mathcal{P}_{\mathscr{A}}^{\top} \Gamma_{Y|X} \mathcal{P}_{\mathscr{A}}\right)$. ζ_j *without any superscript* will continue to represent any eigenvalue of the pencil $\left(\Gamma_Y, \Gamma_{Y|X}\right)$. Since we consider mutual information as the objective, $F(\mathscr{A})$ in terms of the eigenvalues $\zeta_j^{\mathscr{A}}$ is simply ((11) and (13)),

$$F(\mathscr{A}) = \frac{1}{2} \log \prod_{j=1}^{|\mathscr{A}|} \zeta_j^{\mathscr{A}}. \tag{62}$$

We now state a simple corollary of Theorem 39. We omit the proof since it is quite trivial. For any two index sets $\mathscr{A}, \mathscr{B}$ we have,

$$\zeta_j^{\mathscr{A} \cup \mathscr{B}} \geq \zeta_j^{\mathscr{B}} \geq \zeta_{|\mathscr{A} \cup \mathscr{B}| - |\mathscr{B}| + j}^{\mathscr{A} \cup \mathscr{B}}, \qquad j = 1, \ldots, |\mathscr{B}|. \tag{63}$$

As in Theorem 39, in the above statement we have ordered the eigenvalues such that $\zeta_1^{\mathscr{B}} \geq \cdots \geq \zeta_{|\mathscr{B}|}^{\mathscr{B}}$ and $\zeta_1^{\mathscr{A} \cup \mathscr{B}} \geq \cdots \geq \zeta_{|\mathscr{A} \cup \mathscr{B}|}^{\mathscr{A} \cup \mathscr{B}}$. This simple result proves handy in bounding $F(\mathscr{A} \cup \mathscr{B}) - F(\mathscr{B})$.

$$F(\mathscr{A} \cup \mathscr{B}) - F(\mathscr{B}) = \frac{1}{2} \log \prod_{j=1}^{|\mathscr{A} \cup \mathscr{B}|} \zeta_j^{\mathscr{A} \cup \mathscr{B}} - \frac{1}{2} \log \prod_{j=1}^{|\mathscr{B}|} \zeta_j^{\mathscr{B}}, \tag{64a}$$

$$= \frac{1}{2} \log \prod_{j=1}^{|\mathscr{A} \cup \mathscr{B}| - |\mathscr{B}|} \zeta_j^{\mathscr{A} \cup \mathscr{B}} + \frac{1}{2} \log \prod_{j=1}^{|\mathscr{B}|} \frac{\zeta_{|\mathscr{A} \cup \mathscr{B}| - |\mathscr{B}| + j}^{\mathscr{A} \cup \mathscr{B}}}{\zeta_j^{\mathscr{B}}}, \tag{64b}$$

$$\leq \frac{1}{2} \log \prod_{j=1}^{|\mathscr{A} \setminus \mathscr{B}|} \zeta_j^{\mathscr{A} \cup \mathscr{B}}, \quad \because \quad (63) \Rightarrow \frac{\zeta_{|\mathscr{A} \cup \mathscr{B}| - |\mathscr{B}| + j}^{\mathscr{A} \cup \mathscr{B}}}{\zeta_j^{\mathscr{B}}} \leq 1, \tag{64c}$$

$$\leq \frac{1}{2} \log \left(\zeta_1^{\mathscr{A} \cup \mathscr{B}}\right)^{|\mathscr{A} \setminus \mathscr{B}|}, \quad \because \quad \zeta_1^{\mathscr{A} \cup \mathscr{B}} \geq \zeta_j^{\mathscr{A} \cup \mathscr{B}} \quad \forall j, \tag{64d}$$

$$\leq \frac{|\mathscr{A} \setminus \mathscr{B}|}{2} \log \zeta_1^{\mathscr{A} \cup \mathscr{B}}. \tag{64e}$$

$$F(\mathscr{A} \cup \mathscr{B}) - F(\mathscr{B}) = \frac{1}{2} \log \prod_{j=1}^{|\mathscr{A} \cup \mathscr{B}| - |\mathscr{B}|} \zeta_{|\mathscr{B}| + j}^{\mathscr{A} \cup \mathscr{B}} + \frac{1}{2} \log \prod_{j=1}^{|\mathscr{B}|} \frac{\zeta_j^{\mathscr{A} \cup \mathscr{B}}}{\zeta_j^{\mathscr{B}}}, \tag{65a}$$

$$\geq \frac{1}{2} \log \prod_{j=1}^{|\mathscr{A} \setminus \mathscr{B}|} \zeta_{|\mathscr{B}| + j}^{\mathscr{A} \cup \mathscr{B}}, \quad \because \quad (63) \Rightarrow \frac{\zeta_j^{\mathscr{A} \cup \mathscr{B}}}{\zeta_j^{\mathscr{B}}} \geq 1, \tag{65b}$$

