Two-Sided Weak Submodularity for Matroid Constrained Optimization and Regression

Theophile Thiery   T.F.THIERY@QMUL.AC.UK
School of Mathematical Sciences, Queen Mary University of London, United Kingdom

Justin Ward   JUSTIN.WARD@QMUL.AC.UK
School of Mathematical Sciences, Queen Mary University of London, United Kingdom

Abstract

We study the following problem: Given a variable of interest, we would like to find a best linear predictor for it by choosing a subset of k relevant variables obeying a matroid constraint. This problem is a natural generalization of subset selection problems where it is necessary to spread observations amongst multiple different classes. We derive new, strengthened guarantees for this problem by improving the analysis of the residual random greedy algorithm and by developing a novel distorted local-search algorithm. To quantify our approximation guarantees, we refine the definition of weak submodularity by Das and Kempe (2011) and introduce the notion of an *upper submodularity ratio*, which we connect to the minimum k-sparse eigenvalue of the covariance matrix. More generally, we look at the problem of maximizing a set function f with lower and upper submodularity ratio γ and β under a matroid constraint. For this problem, our algorithms have asymptotic approximation guarantee $\frac{1}{2}$ and $1 - e^{-1}$ as the function is closer to being submodular. As a second application, we show that the Bayesian A-optimal design objective falls into our framework, leading to new guarantees for this problem as well.

1. Introduction

In the subset selection problem for linear regression, we are given a collection $\mathcal{X}$ of predictor variables and a target variable Z, as well as known covariances between each pair of variables. The goal is to find a small collection $\mathcal{S} \subseteq \mathcal{X}$ of at most k predictor variables that gives the best linear predictor for Z. When $|\mathcal{X}|$ is very large, the *forward regression* algorithm is commonly employed as a heuristic. It constructs $\mathcal{S}$ iteratively, at each step adding a variable that greedily maximizes the squared multiple correlation objective. To explain the success of this approach in practice, Das and Kempe (2008) connected subset selection problems with submodular optimization. They showed that the squared multiple correlation objective function, also known as the R^2 objective, is *submodular* in the absence of suppressor variables. Intuitively, a variable $X \in \mathcal{X}$ is a suppressor if there is some other variable $Y \in \mathcal{X}$ such that observing X *increases* the (conditional) correlation between Y and the target variable Z. We give an example of such a situation in Example 1. Even in the presence of suppressors, Das and Kempe (2011, 2018) showed that a weaker property they deemed *weak submodularity* can be used to derive meaningful guarantees. The *submodularity ratio* γ measures how far a function deviates from submodularity when considering the aggregate effect of adding elements. By treating the forward regression algorithm as a variant of the standard greedy algorithm, they showed that it has approximation guarantee of $(1 - e^{-\gamma})$, where $\gamma \in [0, 1]$ can be lower bounded by the smallest $2k$-sparse eigenvalue $\lambda_{\min}(C_{\mathcal{X}}, 2k)$ of the covariance matrix for $\mathcal{X}$.
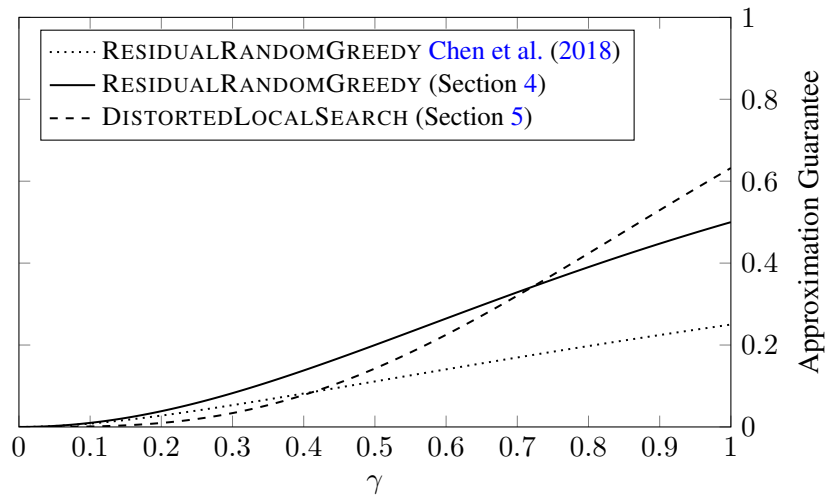
Figure 1: Guarantees for $(\gamma, 1/\gamma)$-weakly submodular maximization under a matroid constraint.

Here, we consider a natural generalization of this problem in which we must select a subset $\mathcal{S}$ that is independent in a general matroid constraint. Such constraints naturally capture settings in which some observations are mutually exclusive (for example, sensors that may be placed in different configurations) or in which it is desirable or necessary to spread observations amongst multiple different classes (for example by time or location). In contrast to cardinality constraints, the best known guarantee for maximizing the R^2 objective in a general matroid is a randomized $1/(1 + \gamma^{-1})^2$-approximation via the RESIDUALRANDOMGREEDY algorithm due to Chen et al. (2018). However, as γ tends to 1 (i.e. as the function becomes closer to a submodular function) this bound tends to only $1/4$, while both RESIDUALRANDOMGREEDY and the standard greedy algorithm are known to provide a $1/2$ approximation for submodular objectives under a matroid constraint. The state of the art in this setting is a $(1 - 1/e)$-approximation (Calinescu et al., 2011; Filmus and Ward, 2014), which is known to be tight (Nemhauser and Wolsey, 1978; Feige, 1998).

A key difficulty is that the definition of weak submodularity considers only the effect of *adding* elements to the current solution. In contrast, the analysis of RESIDUALRANDOMGREEDY as well other state-of-the art procedures for submodular optimization in a matroid requires bounding the losses when elements are *removed* or *swapped* from some solution. Stronger "element-wise" notions of weak submodularity have been proposed that allow the adaptation of such algorithms but in general, these notions may give weaker bounds than those obtained when the submodularity ratio γ can be utilized instead.

1.1. Our results

Here, we give a natural extension of the submodularity ratio γ by considering an *upper submodularity ratio* $\beta > 0$, that bounds how far a function deviates from submodularity when considering the effect of *removing* elements. Intuitively, the parameter β compares the loss by removing an entire set compared to the aggregate individual losses for each element. We show that, as with the submodularity ratio γ, our β can be bounded by spectral quantities in the setting of regression. Specifically, we show that the R^2 objective satisfies $\beta \leq \lambda_{\min}^{-1}(C_{\mathcal{X}}, k) \leq \gamma^{-1}$. Since β is defined in

terms of removing elements from the solution, this requires a different spectral argument than that used to bound γ in Das and Kempe (2018). While their bound for γ follows directly by considering an appropriate Rayleigh quotient, here we must relate the average value of the quadratic forms obtained from the inverses of all rank $k-1$ principle submatrices of a matrix C to that obtained from C^{-1}.

We also consider the problem of Bayesian A-optimal design, which has been previously studied via weak submodularity by Bian et al. (2017); Harshaw et al. (2019); Hashemi et al. (2019). In Appendix C, we show that our parameter β can be bounded by γ^{-1} for this problem, as well.

Using the connection between subset selection problem and submodular maximization, we consider the more general problem of maximizing a (γ, β)-*weakly submodular* set function for which both ratios are bounded. We derive improved guarantees for matroid constrained maximization problems via RESIDUALRANDOMGREEDY. Our guarantee is $\gamma/(\gamma+\beta)$, which approaches $1/2$ as the function f becomes closer to submodular (i.e. as $\gamma, \beta \to 1$). It is natural, then, to ask whether it is possible to obtain an algorithm with guarantee approaching the optimal result of $1 - e^{-1}$ for submodular functions. We answer this affirmatively by giving a local search algorithm guided by a distorted potential, and show it achieves a guarantee approaching $(1 - e^{-1} - \mathcal{O}(\varepsilon))$ for (γ, β)-weakly submodular functions as $\gamma, \beta \to 1$, where $\varepsilon > 0$ is a constant parameter that can be chosen independently of γ, β. Combining these with our spectral bounds, our first analysis improves on the current state of the art for matroid constrained subset selection and A-optimal design problems for all $\gamma > 0$, and our second gives further improvements whenever $\gamma > 0.7217$ (see Figure 1).

Our distorted local-search algorithm builds upon similar techniques from the submodular case presented in Filmus and Ward (2014). There submodularity of f implies submodularity of the potential g, which is used to derive the bounds on g necessary for convergence and sampling, as well the crucial bound linking the local optimality of g to the value of f. Here, however, since f is only approximately submodular, these techniques will not work, and so we require a more delicate analysis for each of these components. A further complication in our setting is that the correct potential g depends on the values of γ and β, which may not be known a priori. We give an approach that is based on guessing the value of a joint parameter in γ and β. Each such guess gives a different distorted potential. Inspired (broadly) by simulated annealing, we show that if each such new potential is initialized by the local optimum of the previous potential, then the overall running time can be amortized over all guesses. We present a simplified version of the algorithm and potential function in Section 5, and defer the more technical details to Appendices A and B.

Finally, given the relationship between γ and β in both problems we consider, it is natural to conjecture that β may be bounded in terms of γ in some generic fashion for *every* weakly submodular function. However, we show that this is not the case by exhibiting (in Appendix D) a function on a ground set of size k for which β must be $\Theta(k^{1-\gamma})$.

1.2. Additional Related Work

The submodularity ratio γ and the corresponding notion of weak submodularity was first introduced to analyze the forward regression and orthogonal matching pursuit algorithms for linear regression by Das and Kempe (2018). It was latter related to restricted strong convexity by Elenberg et al. (2018), leading to similar guarantees for generalized linear model (GLM) likelihood, graphical model learning objectives, or an arbitrary M-estimator. The submodularity ratio has also been applied to the analysis greedy algorithms in other modes of computation (Khanna et al., 2017b;

Elenberg et al., 2017). Together with related algorithmic techniques, it has also lead to algorithms for sensor placement problems (Hashemi et al., 2020), experimental design (Harshaw et al., 2019; Bian et al., 2017), low rank optimization (Khanna et al., 2017a), document summarization (Chen et al., 2018), and interpretation of neural networks (Elenberg et al., 2017)

One motivation for the study of weak submodular functions is to bridge the gap between worst-case theory and the performance of algorithms practice. At the other extreme, one can consider the deviation of a submodular function from linearity. This leads to the notion of *curvature* which can be used to strengthen approximation bounds for both for various submodular optimization problems (Conforti and Cornuéjols, 1984; Vondrák, 2010; Sviridenko et al., 2015; Yoshida, 2018; Friedrich et al., 2019), as well as in combination with weak submodularity (Bian et al., 2017). Inspired by insights from continuous optimization, Pokutta et al. (2020) have recently introduced a new notion of *sharpness*, which provides further explanation for the empirically good performance of the greedy algorithm on submodular objectives.

There have been various approaches based on considering *element-wise* bounds on the deviation of a function from submodularity (Bogunovic et al., 2018; Nong et al., 2019; Gong et al., 2019) including generalizations to functions over the integer lattice (Qian et al., 2018; Kuhnle et al., 2018). These approaches all involve relaxing the notion of decreasing marginal returns by requiring that a function f satisfy $f(A \cup \{e\}) - f(A) \geq \gamma_e \cdot (f(B \cup \{e\}) - f(B))$ for all $e \notin B$ and $A \subseteq B$, where $\gamma_e \in [0, 1]$ has been variously dubbed the *inverse curvature* (Bogunovic et al., 2018), *DR ratio* (Kuhnle et al., 2018), or *generic submodularity ratio* (Nong et al., 2019). For such functions, it is easy to show that our parameters satisfy $\gamma \geq \gamma_e$ and $\beta \leq 1/\gamma_e$. Unfortunately, as we show in Section 3, the resulting inequalities may be very far from tight in our setting. In particular, analyses relying on γ_e may fail to give any non-trivial approximation bounds for regression problems, even in situations when $\lambda_{\min}(C_{\mathcal{X}})$ and the submodularity ratio γ are positive. This observation has motivated our consideration of the more general parameter β, which allows spectral bounds to be utilized.

Finally, we note that all the definitions introduced here assume that the objective is monotone. Recently, Santiago and Yoshida (2020) have proposed a notion of approximate submodularity that extends to the non-monotone case, as well.