$$\geq \frac{1}{2} \log \left(\zeta_{|\mathscr{A} \cup \mathscr{B}|}^{\mathscr{A} \cup \mathscr{B}}\right)^{|\mathscr{A} \setminus \mathscr{B}|}, \quad \because \quad \zeta_{|\mathscr{A} \cup \mathscr{B}|}^{\mathscr{A} \cup \mathscr{B}} \leq \zeta_{|\mathscr{B}| + j}^{\mathscr{A} \cup \mathscr{B}} \quad \forall j, \tag{65c}$$

$$\geq \frac{|\mathscr{A} \setminus \mathscr{B}|}{2} \log \zeta_{|\mathscr{A} \cup \mathscr{B}|}^{\mathscr{A} \cup \mathscr{B}}. \tag{65d}$$

Specializing (63) to the case when the set $\mathscr{A}$ is a singleton set consisting of the element ν we have the following result,

$$\zeta_j^{\nu \cup \mathscr{B}} \geq \zeta_j^{\mathscr{B}} \geq \zeta_{j+1}^{\nu \cup \mathscr{B}}, \qquad j = 1, \ldots, |\mathscr{B}|. \tag{66}$$

Using (66) we can bound the term $\sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} F(\nu \cup \mathscr{B}) - F(\mathscr{B})$ using similar arguments as in (64) and (65).

$$\sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} F(\nu \cup \mathscr{B}) - F(\mathscr{B}) = \sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} \left(\frac{1}{2} \log \prod_{j=1}^{|\mathscr{B}|+1} \zeta_j^{\nu \cup \mathscr{B}} - \frac{1}{2} \log \prod_{j=1}^{|\mathscr{B}|} \zeta_j^{\mathscr{B}} \right), \tag{67a}$$

$$= \sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} \left(\frac{1}{2} \log \zeta_{|\mathscr{B}|+1}^{\nu \cup \mathscr{B}} + \frac{1}{2} \log \prod_{j=1}^{|\mathscr{B}|} \frac{\zeta_j^{\nu \cup \mathscr{B}}}{\zeta_j^{\mathscr{B}}} \right), \tag{67b}$$

$$\geq \frac{|\mathscr{A} \setminus \mathscr{B}|}{2} \log \zeta_{|\mathscr{B}|+1}^{\nu^\dagger \cup \mathscr{B}}, \quad \because \quad \frac{\zeta_j^{\nu \cup \mathscr{B}}}{\zeta_j^{\mathscr{B}}} \geq 1, \tag{67c}$$

where $\nu^\dagger = \arg\min_{\nu \in \mathscr{A} \setminus \mathscr{B}} \zeta_{|\mathscr{B}|+1}^{\nu \cup \mathscr{B}}$.

$$\sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} F(\nu \cup \mathscr{B}) - F(\mathscr{B}) = \sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} \left(\frac{1}{2} \log \zeta_1^{\nu \cup \mathscr{B}} + \frac{1}{2} \log \prod_{j=1}^{|\mathscr{B}|} \frac{\zeta_{j+1}^{\nu \cup \mathscr{B}}}{\zeta_j^{\mathscr{B}}} \right), \tag{68a}$$

$$\leq \frac{|\mathscr{A} \setminus \mathscr{B}|}{2} \log \zeta_1^{\nu^\ddagger \cup \mathscr{B}}, \quad \because \quad \frac{\zeta_{j+1}^{\nu \cup \mathscr{B}}}{\zeta_j^{\mathscr{B}}} \leq 1, \tag{68b}$$

where $\nu^\ddagger = \arg\max_{\nu \in \mathscr{A} \setminus \mathscr{B}} \zeta_1^{\nu \cup \mathscr{B}}$. Gathering the results from (64), (65), (67) and (68) we have,

$$\frac{|\mathscr{A} \setminus \mathscr{B}|}{2} \log \zeta_{|\mathscr{B}|+1}^{\nu^\dagger \cup \mathscr{B}} \leq \sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} F(\nu \cup \mathscr{B}) - F(\mathscr{B}) \leq \frac{|\mathscr{A} \setminus \mathscr{B}|}{2} \log \zeta_1^{\nu^\ddagger \cup \mathscr{B}}, \tag{69}$$

$$\frac{|\mathscr{A} \setminus \mathscr{B}|}{2} \log \zeta_{|\mathscr{A} \cup \mathscr{B}|}^{\mathscr{A} \cup \mathscr{B}} \leq F(\mathscr{A} \cup \mathscr{B}) - F(\mathscr{B}) \leq \frac{|\mathscr{A} \setminus \mathscr{B}|}{2} \log \zeta_1^{\mathscr{A} \cup \mathscr{B}}. \tag{70}$$