2. Preliminaries and Key Definitions

Throughout the remainder of the paper, all set functions $f : 2^X \to \mathbb{R}_{\geq 0}$ that we consider will be *monotone*, satisfying $f(B) \geq f(A)$ for all $A \subseteq B$. Because we are focusing on maximization problems, we will further assume without loss of generality that our objective functions are *normalized*, so $f(\emptyset) = 0$. When there is no risk of confusion, we will use the shorthands $A + e$ for $A \cup \{e\}$ and $A - e$ for $A \setminus \{e\}$. We use the notation $f(e|A) \triangleq f(A + e) - f(A)$ for the marginal gain obtained in f when adding an element $e \notin A$ to A.

The problems that we consider will be constrained by an arbitrary matroid $\mathcal{M} = (X, \mathcal{I})$. Here $\mathcal{I} \subseteq 2^X$ is a family of *independent sets*, satisfying $\emptyset \in \mathcal{I}$, $A \subseteq B \in \mathcal{I} \Rightarrow A \in \mathcal{I}$, and for all $A, B \in \mathcal{I}$ with $|A| < |B|$, there exists $e \in B \setminus A$ such that $A + e \in \mathcal{I}$. The maximal independent sets of $\mathcal{I}$ are called *bases* of $\mathcal{M}$, and the last condition implies that they all have the same size, called the *rank* of $\mathcal{M}$, which we typically denote by k. Our goal will be to find some $S \in \mathcal{I}$ that maximizes the objective f. Since f is monotone, we can assume without loss of generality that S

and the maximizer of f is a base of $\mathcal{M}$. Throughout, we will make use of the following standard result (in fact, this is the only property of matroids that we will use for our analyses):

Proposition 1 *Let $\mathcal{M} = (X, \mathcal{I})$ be a matroid. Then for any pair of bases A, B of $\mathcal{M}$, there exists a bijection $\pi : A \to B$ so that $A - a + \pi(a) \in \mathcal{I}$ for all $a \in A$.*

A set function $f : 2^X \to \mathbb{R}_+$ is *submodular* if and only if $f(e|A) \geq f(e|B)$ for all $A \subseteq B \subseteq X$ and $e \notin B$. It can be shown that this is equivalent to requiring that $f(B) - f(A) \leq \sum_{e \in B \setminus A} f(e|A)$ for any $A \subseteq B \subseteq X$. The definition of *weak-submodularity*[1] relaxes this inequality by requiring:

$$\gamma \cdot (f(B) - f(A)) \leq \sum_{e \in B \setminus A} f(e|A), \tag{1}$$

for any set $A \subseteq B$, where $\gamma \in [0, 1]$ is called the submodularity ratio of f. Note that when $\gamma \geq 1$, (1) iff the function f is submodular. Submodularity can equivalently be characterized by $f(B) - f(A) \geq \sum_{e \in B \setminus A} f(e|B - e)$ for all $A \subseteq B \subseteq X$. Thus, another natural approach is to consider functions that satisfy:

$$\beta \cdot (f(B) - f(A)) \geq \sum_{e \in B \setminus A} f(e|B - e) \tag{2}$$

for some $\beta \geq 1$. We call this property β-*weak submodularity from above* to distinguish it from (1), which we will now refer to as γ-*weak submodularity from below*. Here, $\beta \leq 1$ if and only if f is submodular. Note that by monotonicity, $f(B) - f(A) \geq f(B) - f(B - e) \geq 0$ for any $e \in B \setminus A$ and so every monotone function f satisfies (2) for $\beta = |B|$. We say that a set function f is (γ, β)-*weakly submodular* if it is γ-weakly submodular from below and β-weakly submodular from above (i.e. it satisfies both (1) and (2)).

3. Subset Selection

We now turn to the subset selection problem. Let $\mathcal{M} = (X, \mathcal{I})$ be a matroid. Let Z be a target random variable we wish to predict, and let $\mathcal{X} = \{X_1, \ldots, X_n\}$ be a set of n predictor variables (where here and throughout this section we use calligraphic letters to denote sets of random variables to avoid confusion). We suppose that Z and all X_i have been normalized to have mean 0 and variance 1, and let $C_{\mathcal{X}}$ be the $n \times n$ covariance matrix for the variables X_i. Our goal is to find a set $\mathcal{S} \subseteq \mathcal{X}$, that is independent in some given matroid over $\mathcal{X}$ and gives the best linear predictor for Z. In other words, we want to solve the following optimization problem:

$$\underset{\mathcal{S} \in \mathcal{I}}{\arg\max} \, R_{Z,\mathcal{S}}^2 = \underset{\mathcal{S} \in \mathcal{I}}{\arg\max} (\text{Var}(Z) - \mathbb{E}[(Z - Z_{\mathcal{S}})^2]) / \text{Var}(Z),$$

where R^2 is a measure of fitness of the linear predictor using the *squared multiple correlation*, and $Z_{\mathcal{S}} = \sum_{X_i \in \mathcal{S}} \alpha_i X_i$ is the linear predictor over $\mathcal{S}$ which optimally minimizes the mean square prediction error for Z. The coefficients of this best linear predictor are given by $\boldsymbol{\alpha} = C_{\mathcal{S}}^{-1} \mathbf{b}_{Z,\mathcal{S}}$,

---

1. We note that the definition given here, which is also used in Bian et al. (2017); Chen et al. (2018); Elenberg et al. (2017); Harshaw et al. (2019); Santiago and Yoshida (2020), is slightly adapted from the original definition given in Das and Kempe (2011).

where $C_{\mathcal{S}}$ is the principle submatrix of $C_{\mathcal{X}}$ corresponding to variables in $\mathcal{S}$, and $\mathbf{b}_{Z,\mathcal{S}}$ is a vector of covariances between Z, and each $X_i \in \mathcal{S}$, i.e. $(C_{\mathcal{S}})_{i,j} = \text{Cov}(X_i, X_j)$ and $(\mathbf{b}_{Z,\mathcal{S}})_i = \text{Cov}(X_i, Z)$. Therefore, if we let $X_{\mathcal{S}}$ denote the corresponding vector of random variables in $\mathcal{S}$, the best linear predictor can be written as: $Z_{\mathcal{S}} = X_{\mathcal{S}}^T C_{\mathcal{S}}^{-1} \mathbf{b}_{Z,\mathcal{S}}$. Because Z has unit variance, the objective simplifies to $R^2_{Z,\mathcal{S}} = 1 - \mathbb{E}\big[(Z - Z_{\mathcal{S}})^2\big]$, and so the R^2 objective can be regarded as a measure of the fraction of variance of Z that is explained by $\mathcal{S}$. In addition, we can define the *residual* of Z with respect to this predictor as the random variable $\text{Res}(Z, \mathcal{S}) = Z - Z_{\mathcal{S}} = Z - X_{\mathcal{S}}^T C_{\mathcal{S}}^{-1} \mathbf{b}_{Z,\mathcal{S}}$. Therefore, $R^2_{Z,\mathcal{S}} = 1 - \text{Var}(\text{Res}(Z, \mathcal{S})) = \mathbf{b}_{Z,\mathcal{S}}^T C_{\mathcal{S}}^{-1} \mathbf{b}_{Z,\mathcal{S}}$.

Das and Kempe (2018) show that the R^2 objective satisfies (1) for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{X}$ with $\gamma \geq \lambda_{\min}(C_{\mathcal{X}}, |\mathcal{B}|) \geq \lambda_{\min}(C_{\mathcal{X}})$, where $\lambda_{\min}(C_{\mathcal{X}})$ is the smallest eigenvalue of $C_{\mathcal{X}}$ and $\lambda_{\min}(C_{\mathcal{X}}, |\mathcal{B}|)$ is the smallest $|\mathcal{B}|$-*sparse eigenvalue* of $C_{\mathcal{X}}$. In this section we derive an analogous result.

Theorem 2 *For any $\mathcal{B} \subseteq \mathcal{X}$, the R^2 objective satisfies (2) with $\beta \leq \frac{1}{\lambda_{\min}(C_{\mathcal{X}}, |\mathcal{B}|)} \leq \frac{1}{\lambda_{\min}(C_{\mathcal{X}})}$.*

Combined with existing bounds for γ, it shows that the R^2 objective is $(\gamma, 1/\gamma)$-weakly submodular for $\gamma = \lambda_{\min}(C_{\mathcal{X}})$. To analyze the greedy algorithm, it suffices to let $\mathcal{B} \setminus \mathcal{A}$ and $\mathcal{A}$ both be subsets containing at most k variables, so $|\mathcal{B}| = 2k$, which can lead to tighter bounds on γ. In fact, for both the algorithms we consider in the next two sections, it suffices to consider sets B of size k in (2) and so in practice the tighter bound of $\beta \leq 1/\lambda_{\min}(C_{\mathcal{X}}, k)$ holds.

First, we consider the following small example that illustrates that the *element-wise* bounds on γ_e (inverse curvature, DR ratio, or generic submodularity ratio) are in general not bounded by $\lambda_{\min}(C_{\mathcal{X}})$. In fact, we may have $\gamma_e = 0$ (and so approximation bounds based on γ_e fail) even when $\lambda_{\min}(C_{\mathcal{X}})$ is bounded away from 0.

Example 1 *Let Z, X_1, X_2 be random variables with unit variance and zero mean. Suppose that X_1 is uncorrelated with Z, and $X_2 = (Z + X_1)/\sqrt{2}$. Then, $\text{Cov}(X_1, Z) = 0$ and $\text{Cov}(X_1, X_2) = \text{Cov}(X_2, Z) = 1/\sqrt{2}$. Let $f(\mathcal{S}) = R^2_{Z,\mathcal{S}}$. Then, it can be verified that $f(X_1|\emptyset) = 0$ and $f(X_1|\{X_2\}) = 1/2$. Thus $f(X_1|\emptyset) \geq \gamma_e \cdot f(X_1|\{X_2\})$ is satisfied only for $\gamma_e = 0$. However, $\lambda_{\min}(C_{\mathcal{X}}^{-1}) = 1 - 1/\sqrt{2}$ and, in fact, explicitly computing γ gives $\gamma = 1/2$.*

Next, we turn to the proof of Theorem 2. In order to prove our bounds, we will use the following facts stated in Das and Kempe (2018):

Lemma 3 *Given two sets of random variables $\mathcal{S} = \{X_1, \ldots, X_n\}$, and $\mathcal{A}$, and a random variable Z we have:* $\text{Res}(Z, \mathcal{A} \cup \mathcal{S}) = \text{Res}(\text{Res}(Z, \mathcal{A}), \{\text{Res}(X_i, \mathcal{A})\}_{X_i \in \mathcal{S}})$.

Lemma 4 *Given two sets of random variables $\mathcal{S} = \{X_1, \ldots, X_n\}$, and $\mathcal{A}$, and a random variable Z we have:* $R^2_{Z,\mathcal{A} \cup \mathcal{S}} = R^2_{Z,\mathcal{A}} + R^2_{Z,\{\text{Res}(X_i,\mathcal{A})\}_{X_i \in \mathcal{S}}}$.