Using (69) and (70) we have,

$$\frac{\sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} F(\nu \cup \mathscr{B}) - F(\mathscr{B})}{F(\mathscr{A} \cup \mathscr{B}) - F(\mathscr{B})} \geq \frac{\log \zeta_{|\mathscr{B}|+1}^{\nu^\dagger \cup \mathscr{B}}}{\log \zeta_1^{\mathscr{A} \cup \mathscr{B}}} \geq \frac{\log \zeta_{\min}}{\log \zeta_{\max}}. \tag{71}$$

$$\frac{F(\mathscr{A} \cup \mathscr{B}) - F(\mathscr{B})}{\sum_{\nu \in \mathscr{A} \setminus \mathscr{B}} F(\nu \cup \mathscr{B}) - F(\mathscr{B})} \geq \frac{\log \zeta_{|\mathscr{A} \cup \mathscr{B}|}^{\mathscr{A} \cup \mathscr{B}}}{\log \zeta_1^{\nu^\ddagger \cup \mathscr{B}}} \geq \frac{\log \zeta_{\min}}{\log \zeta_{\max}}. \tag{72}$$

$$\blacksquare$$

51

B.3 Proofs concerning MM algorithms for linear Bayesian experimental design

This subsection is devoted to proving Proposition 13 and its corollaries. We begin by stating certain standard results in linear algebra, subsequently deducing useful corollaries, and some helpful lemmas from them, all of which contribute to the argument of the main result.

Theorem 40 (Hadamard's inequality) *(Horn and Johnson, 2012) Let $A \in \mathbb{R}^{m \times m}$ be any positive definite Hermitian matrix, with a_{ij} representing its individual entries. Then,*

$$\det A \leq a_{11} \cdots a_{mm}$$

Corollary 41 *Let $A \in \mathbb{R}^{m \times m}$ be any positive definite Hermitian matrix, and $\mathcal{P} \in \mathbb{R}^{m \times k}, k < m$ be a selection operator (Definition 2), then:*

$$\log\left(\det\left(\mathcal{P}^\top A \mathcal{P}\right)\right) \leq \operatorname{trace}\left(\mathcal{P}^\top \log\left(\operatorname{diag}\left(A\right)\right)\mathcal{P}\right)$$

Proof of Corollary 41 The corollary follows immediately from Theorem 40. ■

Remark 42 *The $\log \det$ of a principal submatrix is a submodular function with respect to the indices defining the submatrices (Gantmacher and Kreĭn, 1960; Kotelyanskiĭ, 1950; Fan, 1967, 1968; Kelmans and Kimelfeld, 1983; Johnson and Barrett, 1985). We know that submodularity implies subadditivity for nonnegative functions, thus the modular upper bound in the statement of Corollary 41 can be understood in that sense as well.*

Theorem 43 (Operator concave inequality) *(Davis, 1957; Hansen, 1980; Hansen and Kjærgård Pedersen, 1982; Hansen and Pedersen, 2003; Bourin, 2006) Let A be any Hermitian operator defined on a Hilbert space, and K be an isometry on a subspace of the Hilbert space. Denote the compression of A by K as K^*AK, then for any operator concave function $\phi(\cdot)$ we have that,*

$$K^*\phi(A)K \preceq \phi(K^*AK)$$

Theorem 44 (Matrix versions of the Cauchy and Kantorovich inequalities) *(Marshall and Olkin, 1990; Mond and Pečarič, 1994; Bourin, 2005) Let $A \in \mathbb{R}^{m \times m}$ be any positive definite Hermitian matrix, with eigenvalues contained in the interval $[\lambda_s, \lambda_l] \subset \mathbb{R}_{>0}$. If K is an isometry then:*

$$\frac{4\lambda_s\lambda_l}{(\lambda_s + \lambda_l)^2}K^*A^{-1}K \preceq (K^*AK)^{-1} \preceq K^*A^{-1}K$$

As pointed out by Marshall and Olkin (1990), if K is merely a selection operator, then the upper bound of Theorem 44 reduces to a standard result in linear algebra, first established by Chollet (1982) and now a common theorem in texts (see Horn and Johnson, 2012, Theorem 7.7.15.); the inverse of a principal submatrix of any definite matrix is less than or equal to the corresponding submatrix of the inverse.

Corollary 45 *Let $A \in \mathbb{R}^{m \times m}$ be any positive definite Hermitian matrix with an eigenvalue interval $[\lambda_s, \lambda_l]$ as prescribed in Theorem 44. Define the scalar parameter $\varrho := \frac{4\lambda_s \lambda_l}{(\lambda_s + \lambda_l)^2}$. If $K \in \mathbb{R}^{m \times k}, k < m$ is an isometry, $K^*K = I_k$, then:*

$$\text{trace}\left(K^* \log\left(A\right) K\right) \le \log\det\left(K^*AK\right) \le \text{trace}\left(K^*\left(\log\left(A\right) - I_m \log\left(\varrho\right)\right) K\right)$$

Proof of Corollary 45 Since $A \succ 0$, its compression by K, $K^*AK \succ 0$. Hence we have the identity $\text{trace}\left(\log\left(K^*AK\right)\right) = \log\det\left(K^*AK\right)$. Since log is a concave function, we claim using Theorem 43 $\log\left(K^*AK\right) \succeq K^*\log\left(A\right) K$, and this implies the lower bound.

To prove the upper bound, we begin by using the identity

$$\log\det\left(K^*AK\right) = -\log\det\left(K^*AK\right)^{-1} = -\text{trace}\left(\log\left(K^*AK\right)^{-1}\right). \tag{73}$$

Using the lower bound in Theorem 44 we have,

$$\log\det\left(\varrho K^* A^{-1} K\right) \le \log\det\left(K^*AK\right)^{-1}. \tag{74}$$

$$\Rightarrow -\text{trace}\left(\log\left(K^*AK\right)^{-1}\right) \le -\text{trace}\left(\log\left(K^* A^{-1} K\right)\right) - \text{trace}\left(I_k \log\left(\varrho\right)\right). \tag{75}$$

$K^* A^{-1} K$ is a compression by K of $A^{-1} \succ 0$. Using Theorem 43 we claim,

$$\log\left(K^* A^{-1} K\right) \succeq K^* \log\left(A^{-1}\right) K = -K^* \log\left(A\right) K. \tag{76}$$

$$\Rightarrow -\text{trace}\left(\log\left(K^* A^{-1} K\right)\right) \le \text{trace}\left(K^* \log\left(A\right) K\right). \tag{77}$$

Using (77) in (75) we have,

$$-\text{trace}\left(\log\left(K^*AK\right)^{-1}\right) \le \text{trace}\left(K^* \log\left(A\right) K\right) - \text{trace}\left(I_k \log\left(\varrho\right)\right). \tag{78}$$

The upper bound in the corollary statement is now immediately obtained by linearity of the trace operator and the isometry property of K. ∎

Proof of Proposition 13 Equation (13) provides an expression for the mutual information $\mathcal{I}(X; Y_{\mathcal{P}})$,

$$\mathcal{I}(X; Y_{\mathcal{P}}) = \frac{1}{2} \log\left(\frac{\det\left(\mathcal{P}^\top \Gamma_Y \mathcal{P}\right)}{\det\left(\mathcal{P}^\top \Gamma_{Y|X} \mathcal{P}\right)}\right).$$

The proposition statement and the alternative modular bounds in (20) are easily obtained by using Corollaries 41 and 45 on the two log det terms. ∎

Proof of Corollary 15 Eq. (21) is simply a restatement of the lower bound in Proposition 13. The equivalence between (21) and (22) is trivial. Rewriting (22) with $Y \setminus Y_{\mathcal{P}}$ as the argument, and subtracting that term from $\mathcal{I}(X; Y)$ we obtain (21). In a similar way we can obtain (22) from (21). ∎

Proof of Corollary 14 If $Y_{i_k}|X$ are independent, then we know $\Gamma_{Y|X}$ is diagonal, and hence $\log \operatorname{diag}\left(\Gamma_{Y|X}\right) = \log\left(\Gamma_{Y|X}\right)$. Thus the corollary statement is apparent as a consequence of Corollary 15. ∎

Proof of Proposition 16 Observe that

$$\mathcal{I}(X; Y \setminus Y_{\mathcal{P}}|Y_{\mathcal{P}_1}) - \mathcal{I}(X; Y \setminus Y_{\mathcal{P}}, Y_{\mathcal{P}_1}) = \mathcal{I}(Y_{\mathcal{P}_1}; Y \setminus Y_{\mathcal{P}}, Y_{\mathcal{P}_1}|X) - \mathcal{I}(Y_{\mathcal{P}_1}; Y \setminus Y_{\mathcal{P}}, Y_{\mathcal{P}_1}). \quad (79)$$