We define following quantities, which we use for the rest of the section. Let $\mathcal{A}, \mathcal{B}$ be some fixed sets of random variables with $\mathcal{A} \subseteq \mathcal{B}$. Let $\mathcal{T} = \mathcal{B} \setminus \mathcal{A}$ and suppose without loss of generality that $\mathcal{T} = \{X_1, \ldots, X_t\}$. For each $X_i \in \mathcal{T}$, let $\hat{X}_i = \text{Res}(X_i, \mathcal{A})$ and suppose further that each $\hat{X}_i$ has been renormalized to have unit variance. Let $\hat{\mathcal{T}} = \{\hat{X}_i, \ldots, \hat{X}_t\}$, $\hat{C}$ to be the covariance matrix for $\hat{\mathcal{T}}$, and $\hat{\mathbf{b}}$ to be the vector of covariances between Z and each $\hat{X}_i \in \hat{\mathcal{T}}$. We fix a single random variable X_i. For ease of notation, in the next two Lemmas we assume without loss of generality that $\hat{C}$ and $\hat{\mathbf{b}}$ have been permuted so that X_i corresponds to the last row and column of $\hat{C}$. Then, we define $\mathcal{T}_{-i} = \mathcal{T} \setminus \{X_i\}$, $\hat{\mathcal{T}}_{-i} = \hat{\mathcal{T}} \setminus \{\hat{X}_i\}$, and let $\hat{X}_{-i}$ denote the vector containing the variables

of $\hat{\mathcal{T}}_{-i}$ (ordered as in $\hat{C}$ and $\hat{\mathbf{b}}$). Similarly, let $\hat{C}_{-i}$ be the principle submatrix of $\hat{C}$ obtained by excluding the row and column corresponding to $\hat{X}_i$ (i.e., the last row and column), and $\hat{\mathbf{b}}_{-i}$ be the vector obtained from $\hat{\mathbf{b}}$ by excluding the entry for $\hat{X}_i$ (i.e., the last entry). Finally, we let $\mathbf{u}_i$ be the vector of covariances between $\hat{X}_i$ and each $\hat{X}_j \in \hat{\mathcal{T}}_{-i}$. Note that $\mathbf{u}_i$ corresponds to the last column of $\hat{C}$ with its last entry (corresponding to $\mathrm{Var}(\hat{X}_i)$) removed. We begin by computing the loss in $R^2_{Z,\mathcal{B}}$ when removing X_i from $\mathcal{B}$:

Lemma 5 $R^2_{Z,\mathcal{B}} - R^2_{Z,\mathcal{B}\setminus\{X_i\}} = \mathrm{Cov}(Z, \mathrm{Res}(\hat{X}_i, \hat{\mathcal{T}}_{-i}))^2 / \mathrm{Var}(\mathrm{Res}(\hat{X}_i, \hat{\mathcal{T}}_{-i})) = \hat{\mathbf{b}}^T H_i \hat{\mathbf{b}} / s_i$, where $H_i = \begin{pmatrix} \hat{C}_{-i}^{-1} \mathbf{u}_i \mathbf{u}_i^T \hat{C}_{-i}^{-1} & -\hat{C}_{-i}^{-1} \mathbf{u}_i \\ -\mathbf{u}_i^T \hat{C}_{-i}^{-1} & 1 \end{pmatrix}$ and $s_i = 1 - \mathbf{u}_i^T \hat{C}_{-i}^{-1} \mathbf{u}_i$.

Proof Note that $\mathcal{B} \setminus \{X_i\} = \mathcal{A} \cup \mathcal{T}_{-i}$. Thus, by Lemma 4 and Lemma 3, respectively:

$$R^2_{Z,\mathcal{B}} - R^2_{Z,\mathcal{B}\setminus\{X_i\}} = R^2_{Z,\mathrm{Res}(X_i, \mathcal{A} \cup \mathcal{T}_{-i})} = R^2_{Z,\mathrm{Res}(\mathrm{Res}(X_i, \mathcal{A}), \{\mathrm{Res}(X_j, \mathcal{A})\}_{X_j \in \mathcal{T}_{-i}})}. \tag{3}$$

Recall that each $\hat{X}_j$ is obtained from $\mathrm{Res}(X_j, \mathcal{A})$ by renormalization and that $\hat{\mathcal{T}}_{-i} = \hat{\mathcal{T}} \setminus \{\hat{X}_i\} = \{\hat{X}_j\}_{X_j \in \mathcal{T}_{-i}}$. Thus, $\mathrm{Res}(\hat{X}_i, \hat{\mathcal{T}}_{-i})$ is a rescaling of $\mathrm{Res}(\mathrm{Res}(X_i, \mathcal{A}), \{\mathrm{Res}(X_j, \mathcal{A})\}_{X_j \in \mathcal{T}_{-i}})$. Since the R^2 objective is invariant under scaling of the predictor variables, (3) then implies that

$$R^2_{Z,\mathcal{B}} - R^2_{Z,\mathcal{B}\setminus\{X_i\}} = R^2_{Z,\mathrm{Res}(\hat{X}_i, \hat{\mathcal{T}}_{-i})} = \mathrm{Cov}(Z, \mathrm{Res}(\hat{X}_i, \hat{\mathcal{T}}_{-i}))^2 / \mathrm{Var}(\mathrm{Res}(\hat{X}_i, \hat{\mathcal{T}}_{-i})), \tag{4}$$

where the last line follows directly from the definition of the R^2 objective. It remains to express (4) in terms of $\hat{C}$, $\hat{\mathbf{b}}$ and $\mathbf{u}$. By definition, $\mathrm{Res}(\hat{X}_i, \hat{\mathcal{T}}_{-i}) = \hat{X}_i - \hat{X}_{-i}^T \hat{C}_{-i}^{-1} \mathbf{u}_i$. Hence,

$$\begin{aligned} \mathrm{Var}(\mathrm{Res}(\hat{X}_i, \hat{\mathcal{T}}_{-i})) &= \mathrm{Var}(\hat{X}_i - \hat{X}_{-i}^T \hat{C}_{-i}^{-1} \mathbf{u}_i) \\ &= \mathbb{E}[\hat{X}_i^2] - 2\mathbb{E}[\hat{X}_i \hat{X}_{-i}^T] \hat{C}_{-i}^{-1} \mathbf{u}_i + \mathbf{u}_i^T \hat{C}_{-i}^{-1} \mathbb{E}[\hat{X}_{-i} \hat{X}_{-i}^T] \hat{C}_{-i}^{-1} \mathbf{u}_i = 1 - \mathbf{u}_i^T \hat{C}_{-i}^{-1} \mathbf{u}_i, \end{aligned}$$

where the last equality follows from normalization of $\hat{X}_i$, $\mathbb{E}[\hat{X}_i \hat{X}_{-i}^T] = \mathbf{u}_i^T$ and $\mathbb{E}[\hat{X}_{-i} \hat{X}_{-i}^T] = \hat{C}_{-i}$. Furthermore,

$$\begin{aligned} \mathrm{Cov}(Z, \mathrm{Res}(\hat{X}_i, \hat{\mathcal{T}}_{-i}))^2 &= \mathrm{Cov}(Z, \hat{X}_i - \hat{X}_{-i}^T \hat{C}_{-i}^{-1} \mathbf{u}_i)^2 = \left(\mathrm{Cov}(Z, \hat{X}_i) - \mathrm{Cov}(Z, \hat{X}_{-i}^T \hat{C}_{-i}^{-1} \mathbf{u}_i) \right)^2 \\ &= \left(\hat{b}_i - \hat{\mathbf{b}}_{-i}^T C_{-i}^{-1} \mathbf{u}_i \right)^2 = \hat{\mathbf{b}}^T \begin{pmatrix} \hat{C}_{-i}^{-1} \mathbf{u}_i \mathbf{u}_i^T \hat{C}_{-i}^{-1} & -\hat{C}_{-i}^{-1} \mathbf{u}_i \\ -\mathbf{u}_i^T \hat{C}_{-i}^{-1} & 1 \end{pmatrix} \hat{\mathbf{b}}. \end{aligned}$$

Substituting the above 2 expressions into (4) completes the proof. ∎

In the next lemma we show that the previous lemma can be simplified for eigenvectors of $\hat{C}^{-1}$.

Lemma 6 *Let* $(\lambda, \mathbf{v}), (\mu, \mathbf{w})$ *be any 2 eigenpairs of* $\hat{C}^{-1}$. *Then,* $\mathbf{v}^T H_i \mathbf{w} = \lambda \mu s_i^2 v_i w_i$, *where* H_i *and* s_i *are as defined in the statement of Lemma 5.*

Proof Applying the formula for block matrix inversion (Lemma 21) to $\hat{C}^{-1}$, we have

$$\hat{C}^{-1} = \begin{pmatrix} \hat{C}_{-i} & \mathbf{u}_i \\ \mathbf{u}_i^T & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \hat{C}_{-i}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{1 - \mathbf{u}_i^T \hat{C}_{-i}^{-1} \mathbf{u}_i} \begin{pmatrix} \hat{C}_{-i}^{-1} \mathbf{u}_i \mathbf{u}_i^T \hat{C}_{-i}^{-1} & -\hat{C}_{-i}^{-1} \mathbf{u}_i \\ -\mathbf{u}_i^T \hat{C}_{-i}^{-1} & 1 \end{pmatrix}. \tag{5}$$

Now, because $(\mu, \mathbf{w})$ is an eigenpair of $\hat{C}^{-1}$, we must have $(\hat{C}^{-1}\mathbf{w})_i = \mu w_i$. By (5), this is equivalent to $(-\mathbf{u}_i^T \hat{C}_{-i}^{-1}\mathbf{w}_{-i} + w_i)/s_i = \mu w_i$ (where, as usual, we let $\mathbf{w}_{-i}$ be the vector obtained from $\mathbf{w}$ by discarding its ith entry). Rearranging this equation gives $\mathbf{u}_i^T \hat{C}_{-i}^{-1}\mathbf{w}_{-i} = w_i(1 - \mu s_i)$. Since $\hat{C}^{-1}$ is symmetric, the same argument implies that $\mathbf{v}_{-i}^T \hat{C}_{-i}^{-1}\mathbf{u}_i = v_i(1 - \lambda s_i)$. Thus,

$$\begin{aligned}
\mathbf{v}^T H_i \mathbf{w} &= \mathbf{v}_{-i}^T \hat{C}_{-i}^{-1}\mathbf{u}_i \mathbf{u}_i^T \hat{C}_{-i}^{-1}\mathbf{w}_{-i} - w_i(\mathbf{v}_{-i}^T \hat{C}_{-i}^{-1}\mathbf{u}_i) - v_i(\mathbf{u}_i^T \hat{C}_{-i}^{-1}\mathbf{w}_{-i}) + v_i w_i \\
&= v_i w_i(1 - \lambda s_i)(1 - \mu s_i) - v_i w_i(1 - \lambda s_i) - v_i w_i(1 - \mu s_i) + v_i w_i \\
&= v_i w_i \left((1 - \lambda s_i)(1 - \mu s_i) - (1 - \lambda s_i) - (1 - \mu s_i) + 1\right) = \lambda \mu s_i^2 v_i w_i,
\end{aligned}$$

as claimed. ■

We can now complete the proof of our main result from this section (Theorem 2).

Proof of Theorem 2 Let $\{\mathbf{v}_1, \dots, \mathbf{v}_t\}$ be an eigenbasis of $\hat{C}^{-1}$ with corresponding eigenvalues $\lambda_1, \dots, \lambda_t$. Let V be a matrix with columns given by these $\mathbf{v}_i$. Since $\hat{C}^{-1}$ is a symmetric positive semidefinite matrix, the matrix V is orthonormal. Hence, we can write $\hat{\mathbf{b}} = V\mathbf{y}$ for some vector $\mathbf{y}$. By Lemma 5, $\hat{\mathbf{b}}^T H_i \hat{\mathbf{b}} = \text{Cov}(Z, \text{Res}(\hat{X}_i, \hat{\mathcal{T}}_{-i}))^2 \geq 0$ and $s_i = \text{Var}(\text{Res}(\hat{X}_i, \hat{\mathcal{T}}_{-i})) \leq 1$, for each $i = 1, \dots, t$ and $R_{Z,\mathcal{B}}^2 - R_{Z,\mathcal{B}\backslash\{X_i\}}^2 = \hat{\mathbf{b}}^T H_i \hat{\mathbf{b}}/s_i \leq \hat{\mathbf{b}}^T H_i \hat{\mathbf{b}}/s_i^2 = \mathbf{y}^T V^T H_i V \mathbf{y}/s_i^2$. Finally, by Lemma 6, $(V^T H_i V)_{\ell,m} = \lambda_\ell \lambda_m s_i^2 (\mathbf{v}_\ell)_i (\mathbf{v}_m)_i$. Thus, summing over all i we have:

$$\begin{aligned}
\sum_{i=1}^t R_{Z,\mathcal{B}}^2 - R_{Z,\mathcal{B}\backslash\{X_i\}}^2 &\leq \sum_{i=1}^t \sum_{\ell,m=1}^t (y_\ell y_m \lambda_\ell \lambda_m)(\mathbf{v}_\ell)_i (\mathbf{v}_m)_i \,. \\
&= \sum_{\ell,m=1}^t (y_\ell y_m \lambda_\ell \lambda_m) \sum_{i=1}^t (\mathbf{v}_\ell)_i (\mathbf{v}_m)_i = \sum_{i=1}^t y_i^2 \lambda_i^2 \leq \lambda_{\max}(\hat{C}^{-1}) \sum_{i=1}^t y_i^2 \lambda_i, \quad (6)
\end{aligned}$$

where the last equation follows from the orthonormality of the eigenvectors $\mathbf{v}_i$. Moreover, by Lemma 4

$$R_{Z,\mathcal{B}}^2 - R_{Z,\mathcal{A}}^2 = R_{Z,\hat{\mathcal{T}}}^2 = \hat{\mathbf{b}}^T \hat{C}^{-1}\hat{\mathbf{b}} = \sum_{i=1}^t y_i^2 \lambda_i. \qquad (7)$$