This implies

$$\mathcal{I}(X; Y \setminus Y_{\mathcal{P}}|Y_{\mathcal{P}_1}) - \mathcal{I}(X; Y \setminus Y_{\mathcal{P}}, Y_{\mathcal{P}_1}) \leq \mathcal{I}(Y_{\mathcal{P}_1}; Y \setminus Y_{\mathcal{P}}, Y_{\mathcal{P}_1}|X). \quad (80)$$

Using (80), we have

$$\mathcal{I}(X; Y|Y_{\mathcal{P}_1}) - \mathcal{I}(X; Y \setminus Y_{\mathcal{P}_1}, Y_{\mathcal{P}}) \leq \mathcal{I}(X; Y|Y_{\mathcal{P}_1}) - \mathcal{I}(X; Y \setminus Y_{\mathcal{P}}|Y_{\mathcal{P}_1}) + \mathcal{I}(Y_{\mathcal{P}_1}; Y \setminus Y_{\mathcal{P}}, Y_{\mathcal{P}_1}|X). \quad (81)$$

Consider the term $\mathcal{I}(X; Y|Y_{\mathcal{P}_1}) - \mathcal{I}(X; Y \setminus Y_{\mathcal{P}}|Y_{\mathcal{P}_1})$; adapting the result in Corollary 15 we have the following bound:

$$
\begin{aligned}
\mathcal{I}(X; Y|Y_{\mathcal{P}_1}) - \mathcal{I}(X; Y \setminus Y_{\mathcal{P}}|Y_{\mathcal{P}_1}) \leq{}& \operatorname{trace}\left(\widehat{\mathcal{P}}^{\top}\left(\log\left(\Gamma_{Y|Y_{\mathcal{P}_1}}\right) - \log\left(\Gamma_{Y|X,Y_{\mathcal{P}_1}}\right)\right)\widehat{\mathcal{P}}\right) \\
&+ \operatorname{trace}\left(\widehat{\mathcal{P}}^{c\top}\left(\log\operatorname{diag}\left(\Gamma_{Y|X,Y_{\mathcal{P}_1}}\right) - \log\left(\Gamma_{Y|X,Y_{\mathcal{P}_1}}\right)\right)\widehat{\mathcal{P}}^c\right).
\end{aligned}
\quad (82)
$$

Now consider the term $\mathcal{I}(Y_{\mathcal{P}_1}; Y \setminus Y_{\mathcal{P}}, Y_{\mathcal{P}_1}|X)$. Using Corollaries 41 and 45 we have the following bound:

$$\mathcal{I}(Y_{\mathcal{P}_1}; Y \setminus Y_{\mathcal{P}}, Y_{\mathcal{P}_1}|X) \leq \operatorname{trace}\left(\widehat{\mathcal{P}}^{c\top}\left(\log\operatorname{diag}\left(\Gamma_{Y \setminus Y_{\mathcal{P}_1}|X}\right) - \log\left(\Gamma_{Y|X,Y_{\mathcal{P}_1}}\right)\right)\widehat{\mathcal{P}}^c\right). \quad (83)$$

Adding (80) and (82), we get the proposition statement. ∎

References

Alen Alexanderian, Philip J. Gloor, and Omar Ghattas. On Bayesian A- and D- optimal experimental designs in infinite dimensions. *Bayesian Anal.*, 11(3):671–695, 09 2016.

Francis Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2-3):145–373, 2013. ISSN 1935-8237.

Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3(Jul):1–48, 2002.

Zhaojun Bai, Gark Fahey, and Gene Golub. Some large-scale matrix computation problems. *Journal of Computational and Applied Mathematics*, 74(1):71 – 89, 1996. ISSN 0377-0427.

Eric Balkanski and Yaron Singer. Approximation guarantees for adaptive sampling. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 384–393. PMLR, 10–15 Jul 2018a.

Eric Balkanski and Yaron Singer. The adaptive complexity of maximizing a submodular function. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, pages 1138–1151, New York, NY, USA, 2018b. Association for Computing Machinery. ISBN 9781450355599.

Eric Balkanski, Adam Breuer, and Yaron Singer. Non-monotone submodular maximization in exponentially fewer iterations. *arXiv preprint arXiv:1807.11462*, 2018.

Eric Balkanski, Aviad Rubinstein, and Yaron Singer. An optimal approximation for submodular maximization under a matroid constraint in the adaptive complexity model. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, pages 66–77, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367059.

Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer-Verlag New York, 1997.

Andrew An Bian, Joachim M Buhmann, Andreas Krause, and Sebastian Tschiatschek. Guarantees for greedy maximization of non-submodular functions with applications. *arXiv preprint arXiv:1703.02100*, 2017.

Y. Bian, A. Gronskiy, and J. M. Buhmann. Greedy maxcut algorithms and their information content. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5, April 2015.

Ilija Bogunovic, Junyao Zhao, and Volkan Cevher. Robust maximization of non-submodular objectives. *arXiv preprint arXiv:1802.07073*, 2018.

Jean-Christophe Bourin. Symmetric norms and reverse inequalities to Davis and Hansen-Pedersen characterizations of operator convexity. *arXiv preprint math/0511734*, 2005.

Jean-Christophe Bourin. A concavity inequality for symmetric norms. *Linear Algebra and its Applications*, 413(1):212 – 217, 2006. ISSN 0024-3795.

Christos Boutsidis, Michael W. Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the 2009 Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 968–977, 2009.

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

Samprit Chatterjee and Ali S. Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statist. Sci.*, 1(3):379–393, 08 1986.

Chandra Chekuri and Kent Quanrud. Parallelizing greedy for submodular set function maximization in matroids and beyond. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, pages 78–89, New York, NY, USA, 2019a. Association for Computing Machinery. ISBN 9781450367059.

Chandra Chekuri and Kent Quanrud. Submodular function maximization in parallel via the multilinear relaxation. In *Proceedings of the 2019 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 303–322, 2019b.

Lin Chen, Moran Feldman, and Amin Karbasi. Unconstrained submodular maximization with constant adaptive complexity. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, pages 102–113, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367059.

John Chollet. On principal submatrices. *Linear and Multilinear Algebra*, 11(3):283–285, 1982.

Michele Conforti and Gérard Cornuéjols. Submodular set functions, matroids and the greedy algorithm: Tight worst-case bounds and some generalizations of the rado-edmonds theorem. *Discrete Applied Mathematics*, 7(3):251 – 274, 1984. ISSN 0166-218X.

Christophe Couvreur and Yoram Bresler. On the optimality of the backward greedy algorithm for the subset selection problem. *SIAM Journal on Matrix Analysis and Applications*, 21(3):797–808, 2000.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, 2 edition, 2006. ISBN 0471241954.

C. R. Crawford. A stable generalized eigenvalue problem. *SIAM Journal on Numerical Analysis*, 13(6):854–860, 1976. ISSN 00361429.

Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 1057–1064, New York, NY, USA, June 2011. ACM. ISBN 978-1-4503-0619-5.

Chandler Davis. A Schwarz inequality for convex operator functions. *Proceedings of the American Mathematical Society*, 8(1):42–44, 1957.

Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011.

Jack Edmonds. Matroids and the greedy algorithm. *Mathematical Programming*, 1(1): 127–136, Dec 1971. ISSN 1436-4646.

Ethan R. Elenberg, Rajiv Khanna, Alexandros G. Dimakis, and Sahand Negahban. Restricted strong convexity implies weak submodularity. *The Annals of Statistics*, 46(6B): 3539 – 3568, 2018.

Matthew Fahrbach, Vahab Mirrokni, and Morteza Zadimoghaddam. Non-monotone submodular maximization with nearly optimal adaptivity and query complexity. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1833–1842. PMLR, 09–15 Jun 2019.

Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

Jianqing Fan, Yuan Liao, and Martina Mincheva. High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics*, 39(6):3320, 2011.

Ky Fan. Subadditive functions on a distributive lattice and an extension of szász's inequality. *Journal of Mathematical Analysis and Applications*, 18(2):262 – 268, 1967. ISSN 0022-247X.

Ky Fan. An inequality for subadditive functions on a distributive lattice, with application to determinantal inequalities. *Linear Algebra and its Applications*, 1(1):33 – 38, 1968. ISSN 0024-3795.

Valerii Vadimovich Fedorov. *Theory of Optimal Experiments*. Elsevier, 2013.

Ronald Aylmer Fisher. Design of experiments. *Br Med J*, 1(3923):554–554, 1936.

Satoru Fujishige. *Submodular Functions and Optimization*, volume 58 of *Annals of Discrete Mathematics*. Elsevier, 2 edition, 2005.

F.R. Gantmacher and M.G. Kreĭn. *Oszillationsmatrizen, Oszillationskerne Und Kleine Schwingungen Mechanischer Systeme*, volume 5. Akademie-Verlag, 1960.

Loïc Giraldi, Olivier P Le Maître, Ibrahim Hoteit, and Omar M Knio. Optimal projection of observations in a Bayesian setting. *arXiv:1709.06606*, 2017.

Gene H Golub and Gérard Meurant. Matrices, moments and quadrature. In D F Griffiths and G A Watson, editors, *Numerical Analysis 1993, Proceedings of the 15th Dundee Conference, June-July*, Pitman research notes in mathematics series, pages 105–156, Harlow, Essex, England, 1994. Longman Scientific & Technical.