Combining (7) and (6), we have $\sum_{i\in\mathcal{S}} R_{Z,\mathcal{B}}^2 - R_{Z,\mathcal{B}\backslash\{X_i\}}^2 \leq \lambda_{\max}(\hat{C}^{-1})[R_{Z,\mathcal{B}}^2 - R_{Z,\mathcal{A}}^2]$ and so inequality (2) is satisfied for $\beta = \lambda_{\max}(\hat{C}^{-1}) = 1/\lambda_{\min}(\hat{C})$. It remains to bound $1/\lambda_{\min}(\hat{C})$ in terms of the eigenvalues of $C_{\mathcal{X}}$. Recall that $\hat{C}$ is a normalized covariance matrix for the random variables $\{\text{Res}(X_i, \mathcal{B}\backslash\mathcal{A})\}_{X_i\in\mathcal{A}}$. As shown in Das and Kempe (2018) (see Lemma 23 in Appendix E for a formal statement), this implies that $\lambda_{\min}(\hat{C}) \geq \lambda_{\min}(C_{(\mathcal{B}\backslash\mathcal{A})\cup\mathcal{A}}) \geq \lambda_{\min}(C_{\mathcal{X}}, |\mathcal{B}|) \geq \lambda_{\min}(C_{\mathcal{X}})$. The claimed bound on β then follows. ■

4. Improved Analysis of RESIDUALRANDOMGREEDY

In this section, we show that we can derive stronger approximation guarantees for (γ, β)-weakly submodular functions by using the RESIDUALRANDOMGREEDY algorithm considered in Buchbinder et al. (2014); Chen et al. (2018). Combined with the results from the previous section, this gives improved approximation bounds for the subset selection problem with an arbitrary matroid

constraint $\mathcal{M}$. The algorithm RESIDUALRANDOMGREEDY (shown in Algorithm 1) proceeds over k iterations. In iteration i, it greedily extends the current solution S_{i-1} to a base $S_{i-1} \cup M_i$ of $\mathcal{M}$ by selecting a set M_i of the $k - |S_{i-1}| = k - i + 1$ elements with the largest marginal contribution with respect to the S_{i-1}. Then, it chooses an element s_i uniformly at random from M_i which is added to S_{i-1} to obtain a new solution S_i. After k iterations, the final set S_k is returned.

Algorithm 1: RESIDUALRANDOMGREEDY$(\mathcal{M}, X, f)$

$S_0 \leftarrow \emptyset$;
for $i = 1, 2, \ldots, k$ **do**
    $M_i \leftarrow \arg\max\{\sum_{e \in T} f(e \mid S) : T \subseteq X, S \cup T \text{ is a base of } \mathcal{M}\}$;
    $s_i \leftarrow$ an element of M_i chosen uniformly at random;
    $S_i \leftarrow S_{i-1} \cup \{s_i\}$;
return S_k;

Theorem 7 *Suppose that $f : 2^X$ is (α, β)-weakly submodular and $\mathcal{M} = (X, \mathcal{I})$ is a matroid and let $O = \arg\max_{A \in \mathcal{I}} f(A)$. Then, RESIDUALRANDOMGREEDY$(\mathcal{M}, X, f)$ returns a solution $S \in \mathcal{I}$ that satisfies $\mathbb{E}[f(S)] \geq \frac{\gamma}{\gamma + \beta} \cdot f(O)$.*

Proof We begin by introducing some auxiliary sets used in the analysis. For each $i = 0, 1, \cdots, k$, we let O_i to be a subset of O of size $k - i$ such that $S_i \cup O_i$ is a base of $\mathcal{M}$, as follows. Let $O_0 = O$. For each $i \geq 1$, suppose that $S_{i-1} \cup O_{i-1}$ is a base and consider the bijection $\pi_i : S_{i-1} \cup M_i \rightarrow S_{i-1} \cup O_{i-1}$ guaranteed by Proposition 1. We set $O_i = O_{i-1} - \pi_i(s_i)$. Then, $S_i \cup O_i = S_{i-1} \cup O_{i-1} + s_i - \pi_i(s_i)$ is a base, as required. Moreover, note the choice of π_i is independent of the random choice s_i, which implies that $\pi_i(s_i)$ is an element of O_{i-1} chosen uniformly at random. Let $\mathcal{E}$ be the event which fixes the random decisions of the algorithm up to iteration $i - 1$. Conditioned on $\mathcal{E}$, we have:

$$
\mathbb{E}[f(S_i) - f(S_{i-1})] = \frac{1}{|M_i|} \sum_{e \in M_i} f(e|S_{i-1}) = \frac{1}{k - i + 1} \sum_{e \in M_i} f(e|S_{i-1})
$$
$$
\geq \frac{1}{k - i + 1} \sum_{e \in O_{i-1}} f(e|S_{i-1}) \geq \frac{\gamma}{k - i + 1}(f(O_{i-1} \cup S_{i-1}) - f(S_{i-1})). \quad (8)
$$

Here, the third inequality follows the fact that $S_{i-1} \cup O_{i-1}$ is a base and so O_{i-1} is a candidate for M_i. The fourth inequality follows from (1) since f is (γ, β)-weakly submodular. Similarly, (2) together with the fact that $\pi_i(s_i)$ is a uniformly random element of O_{i-1} implies

$$
\frac{1}{k - i + 1}(f(O_{i-1} \cup S_{i-1}) - f(S_{i-1})) \geq \frac{\beta^{-1}}{k - i + 1} \sum_{e \in O_{i-1}} f(e|O_{i-1} \cup S_{i-1} - e),
$$
$$
= \beta^{-1} \cdot \mathbb{E}[f(\pi_i(s_i)|O_{i-1} \cup S_{i-1} - \pi_i(s_i)]. \quad (9)
$$

Algorithm 2: DISTORTEDLOCALSEARCH($\mathcal{M}, X, f$)

Suppose that f is (γ, β)-weakly submodular and let $\phi = \gamma^2 + \beta(1 - \gamma)$;
$A \leftarrow$ an arbitrary base of $\mathcal{M}$;
while $\exists a \in S, b \in X \setminus S$ *with* $S - a + b \in \mathcal{I}$ *and* $g_\phi(A - a + b) > g_\phi(A)$ **do**
$\quad \lfloor \; A \leftarrow S - a + b$;
return A;

We can bound the expected decrease in $f(O_i \cup S_i)$ in iteration i as:

$$
\begin{aligned}
\mathbb{E}&[f(O_i \cup S_i) - f(O_{i-1} \cup S_{i-1})]\\
&= \mathbb{E}[f(O_{i-1} \cup S_{i-1} + s_i - \pi_i(s_i)) - f(O_{i-1} \cup S_{i-1})]\\
&= \mathbb{E}[f(s_i|O_{i-1} \cup S_{i-1} - \pi_i(s_i)) - f(\pi_i(s_i)|O_{i-1} \cup S_{i-1} - \pi_i(s_i))]\\
&\geq -\mathbb{E}[f(\pi_i(s_i)|O_{i-1} \cup S_{i-1} - \pi_i(s_i))],
\end{aligned}
\tag{10}
$$

where the inequality follows by monotonicity of f. Thus

$$
\begin{aligned}
\mathbb{E}[f(S_i) - f(S_{i-1})] &\geq \frac{\gamma}{\beta}\mathbb{E}[f(\pi_i(s_i)|O_{i-1} \cup S_{i-1} - \pi_i(s_i))]\\
&\geq \frac{\gamma}{\beta}\mathbb{E}[f(O_{i-1} \cup S_{i-1}) - f(O_i \cup S_i)],
\end{aligned}
\tag{11}
$$

where the first inequality follows by combining (8) and (9) and the second by (10).

Removing the conditioning on $\mathcal{E}$ and summing the inequalities (11) for $i = 1, \cdots, k$, gives $\mathbb{E}[f(S_k) - f(S_0)] \geq \frac{\gamma}{\beta}\mathbb{E}[f(S_0 \cup O_0) - f(S_k \cup O_k)]$. The claim then follows by observing that $S_0 = \emptyset, S_k = S, O_0 \cup S_0 = O$ and $O_k \cup S_k = S_k$ and so $\left(1 + \frac{\beta}{\gamma}\right)\mathbb{E}[f(S)] \geq f(O)$. ∎

5. Distorted Local Search

Here, we present an algorithm for (γ, β)-weakly submodular functions with a guarantee that smoothly approaches the optimal value of $(1 - 1/e)$ as $\gamma, \beta \to 1$. The algorithm (Algorithm 2), is a local search routine that attempts to swap a single element into the current solution if and only if it improves the following auxiliary potential function parameterized by $\phi \in \mathbb{R}_+$, which we will set appropriately depending on γ and β:

$$
g_\phi(A) = \int_0^1 \frac{\phi e^{\phi p}}{e^\phi - 1} \sum_{B \subseteq A} p^{|B|-1}(1-p)^{|A|-|B|} f(B) \, dp = \sum_{B \subseteq A} m^{(\phi)}_{|A|-1,|B|-1}, f(B)
$$

where we define

$$
m^{(\phi)}_{a,b} \triangleq \int_0^1 \phi e^{\phi p} p^b (1-p)^{a-b}/(e^\phi - 1) \, dp.
$$

In the analysis of Filmus and Ward (2014), it is shown that if f is submodular, its associated potential g is as well, and this plays a crucial role in the analysis. Here, however, f is only *weakly* submodular, which means we must carry out an alternative analysis to bound the quality of a local optimum for g_ϕ. Our analysis will rely on the following properties of the coefficients $m^{(\phi)}_{a,b}$ (see Appendix A.1 for a full proof of each):

Lemma 8 *For any $\phi > 0$, the coefficients $m_{a,b}^{(\phi)}$ satisfy the following:*

1. *$g_\phi(e|A) = \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} f(e|B)$, for any $A \subseteq X$ and $e \notin A$.*

2. *$\sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} = 1$, for all $A \subseteq X$.*

3. *$m_{a,b}^{(\phi)} = m_{a+1,b+1}^{(\phi)} + m_{a+1,b}^{(\phi)}$ for all $0 \leq b \leq a$.*

4. *$\phi m_{a,b}^{(\phi)} = -b m_{a-1,b-1}^{(\phi)} + (a-b) m_{a-1,b}^{(\phi)} + (\phi/(e^{\phi-1}))\mathbf{1}_{b=0} + (\phi e^\phi/(e^\phi - 1))\mathbf{1}_{b=a}$,*
   for all $a > 0$ and $0 \leq b \leq a$.

In order to analyze the performance of Algorithm 2, we consider now two arbitrary bases A and O of the given matroid $\mathcal{M}$. We index the elements $a_i \in A$ and $o_i \in O$ according to the bijection $\pi : A \to O$ guaranteed by Proposition 1 so that $A - a_i + o_i$ is a base for all $1 \leq i \leq |A|$. Our main theorem is the following:

Theorem 9 *Suppose that f is (γ, β)-weakly submodular and let $\phi = \phi(\gamma, \beta) \triangleq \gamma^2 + \beta(1 - \gamma)$. Then, for any bases A, O of a matroid $\mathcal{M}$, $\frac{\phi e^\phi}{e^\phi - 1} f(A) \geq \gamma^2 f(O) + \sum_{i=1}^{|A|} [g_\phi(A) - g_\phi(A - a_i + o_i)]$.*

Note that the base A ultimately returned by Algorithm 2, necessarily has $g_\phi(A) - g_\phi(A - a_i + o_i) \leq 0$ for all i, so Theorem 9 immediately implies that $f(A) \geq \gamma^2 \frac{(1 - e^{-\phi})}{\phi} f(O)$, where $\phi = \gamma^2 + \beta(1 - \gamma)$.