Gene H Golub and Gérard Meurant. *Matrices, Moments and Quadrature with Applications*. Princeton University Press, 2009.

Frank Hansen. An operator inequality. *Mathematische Annalen*, 246(3):249–250, Oct 1980. ISSN 1432-1807.

Frank Hansen and Gert Kjærgård Pedersen. Jensen's inequality for operators and Löwner's theorem. *Mathematische Annalen*, 258(3):229–241, Sep 1982. ISSN 1432-1807.

Frank Hansen and Gert K. Pedersen. Jensen's operator inequality. *Bulletin of the London Mathematical Society*, 35(4):553–564, 2003.

N. Higham. *Functions of Matrices*. Society for Industrial and Applied Mathematics, 2008.

David C. Hoaglin and Roy E. Welsch. The hat matrix in regression and ANOVA. *The American Statistician*, 32(1):17–22, 1978.

Thibaut Horel and Yaron Singer. Maximization of approximately submodular functions. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3045–3053. Curran Associates, Inc., 2016.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

James W Hurrell, Marika M Holland, Peter R Gent, Steven Ghan, Jennifer E Kay, Paul J Kushner, J-F Lamarque, William G Large, D Lawrence, Keith Lindsay, et al. The community earth system model: a framework for collaborative research. *Bulletin of the American Meteorological Society*, 94(9):1339–1360, 2013.

Victor P. Il'ev. An approximation guarantee of the greedy descent algorithm for minimizing a supermodular set function. *Discrete Applied Mathematics*, 114(1):131 – 146, 2001. ISSN 0166-218X.

Rishabh Iyer and Jeff Bilmes. Algorithms for approximate minimization of the difference between submodular functions, with applications. In *Proceedings of the Twenty-Eighth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-12)*, pages 407–417, Corvallis, Oregon, 2012. AUAI Press.

Rishabh Iyer and Jeff Bilmes. Polyhedral aspects of submodularity, convexity and concavity. *arXiv preprint arXiv:1506.07329*, 2015.

Rishabh Iyer, Stefanie Jegelka, and Jeff Bilmes. Mirror descent like algorithms for submodular optimization. In *NIPS Workshop on Discrete Optimization in Machine Learning (DISCML)*, 2012.

Rishabh Iyer, Stefanie Jegelka, and Jeff Bilmes. Fast semidifferential-based submodular function optimization. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 855–863, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

S. Jegelka and J. Bilmes. Submodularity beyond submodular energies: Coupling edges in graph cuts. In *CVPR 2011*, pages 1897–1904, June 2011.

Charles R. Johnson and Wayne W. Barrett. Spanning-tree extensions of the hadamard-fischer inequalities. *Linear Algebra and its Applications*, 66:177 – 193, 1985. ISSN 0024-3795.

Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*, volume 160. Springer Science & Business Media, 2006.

Orcun Karaca and Maryam Kamgarpour. Exploiting weak supermodularity for coalition-proof mechanisms. *arXiv preprint arXiv:1803.11030*, 2018.

A.K. Kelmans and B.N. Kimelfeld. Multiplicative submodularity of a matrix's principal minor as a function of the set of its rows and some combinatorial applications. *Discrete Mathematics*, 44(1):113 – 116, 1983. ISSN 0012-365X.

Rajiv Khanna, Ethan Elenberg, Alex Dimakis, Sahand Negahban, and Joydeep Ghosh. Scalable Greedy Feature Selection via Weak Submodularity. In Aarti Singh and Jerry

Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1560–1568, Fort Lauderdale, FL, USA, 20–22 Apr 2017a. PMLR.

Rajiv Khanna, Ethan R. Elenberg, Alexandros G. Dimakis, Joydeep Ghosh, and Sahand Negahban. On approximation guarantees for greedy low rank optimization. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1837–1846. PMLR, 06–11 Aug 2017b.

D.M. Kotelyanskiĭ. On the theory of nonnegative and oscillating matrices. *Ukrains'kyi Matematychnyi Zhurnal*, 2(2):94–101, 1950.

J. Kovač-Striko and K. Veselić. Trace minimization and definiteness of symmetric pencils. *Linear Algebra and its Applications*, 216:139 – 158, 1995. ISSN 0024-3795.

Andreas Krause and Daniel Golovin. Submodular function maximization. In Bordeaux Lucas, Hamadi Youssef, and Kohli Pushmeet, editors, *Tractability, Practical Approaches to Hard Problems*, chapter 3. Cambridge University Press, 2014.

Daniel Kressner, Marija Miloloža Pandur, and Meiyue Shao. An indefinite variant of LOBPCG for definite matrix pencils. *Numerical Algorithms*, 66(4):681–703, Aug 2014. ISSN 1572-9265.