To prove Theorem 9, we first note that $g_\phi(A - a_i + o_i) - g(A) = g(o_i|A - a_i) - g(a_i|A - a_i)$ and so

$$\sum_{i=1}^{|A|} g(a_i|A - a_i) = \sum_{i=1}^{|A|} [g(A) - g(A - a_i + o_i)] + \sum_{i=1}^{|A|} g(o_i|A - a_i). \tag{12}$$

The final term in (12) can now be bounded as follows:

Lemma 10 *Suppose that f is (γ, β)-weakly submodular, and let $A, O \subseteq X$ with $A = \{a_1, \ldots, a_{|A|}\}$ and $O = \{o_1, \ldots, o_{|A|}\}$ (so $|A| = |O|$). Then,*

$$\sum_{i=1}^{|A|} g(o_i|A - a_i) \geq \gamma^2 f(O) - \left(\gamma^2 + \beta(1 - \gamma)\right) \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} f(B).$$

Proof By parts 1 and 3 of Lemma 8, we have

$$g_\phi(o_i|A - a_i) = \sum_{B \subseteq A - a_i} m_{|A|-1,|B|}^{(\phi)} f(o_i|B) = \sum_{B \subseteq A - a_i} [m_{|A|,|B|+1}^{(\phi)} f(o_i|B) + m_{|A|,|B|}^{(\phi)} f(o_i|B)]. \tag{13}$$

Since f is γ-weakly submodular from below

$$f(o_i|B) + f(a_i|B) \geq \gamma f(B \cup \{o_i, a_i\}) - \gamma f(B) = \gamma f(o_i|B + a_i) + \gamma f(a_i|B),$$

and so $f(o_i|B) \geq \gamma f(o_i|B + a_i) - (1 - \gamma)f(a_i|B)$. Thus, the right-hand side of (13) is at least

$$\sum_{B \subseteq A - a_i} m_{|A|,|B|+1}^{(\phi)} [\gamma f(o_i|B + a_i) - (1 - \gamma)f(a_i|B)] + m_{|A|,|B|}^{(\phi)} f(o_i|B) = P + Q, \tag{14}$$

11

where

$$P = \gamma \sum_{B \subseteq A - a_i} \left[m_{|A|,|B|+1}^{(\phi)} f(o_i|B + a_i) + m_{|A|,|B|}^{(\phi)} f(o_i|B) \right] = \gamma \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} f(o_i|B)$$

$$Q = (1 - \gamma) \sum_{B \subseteq A - a_i} \left[m_{|A|,|B|}^{(\phi)} f(o_i|B) - m_{|A|,|B|+1}^{(\phi)} f(a_i|B) \right] \geq -(1 - \gamma) \sum_{B \subseteq A - a_i} m_{|A|,|B|+1}^{(\phi)} f(a_i|B).$$

In the first equation, we have used that for each set $T \subseteq A$, $f(o_i|T)$ appears in the right-hand summation exactly once: if $a_i \in T$ it appears as $T = B + a_i$ with coefficient $m_{|A|,|B|+1}^{(\phi)} = m_{|A|,|T|}^{(\phi)}$ and if $a_i \notin T$ it appears as $T = B$ with coefficient $m_{|A|,|B|}^{(\phi)} = m_{|A|,|T|}^{(\phi)}$. Summing (14) over each $a_i \in A$ we then have

$$\sum_{i=1}^{|A|} g_\phi(o_i|A - a_i) \geq \gamma \sum_{i=1}^{|A|} \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} f(o_i|B) - (1 - \gamma) \sum_{i=1}^{|A|} \sum_{B \subseteq A - a_i} m_{|A|,|B|+1}^{(\phi)} f(a_i|B). \quad (15)$$

Since f is γ-weakly submodular from below and monotone,

$$\gamma \sum_{i=1}^{|A|} \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} f(o_i|B) = \gamma \sum_{B \subseteq A} \sum_{i=1}^{|A|} m_{|A|,|B|}^{(\phi)} f(o_i|B) \geq \gamma^2 \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} [f(O \cup B) - f(B)]$$

$$\geq \gamma^2 \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} [f(O) - f(B)] = \gamma^2 f(O) - \gamma^2 \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} f(B),$$

where the last equation follows from part 2 of Lemma 8. Similarly, since f is β-weakly submodular from above:

$$(1 - \gamma) \sum_{i=1}^{|A|} \sum_{B \subseteq A - a_i} m_{|A|,|B|+1}^{(\phi)} (f(B + a_i) - f(B)) = (1 - \gamma) \sum_{T \subseteq A} \sum_{i=1}^{|A|} m_{|A|,|T|}^{(\phi)} (f(T) - f(T - a_i))$$

$$\leq \beta(1 - \gamma) \sum_{T \subseteq A} m_{|A|,|T|}^{(\phi)} [f(T) - f(\emptyset)] = \beta(1 - \gamma) \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} f(B),$$

where the first equation can be verified by substituting $B = T - a_i$ for each $a_i \in T$ and noting that $|T| = |B| + 1$, and the last equation simply follows from $f(\emptyset) = 0$ and renaming T to B. Using the two previous inequalities to bound the right-hand side of (15), then gives the claimed result. ∎

Proof of Theorem 9 Applying Lemma 10 to the last term in (12) and rearranging gives:

$$\sum_{i=1}^{|A|} g_\phi(a_i|A - a_i) + \left(\gamma^2 + \beta(1 - \gamma)\right) \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} f(B) \geq \gamma^2 f(O) + \sum_{i=1}^{|A|} [g_\phi(A) - g_\phi(A - a_i + o_i)].$$

$$(16)$$

From part 1 of Lemma 8,

$$\sum_{i=1}^{|A|} g_\phi(a_i|A - a_i) = \sum_{i=1}^{|A|} \sum_{B \subseteq A - a_i} m_{|A|-1,|B|}^{(\phi)} (f(B + a_i) - f(B))$$

$$= \sum_{T \subseteq A} |T| m_{|A|-1,|T|-1}^{(\phi)} f(T) - (|A| - |T|) m_{|A|-1,|T|}^{(\phi)} f(T),$$

TWO-SIDED WEAK SUBMODULARITY FOR MATROID CONSTRAINED OPTIMIZATION AND REGRESSION

where the last equation follows from the fact that each $T \subseteq A$ appears once as $T = B + a_i$ for each $a_i \in T$ (in which case it has coefficient $m_{|A|-1,|B|}^{(\phi)} = m_{|A|-1,|T|-1}^{(\phi)}$) and once as $T = B$ for each $a_i \notin T$ (in which case it has coefficient $m_{|A|-1,|B|}^{(\phi)} = m_{|A|-1,|T|}^{(\phi)}$). Thus, we can rewrite (16) as:

$$\sum_{B \subseteq A} \left(|B| m_{|A|-1,|B|-1}^{(\phi)} - (|A|-|B|) m_{|A|-1,|B|}^{(\phi)} + \left(\gamma^2 + \beta(1-\gamma)\right) m_{|A|,|B|}^{(\phi)} \right) f(B)$$

$$\geq \gamma^2 f(O) + \sum_{i=1}^{|A|} [g_\phi(A) - g_\phi(A - a_i + o_i)]. \quad (17)$$

Since $\phi = \gamma^2 + \beta(1-\gamma)$, the recurrence in part 4 of Lemma 8 implies that the left-hand side vanishes for all B except $B = \emptyset$, in which case it is $\frac{\phi}{e^\phi - 1} f(\emptyset) = 0$ or $B = A$, in which case it is $\frac{\phi e^\phi}{e^\phi - 1} f(A)$. The theorem then follows. ∎

There are several further issues that must be addressed in order to convert Algorithm 2 to a general, polynomial-time algorithm. First, we cannot compute $g_\phi(A)$ directly, as it depends on the values $f(A)$ for all subsets of A. In Appendix A.4 we show that we can efficiently estimate g_ϕ via simple sampling procedure. To bound the number of improvements made, we can instead require that each improvement makes a $(1 + \varepsilon)$ increase in g_ϕ. Then at termination, we will instead have $\sum_{i=1}^{|A|} g_\phi(A) - g_\phi(A - a_i + o_i)] \leq |A| \varepsilon g_\phi(A)$. In order to bound the resulting loss in our guarantee we must bound the value $g_\phi(A)$ in terms of $f(A)$, which we accomplish in Appendix A.2. Finally, we address the fact that γ and β may not be known and so we cannot set ϕ a priori. We show that by initializing the algorithm with a solution produced by RESIDUALRANDOMGREEDY, we can bound the range of values for ϕ that must be considered to obtain our guarantee. It then suffices to enumerate guesses for ϕ from this range. In Appendix A.3 we show that small changes in ϕ result in small changes to $g_\phi(A)$, and so by initializing the run for each subsequent guess of ϕ with the solution produced for the previous guess, we can amortize the total number of improvements (and work) required across all guesses. The final algorithm, presented in Appendix B, has the same guarantee as Algorithm 2 minus a small $\mathcal{O}(\varepsilon)$ term, and requires $\tilde{\mathcal{O}}(nk^4\varepsilon^{-3})$ evaluations of f. Thus, we have the following:

Theorem 11 *Let $\mathcal{M} = (X, \mathcal{I})$ be a matroid, $f : 2^X \to \mathbb{R}_+$ be a (γ, β)-weakly submodular function, and $\varepsilon > 0$. Then, there is a randomized algorithm that with probability $1 - o(1)$ returns a set S satisfying $f(S) \geq \left(\frac{\gamma^2(1 - e^{-\phi(\gamma,\beta)})}{\phi(\gamma,\beta)} - \mathcal{O}(\varepsilon) \right) f(O)$ for any solution $O \in \mathcal{I}$, where $\phi(\gamma, \beta) = \gamma^2 + \beta(1-\gamma)$. The algorithm runs in time $\tilde{\mathcal{O}}(nk^4\varepsilon^{-3})$.*

In both the subset selection problem and Bayesian A-optimal design (considered in Appendix C), we can derive an upper bound for β matching existing spectral bounds on γ^{-1}. Then we have $\phi(\gamma, \beta) = \gamma^2 + \frac{1}{\gamma} - 1$ and so for both problems we obtain a guarantee of $\gamma^2 \cdot \frac{1 - e^{-(\gamma^2 + \gamma^{-1} - 1)}}{\gamma^2 + \gamma^{-1} - 1} - \mathcal{O}(\varepsilon)$. In particular, as γ tends to 1 (and so f becomes closer to submodular) our guarantee approaches $1 - e^{-1} - \mathcal{O}(\varepsilon)$, matching (up to $\mathcal{O}(\varepsilon)$) the optimal guarantee for submodular functions.

Surprisingly, we show that this relationship between γ^{-1} and β does not hold in general. In Appendix D, we prove the following:

13

Theorem 12 *For any $\gamma > 0$ and $k > 0$ there exists a function on a ground set of size k that is γ-weakly submodular from below but not β-weakly submodular from above for any $\beta < \binom{k-\gamma}{k-1} = \Theta(k^{1-\gamma})$.*

Note that as $\gamma \to 0$ (respectively $\gamma \to 1$), we have $\beta \to k$ (respectively $\beta \to 1$). Hence, we recover the trivial upper bound for monotone set functions when $\gamma \to 0$ and as $\gamma \to 1$, our lower bound on β approaches 1, corresponding to submodularity. We conjecture that this bound on β is in fact the tightest achievable.

Acknowledgments

References

Noga Alon and Joel H. Spencer. *The Probabilistic Method*. Wiley Publishing, 4th edition, 2016.

Andrew An Bian, Joachim M Buhmann, Andreas Krause, and Sebastian Tschiatschek. Guarantees for greedy maximization of non-submodular functions with applications. In *Proc. 34th ICML*, pages 498–507, 2017.

Ilija Bogunovic, Junyao Zhao, and Volkan Cevher. Robust maximization of non-submodular objectives. In *Proc. 21st AISTATS*, pages 890–899, 2018.

Niv Buchbinder, Moran Feldman, Joseph Naor, and Roy Schwartz. Submodular maximization with cardinality constraints. In *Proc. 25th SODA*, pages 1433–1452, 2014.

Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM J. Computing*, 40(6):1740–1766, 2011.

Lin Chen, Moran Feldman, and Amin Karbasi. Weakly submodular maximization beyond cardinality constraints: Does randomization help greedy? In *Proc. 35th ICML*, pages 804–813, 2018.

Flavio Chierichetti, Anirban Dasgupta, and Ravi Kumar. On additive approximate submodularity. *arXiv preprint arXiv:2010.02912*, 2020.

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.

Abhimanyu Das and David Kempe. Algorithms for subset selection in linear regression. In *Proc. 40th STOC*, pages 45–54, 2008.

Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proc. 28th ICML*, 2011.

Abhimanyu Das and David Kempe. Approximate submodularity and its applications: Subset selection, sparse approximation and dictionary selection. *J. Mach. Learn. Res.*, 19:3:1–3:34, 2018.

Ethan R Elenberg, Alexandros G Dimakis, Moran Feldman, and Amin Karbasi. Streaming weak submodularity: Interpreting neural networks on the fly. In *Proc. 31st NeurIPS*, pages 4044–4054, 2017.

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

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

Yuval Filmus and Justin Ward. A tight combinatorial algorithm for submodular maximization subject to a matroid constraint. *SIAM J. Computing*, 43(2):514–542, 2014.

M. L. Fisher, G. L. Nemhauser, and L. A. Wolsey. *An analysis of approximations for maximizing submodular set functions—II*, pages 73–87. Springer Berlin Heidelberg, 1978.

Tobias Friedrich, Andreas Göbel, Frank Neumann, Francesco Quinzan, and Ralf Rothenberger. Greedy maximization of functions with bounded curvature under partition matroid constraints. In *Proc. of AAAI*, volume 33(1), pages 2272–2279, 2019.

Suning Gong, Qingqin Nong, Wenjing Liu, and Qizhi Fang. Parametric monotone function maximization with matroid constraints. *J. Global Optimization*, 75(3):833–849, 2019.

Chris Harshaw, Moran Feldman, Justin Ward, and Amin Karbasi. Submodular maximization beyond non-negativity: Guarantees, fast algorithms, and applications. In *Proc. 36th ICML*, volume 97, pages 2634–2643, 2019.

Abolfazl Hashemi, Mahsa Ghasemi, Haris Vikalo, and Ufuk Topcu. Submodular observation selection and information gathering for quadratic models. In *Proc. 36th ICML*, pages 2653–2662, 2019.

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

Thibaut Horel and Yaron Singer. Maximization of approximately submodular functions. *Proc. 29th NeurIPS*, pages 3045–3053, 2016.

Rajiv Khanna, Ethan Elenberg, Alexandros G. Dimakis, and Sahand Negahban. On approximation guarantees for greedy low rank optimization. In *Proc. 34th ICML*, pages 1837–1846, 2017a.

Rajiv Khanna, Ethan Elenberg, Alexandros G Dimakis, Sahand Negahban, and Joydeep Ghosh. Scalable greedy feature selection via weak submodularity. In *Proc. 20th AISTATS*, pages 1560–1568, 2017b.

Andreas Krause, Ajit Paul Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *J. Machine Learning Research*, 9: 235–284, 2008.

Alan Kuhnle, J. David Smith, Victoria G. Crawford, and My T. Thai. Fast maximization of non-submodular, monotonic functions on the integer lattice. In *Proc. 35th ICML*, pages 2791–2800, 2018.

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.

Qingqin Nong, Tao Sun, Suning Gong, Qizhi Fang, Dingzhu Du, and Xiaoyu Shao. Maximize a monotone function with a generic submodularity ratio. In *Proc. International Conference on Algorithmic Applications in Management*, pages 249–260, 2019.

Sebastian Pokutta, Mohit Singh, and Alfredo Torrico. On the unreasonable effectiveness of the greedy algorithm: Greedy adapts to sharpness. In *Proc. 37th ICML*, pages 7772–7782, 2020.

Chao Qian, Yibo Zhang, Ke Tang, and Xin Yao. On multiset selection with size constraints. In *Proc. 32nd AAAI*, pages 1395–1402, 2018.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005. ISBN 026218253X.

Richard Santiago and Yuichi Yoshida. Weakly submodular function maximization using local submodularity ratio. *arXiv preprint arXiv:2004.14650*, 2020.

Maxim Sviridenko, Jan Vondrák, and Justin Ward. Optimal approximation for submodular and supermodular optimization with bounded curvature. In *Proc. 26th SODA*, pages 1134–1148, 2015.

Jan Vondrák. Submodularity and curvature: the optimal algorithm. *RIMS Kôkyûroku Bessatsu*, B23: 253–266, 01 2010.

Yuichi Yoshida. Maximizing a monotone submodular function with a bounded curvature under a knapsack constraint. *SIAM J. Discrete Mathematics*, 33(3):1452–1471, 2018.

Appendix A. Properties of g_ϕ

Here we give further properties of the potential

$$g_\phi(A) = \int_0^1 \frac{\phi e^{\phi p}}{e^\phi - 1} \sum_{B \subseteq A} p^{|B|-1}(1-p)^{|A|-|B|} f(B) \, dp = \sum_{B \subseteq A} m^{(\phi)}_{|A|-1,|B|-1} f(B)$$

defined in Section 5. We recall that the coefficients $m^{(\phi)}_{a,b}$ for $0 \le b \le a$ are defined by

$$m^{(\phi)}_{a,b} = \int_0^1 \frac{\phi e^{\phi p}}{e^\phi - 1} p^b (1-p)^{a-b} \, dp.$$

If we consider a continuous distribution $\mathcal{D}_\phi$ on $[0,1]$ with density function:

$$\mathcal{D}_\phi(x) = \frac{\phi e^{\phi x}}{e^\phi - 1}$$

then we can succinctly express these coefficients as $m^{(\phi)}_{a,b} = \mathbb{E}_{p \sim \mathcal{D}_\phi}[p^a (1-p)^b]$. For convenience, we will define $m^{(\phi)}_{a,b} = 0$ if either $a < 0$ or $b < 0$.

Here, and in the following sections, we define the function $h : \mathbb{R} \to \mathbb{R}$ by $h(x) = \frac{xe^x}{e^x-1}$. Then, by Bernoulli's inequality $\frac{dh}{dx} = \frac{e^x(e^x-1-x)}{(e^x-1)^2} \geq 0$ and so h is an increasing function. Note that our algorithm's claimed guarantee can then be expressed as $\frac{\gamma^2}{h(\phi(\gamma,\beta))}$ where $\phi(\gamma,\beta) = \gamma^2 + \beta(1-\gamma)$. Finally, for $k \in \mathbb{Z}_+$, we let H_k denote the kth harmonic number $H_k = \sum_{i=1}^k 1/i = \Theta(\log k)$.

A.1. Properties of the coefficients $m_{a,b}^{(\phi)}$

Here we provide a proof for the following properties $m_{a,b}^{(\phi)}$ used in Section 5, which we restate here for convenience. We give a proof of each claim in turn.

Lemma 8 *For any $\phi > 0$, the coefficients $m_{a,b}^{(\phi)}$ satisfy the following:*

1. $g_\phi(e|A) = \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} f(e|B)$, *for any $A \subseteq X$ and $e \notin A$.*

2. $\sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} = 1$, *for all $A \subseteq X$.*

3. $m_{a,b}^{(\phi)} = m_{a+1,b+1}^{(\phi)} + m_{a+1,b}^{(\phi)}$ *for all $0 \leq b \leq a$.*

4. $\phi m_{a,b}^{(\phi)} = -b m_{a-1,b-1}^{(\phi)} + (a-b)m_{a-1,b}^{(\phi)} + (\phi/(e^{\phi-1}))\mathbf{1}_{b=0} + (\phi e^\phi/(e^\phi-1))\mathbf{1}_{b=a}$, *for all $a > 0$ and $0 \leq b \leq a$.*

Proof of Claim 1 Note that by the definition of g_ϕ:

$$g_\phi(e|A) = \sum_{B \subseteq A+e} m_{|A|,|B|-1}^{(\phi)} f(B) - \sum_{B \subseteq A} m_{|A|-1,|B|-1}^{(\phi)} f(B)$$

$$= \sum_{B \subseteq A} \left[(m_{|A|,|B|-1}^{(\phi)} - m_{|A|-1,|B|-1}^{(\phi)}) f(B) + m_{|A|,|B|}^{(\phi)} f(B+e) \right].$$

It thus suffices to show $(m_{|A|,|B|-1}^{(\phi)} - m_{|A|-1,|B|-1}^{(\phi)}) f(B) = -m_{|A|,|B|}^{(\phi)} f(B)$. For $B = \emptyset$, we have $f(\emptyset) = 0$ and so $(m_{|A|,-1}^{(\phi)} - m_{|A|-1,-1}^{(\phi)}) f(\emptyset) = 0 = -m_{|A|,0}^{(\phi)} f(\emptyset)$. When $|B| \geq 1$,

$$m_{|A|,|B|-1}^{(\phi)} - m_{|A|-1,|B|-1}^{(\phi)} = \mathop{\mathbb{E}}_{p \sim \mathcal{D}_\phi} \left[p^{|B|-1}(1-p)^{|A|-|B|+1} - p^{|B|-1}(1-p)^{|A|-|B|} \right]$$

$$= \mathop{\mathbb{E}}_{p \sim \mathcal{D}_\phi} \left[-p^{|B|}(1-p)^{|A|-|B|} \right] = -m_{|A|,|B|}^{(\phi)}.$$

∎

Proof of Claim 2 By linearity of expectation:

$$\sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} = \sum_{b=0}^{|A|} \binom{|A|}{b} \mathop{\mathbb{E}}_{p \sim \mathcal{D}_\phi} \left[p^b(1-p)^{|A|-b} \right] = \mathop{\mathbb{E}}_{p \sim \mathcal{D}_\phi} \left[\sum_{b=0}^{|A|} \binom{|A|}{b} p^b(1-p)^{|A|-b} \right] = 1.$$

∎

17

Proof of Claim 3 When $0 \le b \le a$, the definition of $m_{a,b}^{(\phi)}$ immediately gives:

$$m_{a,b}^{(\phi)} = \mathop{\mathbb{E}}_{p \sim \mathcal{D}_\phi} p^b (1-p)^{a-b} = \mathop{\mathbb{E}}_{p \sim \mathcal{D}_\phi} [p^b(1-p)^{a-b}p + p^b(1-p)^{a-b}(1-p)] = m_{a+1,b+1}^{(\phi)} + m_{a+1,b}^{(\phi)}.$$

∎

Proof of Claim 4 For $a > 0$ and $b \le a$, noting that $\mathcal{D}_\phi(p) = \frac{d}{dp}\frac{\mathcal{D}_\phi(p)}{\phi}$ and applying integration by parts

$$\begin{aligned}
m_{a,b}^{(\phi)} &= \int_0^1 \mathcal{D}_\phi(p) \cdot p^b (1-p)^{a-b}\, dp \\
&= \frac{\mathcal{D}_\phi(p)}{\phi} p^b (1-p)^{a-b} \Big|_{p=0}^{p=1} - \int_0^1 \frac{\mathcal{D}_\phi(p)}{\phi} \left(b p^{b-1}(1-p)^{a-b} - (a-b)p^b(1-p)^{a-b-1} \right)\, dp\,.
\end{aligned}$$

Which is equivalent to:

$$\phi m_{a,b}^{(\phi)} = -b m_{a-1,b-1}^{(\phi)} + (a-b)m_{a-1,b}^{(\phi)} + \mathcal{D}_\phi(p)p^b(1-p)^{a-b}\Big|_{p=0}^{p=1}.$$

This follows immediately from the definition of $m_{a,b}^{(\phi)}$ when $b > 0$, and when $b = 0$ it follows from $-b m_{a-1,b-1}^{(\phi)} = 0 = b p^{b-1}(1-p)^{a-b}$.

To complete the claim, we note that $\lim_{p \to 0^+} \mathcal{D}_\phi(p)p^b(1-p)^{a-b}$ is $\mathcal{D}_\phi(0) = \phi/(e^\phi - 1)$ if $b = 0$ and 0 if $b > 0$, and $\lim_{p \to 1^-} \mathcal{D}_\phi(p)p^b(1-p)^{a-b}$ is $\mathcal{D}_\phi(1) = \phi e^\phi/(e^\phi - 1)$ if $a = b$, and 0 if $0 \le b < a$. ∎

A.2. Bounding the value of g

Here we show that the value of $g_\phi(A)$ can be bounded in terms of $f(A)$ for any set A. In the analysis of Filmus and Ward (2014), this follows from submodularity of g, which is inherited from the submodularity of f. Here, we must again adopt a different approach. We begin by proving the following claim. Fix some set $A \subseteq X$ and for all $0 \le j \le |A|$ define $F_j = \sum_{B \in \binom{A}{j}} f(B)$ as the total value of all subsets of A of size j. Note that since we suppose f is normalized, $F_0 = f(\emptyset) = 0$.