Alex Kulesza and Ben Taskar. k-DPPs: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1193–1200, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2-3):123–286, 2012. ISSN 1935-8237.

K. Lange. *MM Optimization Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Y. Liu, Z. Zhang, E. K. P. Chong, and A. Pezeshki. Performance bounds for the k-batch greedy strategy in optimization problems with curvature. In *2016 American Control Conference (ACC)*, pages 7177–7182, July 2016.

László Lovász. *Combinatorial Problems and Exercises*, volume 361. American Mathematical Soc., 2007.

Dan Lu and Daniel Ricciuto. Efficient surrogate modeling methods for large-scale earth system models based on machine learning techniques. *arXiv preprint arXiv:1901.05125*, 2019.

Zelda E. Mariet and Suvrit Sra. Elementary symmetric polynomials for optimal experimental design. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2139–2148. Curran Associates, Inc., 2017.

A. W. Marshall and I. Olkin. Matrix versions of the Cauchy and Kantorovich inequalities. *Aequationes Mathematicae*, 40(1):89–93, Dec 1990. ISSN 1420-8903.

Baharan Mirzasoleiman, Amin Karbasi, Rik Sarkar, and Andreas Krause. Distributed submodular maximization: Identifying representative elements in massive data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

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

B. Mond and J.D. Pečariĉ. A matrix version of the Ky Fan generalization of the Kantorovich inequality. *Linear and Multilinear Algebra*, 36(3):217–221, 1994.

Mukund Narasimhan and Jeff Bilmes. A submodular-supermodular procedure with applications to discriminative structure learning. In *Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 404–412, Arlington, Virginia, 2005. AUAI Press.

G. L. Nemhauser and L. A. Wolsey. Best algorithms for approximating the maximum of a submodular set function. *Mathematics of Operations Research*, 3(3):177–188, 1978. ISSN 0364765X, 15265471.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, Dec 1978. ISSN 1436-4646.

Christos H Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Courier Corporation, 1998.

Friedrich. Pukelsheim. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006.

Sharon Qian and Yaron Singer. Fast parallel algorithms for statistical subset selection problems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Daniel Ricciuto. Simplified version of Energy Exascale Earth System (E3SM) land model, 2020. URL `https://github.com/dmricciuto/OSCM_SciDAC/tree/master/models/simple_ELM`.

Daniel Ricciuto, Khachik Sargsyan, and Peter Thornton. The impact of parametric uncertainties on biogeochemistry in the E3SM land model. *Journal of Advances in Modeling Earth Systems*, 10(2):297–319, 2018.

T. Robertazzi and S. Schwartz. An accelerated sequential algorithm for producing D-optimal designs. *SIAM Journal on Scientific and Statistical Computing*, 10(2):341–358, 1989.

Adam J Rothman, Elizaveta Levina, and Ji Zhu. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962, 2010.

Alexander Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*, volume 24. Springer Science & Business Media, 2003.

Alessio Spantini, Antti Solonen, Tiangang Cui, James Martin, Luis Tenorio, and Youssef Marzouk. Optimal low-rank approximations of Bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 37(6):A2451–A2487, 2015.

G. Stewart and Ji-Guang Sun. *Matrix Perturbation Theory*. Academic Press, 1 edition, 1990.

Andrew M Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

Maxim Sviridenko, Jan Vondrák, and Justin Ward. Optimal approximation for submodular and supermodular optimization with bounded curvature. *Mathematics of Operations Research*, 42(4):1197–1218, 2017.

Vasileios Tzoumas, Luca Carlone, George J Pappas, and Ali Jadbabaie. Sensing-constrained LQG control. *arXiv preprint arXiv:1709.08826*, 2017.

Jan Vondrak. Optimal approximation for the submodular welfare problem in the value oracle model. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, pages 67–74, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-047-0.

Jan Vondrák, Chandra Chekuri, and Rico Zenklusen. Submodular function maximization via the multilinear relaxation and contention resolution schemes. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, pages 783–792, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0691-1.

Abraham Wald. On the efficient design of statistical investigations. *The Annals of Mathematical Statistics*, 14(2):134–140, 1943.

Laurence A. Wolsey. Maximising real-valued submodular functions: Primal and dual heuristics for location problems. *Mathematics of Operations Research*, 7(3):410–425, 1982.

Laurence A Wolsey and George L Nemhauser. *Integer and Combinatorial Optimization*, volume 55. John Wiley & Sons, 1999.

A. L. Yuille and Anand Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.

T. Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Transactions on Information Theory*, 57(7):4689–4708, July 2011. ISSN 0018-9448.