Lemma 13 *If f is γ-weakly submodular from below, then $F_i \ge \binom{|A|-1}{i-1}\gamma f(A)$ for all $1 \le i \le |A|$.*

Proof Let $k = |A|$. Since f is γ-weakly submodular, for any $B \subseteq A$ we have

$$\sum_{e \in A \setminus B} (f(B+e) - f(B)) \ge \gamma(f(A) - f(B)).$$

Rearranging this, we have

$$\sum_{e \in A \setminus B} f(B+e) \ge \gamma f(A) + (|A| - |B| - \gamma)f(B) \ge \gamma f(A) + (|A| - |B| - 1)f(B)\,, \qquad (18)$$

for all $B \subseteq A$. Summing (18) over all $\binom{k}{j}$ possible subsets B of size j, we obtain

$$(j+1)F_{j+1} \geq \gamma \binom{k}{j}f(A) + (k-j-1)F_j, \tag{19}$$

since each set T of size $j+1$ appears once as $B+e$ on the left-hand side of (18) for each of the $j+1$ distinct choices of $e \in T$ with $B = T - e$.

We now show that $F_i \geq \binom{k-1}{i-1}\gamma f(A)$ for all $1 \leq i \leq k$. The proof is by induction on i. For $i = 1$, the claim follows immediately from (19) with $j = 0$, since then $\binom{k}{j} = 1 = \binom{k-1}{i-1}$ and $(k-j-1)F_j = (k-1)F_0 = 0$. For the induction step, (19) and the induction hypothesis imply:

$$F_{i+1} \geq \tfrac{1}{i+1}\left(\gamma\binom{k}{i}f(A) + (k-i-1)F_i\right) \geq \tfrac{1}{i+1}\left(\gamma\binom{k}{i}f(A) + (k-i-1)\gamma\binom{k-1}{i-1}f(A)\right)$$

$$= \tfrac{\gamma}{i+1}\left(\tfrac{k}{i}\binom{k-1}{i-1}f(A) + (k-i-1)\binom{k-1}{i-1}f(A)\right) = \tfrac{\gamma}{i+1}\left(\tfrac{k+ki-i^2-i}{i}\right)\binom{k-1}{i-1}f(A)$$

$$= \tfrac{\gamma}{i+1}\tfrac{k(i+1)-i(i+1)}{i}\binom{k-1}{i-1}f(A) = \gamma\tfrac{k-i}{i}\binom{k-1}{i-1}f(A) = \gamma\binom{k-1}{i}f(A).$$

∎

Using the above claim, we now bound the value of $g_\phi(A)$ for any set A.

Lemma 14 *If f is γ-weakly submodular, then for all $A \subseteq X$, $\gamma f(A) \leq g_\phi(A) \leq h(\phi)H_{|A|}f(A)$.*

Proof Let $k = |A|$. We begin with the lower bound for $g_\phi(A)$. By the definition of the coefficients $m_{a,b}^{(\phi)}$ and Lemma 13:

$$g_\phi(A) = \sum_{B \subseteq A} \mathop{\mathbb{E}}_{p \sim \mathcal{D}_\phi}\left[p^{|B|-1}(1-p)^{|A|-|B|}\right]f(B) = \mathop{\mathbb{E}}_{p \sim \mathcal{D}_\phi}\left[\sum_{i=1}^{k}p^{i-1}(1-p)^{k-i}F_i\right]$$

$$\geq \mathop{\mathbb{E}}_{p \sim \mathcal{D}_\phi}\left[\sum_{i=1}^{k}\gamma p^{i-1}(1-p)^{k-i}\binom{k-1}{i-1}f(A)\right] = \mathop{\mathbb{E}}_{p \sim \mathcal{D}_\phi}\left[\gamma f(A)\sum_{i=0}^{k-1}\binom{k-1}{i}p^i(1-p)^{k-i-1}\right]$$

$$= \mathop{\mathbb{E}}_{p \sim \mathcal{D}_\phi}\left[\gamma f(A)\right] = \gamma f(A).$$

For the upper bound, we similarly have:

$$g_\phi(A) = \mathop{\mathbb{E}}_{p \sim \mathcal{D}_\phi}\left[\sum_{i=1}^{k}p^{i-1}(1-p)^{k-i}F_i\right] \leq \mathop{\mathbb{E}}_{p \sim \mathcal{D}_\phi}\left[\sum_{i=1}^{k}p^{i-1}(1-p)^{k-i}\binom{k}{i}f(A)\right]$$

$$= \int_0^1 \frac{\phi e^{\phi p}}{e^\phi - 1}\frac{\sum_{i=1}^{k}\binom{k}{i}p^i(1-p)^{k-i}f(A)}{p}\,dp = \int_0^1 \frac{\phi e^{\phi p}}{e^\phi - 1}\frac{1-(1-p)^k}{p}f(A)\,dp$$

$$\leq \frac{\phi e^\phi}{e^\phi - 1}\int_0^1\frac{1-(1-p)^k}{p}f(A)\,dp = h(\phi)\int_0^1\sum_{j=0}^{k-1}(1-p)^j f(A)\,dp$$

$$= h(\phi)\sum_{j=0}^{k-1}\frac{1}{j+1}f(A) = h(\phi)H_{|A|}f(A),$$

where the first inequality follows from monotonicity of f and the second inequality from $\frac{1-(1-p)^k}{p} > 0$ for $p \in (0,1]$ and $h(\phi) = \frac{\phi e^\phi}{e^\phi - 1}$ is an increasing function of p. ∎

A.3. Bounding the sensitivity of g_ϕ to ϕ

The following lemma shows that small changes in the parameter ϕ produce relatively small changes in the value $g_\phi(A)$ for any set A.

Lemma 15 *For all ϕ, $\varepsilon > 0$, and $S \subseteq X$,*

1. $g_{\phi(1-\varepsilon)}(S) \geq e^{-\phi\varepsilon} g_\phi(S)$

2. $h(\phi) \leq e^{\phi\varepsilon} h(\phi(1-\varepsilon))$

Proof Both claims will follow from the inequality

$$\frac{\phi(1-\varepsilon)e^{\phi(1-\varepsilon)p}}{e^{\phi(1-\varepsilon)} - 1} \geq e^{-\phi\varepsilon} \frac{\phi e^{\phi p}}{e^\phi - 1}, \tag{20}$$

which we show is valid for all $p \in [0,1]$ and $\varepsilon > 0$. Indeed, under these assumptions,

$$\frac{\phi(1-\varepsilon)e^{\phi(1-\varepsilon)p}}{e^{\phi(1-\varepsilon)} - 1} \cdot \frac{e^\phi - 1}{\phi e^{\phi p}} = (1-\varepsilon)e^{-\phi\varepsilon p} \frac{e^\phi - 1}{e^\phi e^{-\phi\varepsilon} - 1} = (1-\varepsilon)e^{-\phi\varepsilon p} \frac{e^\phi - 1}{e^\phi(1 + (e^{-\phi} - 1))^\varepsilon - 1}$$

$$\geq (1-\varepsilon)e^{-\phi\varepsilon p} \frac{e^\phi - 1}{e^\phi(1 + \varepsilon(e^{-\phi} - 1)) - 1} = (1-\varepsilon)e^{-\phi\varepsilon p} \frac{e^\phi - 1}{(1-\varepsilon)(e^\phi - 1)} = e^{-\phi\varepsilon p} \geq e^{-\phi\varepsilon}.$$

Here the first inequality follows from the generalized Bernoulli inequality $(1 + x)^t \leq (1 + tx)$, which holds for all $x \geq -1$ and $0 \leq t \leq 1$, and the second inequality follows from $p \in [0,1]$.

For the first claim, applying (20) gives

$$g_{\phi(1-\varepsilon)}(A) = \int_0^1 \frac{\phi(1-\varepsilon)e^{\phi(1-\varepsilon)p}}{e^{\phi(1-\varepsilon)p} - 1} \sum_{B \subseteq A} p^{|B|-1}(1-p)^{|A|-|B|} f(B)\, dp$$

$$\geq \int_0^1 e^{-\phi\varepsilon} \frac{\phi e^{\phi p}}{e^\phi - 1} \sum_{B \subseteq A} p^{|B|-1}(1-p)^{|A|-|B|} f(B)\, dp = e^{-\phi\varepsilon} g_\phi(A),$$

as required. For the second claim, setting $p = 1$ in (20) gives $h(\phi(1-\varepsilon)) \geq e^{-\phi\varepsilon} h(\phi)$ or, equivalently, $h(\phi) \leq e^{\phi\varepsilon} h(\phi(1-\varepsilon))$. ∎

A.4. Efficiently estimating g_ϕ via sampling

The definition of g_ϕ requires evaluating $f(B)$ on all $B \subseteq A$, which requires $2^{|A|}$ calls to the value oracle for f. In this section, we show that we can efficiently estimate g_ϕ using only a polynomial number of value queries to f. Our sampling procedure is based on the same general ideas described in Filmus and Ward (2014), but here we focus on evaluating only the *marginals* of g_ϕ, which results in a considerably simpler implementation. In particular, our algorithm does not require computation of the coefficients $m_{a,b}^{(\phi)}$.

Lemma 16 *For any ϕ, N, there is a randomized procedure for obtaining an estimate $\tilde{g}(e|A)$ of $g_\phi(e|A)$ using N queries to the value oracle for f so that for any $\delta > 0$,*

$$\mathbb{P}[\, |g(e|A) - \tilde{g}(e|A)| \geq \delta f(A + e)\,] < 2e^{-\frac{\delta^2 N}{2}},$$

Proof We consider the following 2-step procedure given as an interpretation of g in Filmus and Ward (2014): we first sample $p \sim \mathcal{D}_\phi$, then construct a random $B \subseteq A$ by taking each element of A independently with probability p. The probability that any given $B \subseteq A$ is selected by the procedure is then precisely

$$\int_0^1 \frac{\phi e^{\phi p}}{e^\phi - 1} p^{|B|} (1-p)^{|A|} \, dp = m_{|A|,|B|}^{(\phi)}.$$

Thus, for a random $\tilde{B} \subseteq A$ sampled in this fashion, $\mathbb{E}[f(e|\tilde{B})] = \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} f(e|B) = g(e|A)$, by part 1 of Lemma 8. We remark that for the particular distributions $\mathcal{D}_\phi$ we consider, the first step of the procedure can easily by implemented with inverse transform sampling.

Suppose now that we draw N independent random samples $\{B_i\}_{i=1}^N$ in this fashion and define the random variables $Y_i = \frac{g(e|A) - f(e|B_i)}{f(A+e)}$. Then, $\mathbb{E}[Y_i] = 0$ for all i. Moreover, by monotonicity of f, $0 \leq f(e|B) \leq f(B+e) \leq f(A+e)$ for all $B \subseteq A$ and also $0 \leq \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} f(e|B) = g(e|A)$ and $g(e|A) = \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} f(e|B) \leq \sum_{B \subseteq A} m_{|A|,|B|}^{(\phi)} f(A+e) = f(A+e)$ by part 2 of Lemma 8. Thus, $|Y_i| \leq 1$ for all i. Let $\tilde{g}_\phi(e|A) = \frac{1}{N} \sum_{i=1}^N f(e|B_i)$. Applying the Chernoff bound (Lemma 24), for any $\delta > 0$ we have

$$\mathbb{P}[\, |g(e|A) - \tilde{g}(e|A)| \geq \delta f(A+e) \,] \leq \mathbb{P}\Big[\sum_{i=1}^N Y_i > \delta N\Big] < 2e^{-\frac{\delta^2 N}{2}}.$$

∎

Appendix B. A randomized, polynomial time distorted local-search algorithm

Our final algorithm is shown in Algorithm 3. Before presenting it in detail, we describe the main concerns involved in its formulation.

B.1. Initialization

We initialize the algorithm with a solution S_0 by using the guarantee for RESIDUALRANDOM-GREEDY provided by Chen et al. (2018) when only γ is bounded. In this case, their analysis shows that the *expected* value of the solution produced by the algorithm is at least $\frac{1}{(1+\gamma^{-1})^2} f(O)$, where O is an optimal solution to the problem. Here, however, we will require a guarantee that holds with high probability. This is easily ensured by independently running RESIDUALRANDOMGREEDY a sufficient number of times and taking the best solution found.

Formally, suppose we set $\varepsilon' = \min(\varepsilon, \frac{1}{128})$ and run RESIDUALRANDOMGREEDY $G = \frac{2 \log(n)}{\varepsilon'^2} = \tilde{\mathcal{O}}(\varepsilon^{-2})$ times independently. For each $1 \leq l \leq G$, let T_l be the solution produced by the lth instance of the RESIDUALRANDOMGREEDY. Define the random variables $Z_l = \frac{1}{(1+\gamma^{-1})^2} - \frac{f(T_l)}{f(O)}$, where $O \in \mathcal{I}$ is the optimal solution. Then, $\mathbb{E}[Z_l] = 0$ and $|Z_l| \leq 1$ for all l.

Let $S_0 = \arg\max_{1 \leq l \leq G} f(T_l)$. Then, by the Chernoff bound (Lemma 24),

$$\mathbb{P}\big[f(S_0) < \big((1+\gamma^{-1})^{-2} - \varepsilon'\big) f(O)\big] \leq \mathbb{P}\Big[\tfrac{1}{G} \sum_{l=1}^G f(T_l) < \big((1+\gamma^{-1})^{-2} - \varepsilon'\big) f(O)\Big]$$
$$= \mathbb{P}\Big[\sum_{l=1}^G Z_l > G\varepsilon'\Big] < e^{-\frac{\varepsilon'^2 G}{2}} = \frac{1}{n}.$$

Thus, with probability at least $1 - \frac{1}{n} = 1 - o(1)$, $f(S_0) \geq \Big(\frac{1}{(1+\gamma^{-1})^2} - \varepsilon'\Big) f(O)$.

B.2. Determining ϕ

In Theorem 9, we considered a (γ, β)-weakly submodular function f, and used the potential g_ϕ with $\phi = \phi(\gamma, \beta) = \gamma^2 + \beta(1 - \gamma)$ to guide the search. In general, however, the values of γ and β may not be known in advance. One approach to coping with this would be to make an appropriate series of guesses for each of the values, then run our the algorithm for each guess and return the best solution obtained.

Here we describe an alternative and more efficient approach: we guess the value of $\phi(\gamma, \beta)$ directly from an appropriate range of values. Moreover, when running the algorithm for each subsequent guess, we initialize the local search procedure using the solution produced by the algorithm for the previous guess. Combined with the bounds from 15, this will allow us to amortize the number of improvements made by the algorithm across all guesses.

In the next lemma, we show that if γ or $\phi(\gamma, \beta)$ is very small, then guarantee for RESIDU-ALRANDOMGREEDY is stronger than that required by our analysis (and so S_0 is already a good solution). This will allow us to bound the range of values for both ϕ and γ that we must consider in our algorithm.

Lemma 17 *For all $\gamma \in (0, 1]$ and $\beta \geq 1$, $\phi(\gamma, \beta) \geq \frac{3}{4}$. Moreover, if $\phi(\gamma, \beta) > 4$ or $\gamma < \frac{1}{7}$, then* $\frac{1}{(1+\gamma^{-1})^2} > \frac{\gamma^2(1 - e^{-\phi(\gamma,\beta)})}{\phi(\gamma,\beta)}.$

Proof First, we show that $\phi(\gamma, \beta) \geq 3/4$ for any value of $\gamma \in (0, 1]$ and $\beta \geq 1$. Note that $\frac{\partial \phi}{\partial \beta} = 1 - \gamma \geq 0$, for all $\gamma \in [0, 1]$. Thus, any minimizer of $\phi(\gamma, \beta)$ sets $\beta = 1$. Moreover, $\frac{\partial \phi}{\partial \gamma} = 2\gamma - \beta$ and $\frac{\partial^2 \phi}{\partial \gamma^2} = 2$ so a $\phi(\gamma, \beta)$ is minimized by $\gamma = \frac{\beta}{2} = \frac{1}{2}$. It follows that $\phi(\gamma, \beta) \geq \phi\left(\frac{1}{2}, 1\right) = \frac{3}{4}$ for all $\gamma \in [0, 1]$ and $\beta \geq 1$.

Now suppose that $\phi(\gamma, \beta) > 4$. Then, the claim follows, since

$$\frac{\gamma^2(1 - e^{-\phi(\gamma,\beta)})}{\phi(\gamma,\beta)} < \frac{\gamma^2}{4} \leq \frac{\gamma^2}{(1 + \gamma)^2} = \frac{1}{(1 + \gamma^{-1})^2}.$$

It remains to consider the case in which $\gamma < \frac{1}{7}$. Recall that $h(x) \triangleq \frac{xe^x}{e^x - 1}$ is increasing in x and so $h(\phi(\gamma, \beta)) \geq h(\frac{3}{4}) > \frac{4}{3}$ (where the last inequality follows directly by computation of $h(\frac{3}{4})$). Suppose that $\gamma < \frac{1}{7}$. Then,

$$\frac{\gamma^2(1 - e^{-\phi(\gamma,\beta)})}{\phi(\gamma,\beta)} = \frac{\gamma^2}{h(\phi(\gamma,\beta))} < \frac{3}{4}\gamma^2.$$

Comparing the previous estimation to the approximation ratio of Chen et al. (2018) and using that $\gamma < 1/7$, we have

$$\frac{\frac{3}{4}\gamma^2}{(1 + \gamma^{-1})^{-2}} = \frac{3}{4}(\gamma + 1)^2 < \frac{3}{4}\left(\frac{8}{7}\right)^2 < 1.$$

Thus, $\frac{3}{4}\gamma^2 < \frac{1}{(1+\gamma^{-1})^2}$ and again the claim follows. ∎

Lemma 17 shows that it suffices to consider $\phi(\gamma, \beta) \in [3/4, 4]$ and $\gamma > 1/7$, since otherwise the starting solution already satisfies the claimed guarantee.[2] Thus, our algorithm considers a geomet-

---

2. We remark that the use of the RESIDUALRANDOMGREEDY is not strictly necessary for our results. One can instead initialize the algorithm with a base containing the best singleton as in the standard local search procedure to obtain a guarantee of γ/k for the initial solution. The remaining arguments can then be modified at the cost of a larger running time dependence on the parameter k.

rically decreasing sequence of guesses for the value $\phi \in [3/4, 4]$, given by $\phi_j = 4(1 - \varepsilon)^j$, where $0 \leq j \leq \lceil \log_{1-\varepsilon} \frac{3}{16} \rceil$. For the first guess, we initialize our algorithm with the solution S_0 produced using several runs of RESIDUALRANDOMGREEDY. For each guess after the this, we initialize S with the approximately locally optimal solution produced for the previous guess.

For each guess, the algorithm proceeds by repeatedly searching for single element swaps that significantly improve the potential $g_\phi(S)$. Specifically, we will exchange an element $a \notin S$ with an element $b \in S$ whenever $\tilde{g}_{\phi_j}(a|S - b) > \tilde{g}_{\phi_j}(b|S - b) + \Delta f(S)$, where $\tilde{g}_{\phi_j}(\cdot|S - b)$ is an estimate of $g_{\phi_j}(\cdot|S - b)$ computed using N samples as described in Section A.4 and Δ is an appropriately chosen parameter. We show that by setting N appropriately, we can ensure that with high probability an approximate local optimum of every g_ϕ is reached after at most some total number M of improvements across all guesses.

B.3. The algorithm and its analysis

Our final algorithm is shown in Algorithm 3. Let $\mathcal{M} = (X, \mathcal{I})$ be a matroid, and $f : 2^X \to \mathbb{R}_+$ be a (γ, β)-weakly submodular function. Given some $0 < \varepsilon \leq 1$ we set the parameters:

$$\Delta = \frac{\varepsilon}{k} \qquad \text{(threshold for accepting improvements)}$$

$$\delta = \frac{\Delta}{4h(4) \cdot H_k} = \frac{\varepsilon}{4h(4) \cdot H_k k} = \Theta(\varepsilon/(k \log k)) \qquad \text{(bound on sampling accuracy)}$$

$$L = 1 + \lceil \log_{1-\varepsilon} \tfrac{3}{16} \rceil = \mathcal{O}(\varepsilon^{-1}) \qquad \text{(number of guesses for ϕ)}$$

$$M = \log_{1+\delta}(7 \cdot 128 \cdot e^{4L\varepsilon} h(4) \cdot H_k) = \tilde{\mathcal{O}}(\delta^{-1}) = \tilde{\mathcal{O}}(k\varepsilon^{-1}) \qquad \text{(total number of improvements)}$$

$$N = 4 \cdot 7^2 \delta^{-2} \ln(Mkn) = \tilde{\mathcal{O}}(\delta^{-2}) = \tilde{\mathcal{O}}(k^2 \varepsilon^{-2}) \qquad \text{(number of samples to estimate g_ϕ)}$$

In Algorithm 3, we evaluate potential improvements using an estimate $\tilde{g}_{\phi_j}(\cdot|S - b)$ for the marginals of g that is computed using N samples. By Lemma 16, we then have $|\tilde{g}_{\phi_j}(e|A) - g_{\phi_j}(e|A)| \leq \gamma \delta f(A + e)$ for any A, e considered by the algorithm with probability at least $1 - 2e^{-\frac{\delta^2 \gamma^2 N}{2}}$. If $\gamma \geq 1/7$, this is at least $1 - 2e^{-\frac{\delta^2}{2 \cdot 7^2} N} = 1 - \frac{2}{(Mkn)^2}$. In our algorithm we will limit the total number of improvements made across all guesses for ϕ to be at most M. Note that any improvement can be found by testing at most kn marginal values, so we must estimate at most Mkn marginal values across the algorithm. By a union bound, we then have $|\tilde{g}_{\phi_j}(e|A) - g_{\phi_j}(e|A)| \leq \gamma \delta f(A + e)$ for *all* A, e considered by Algorithm 3 with probability at least $1 - o(1)$ whenever $\gamma \geq 1/7$. Before proving our main result, let us show that if the algorithm terminates and returns S after making M improvements, we must in fact have an *optimal* solution with high probability.

Lemma 18 *Suppose that $\gamma \geq 1/7$. Then, if Algorithm 3 makes M improvements, the set S it returns satisfies $f(S) \geq f(O)$ with probability $1 - o(1)$.*

Proof With probability $1 - o(1)$ we have $|\tilde{g}_{\phi_j}(e|A) - g_{\phi_j}(e|A)| \leq \gamma \delta f(A + e)$ for any e, A considered by Algorithm 3. Whenever the algorithm exchanges some $a \in X \setminus S$ for $b \in S$ for some guess ϕ_j, we have $\tilde{g}_{\phi_j}(a|S - b) - \tilde{g}_{\phi_j}(b|S - b) \geq \Delta f(S)$ and so

$$
\begin{aligned}
g_{\phi_j}(S - b + a) - g_{\phi_j}(S) &= g_{\phi_j}(a|S - b) - g_{\phi_j}(b|S - b) \\
&\geq \tilde{g}_{\phi_j}(a|S - b) - \delta\gamma f(S - b + a) - \tilde{g}_{\phi_j}(b|S - b) - \delta\gamma f(S) \\
&\geq \tilde{g}_{\phi_j}(a|S - b) - \delta g_{\phi_j}(S - b + a) - \tilde{g}_{\phi_j}(b|S - b) - \delta g_{\phi_j}(S) \\
&\geq \Delta f(S) - \delta g_{\phi_j}(S - b + a) - \delta g_{\phi_j}(S),
\end{aligned}
$$

Algorithm 3: Distorted Local Search Implementation

Let $\Delta = \frac{\varepsilon}{k}$, $\delta = \frac{\Delta}{4h(4) \cdot H_k} = \frac{\varepsilon}{4h(4) \cdot H_k k}$, $M = (1 + \delta^{-1})(37 + \ln(H_k))$, $N = 28\delta^{-2} \ln(Mkn)$,
  $G = \log(n)/(2 \min(\varepsilon, \frac{1}{128})^2)$;
$S_0 \leftarrow$ the best output produced by G independent runs of RESIDUALRANDOMGREEDY
  applied to f and $\mathcal{M}$;
$S_{\max} \leftarrow S_0$;
$i \leftarrow 0$;
for $0 \leq j \leq \lceil \log_{1-\varepsilon} 16/3 \rceil$ **do**
  $\phi \leftarrow 4(1 - \varepsilon)^j$;
  $S \leftarrow S_j$;
  repeat
    isLocalOpt $\leftarrow$ **true**;
    foreach $b \in S$ *and* $a \in X \setminus S$ *with* $S - b + a \in \mathcal{I}$ **do**
      Compute $\tilde{g}_{\phi_j}(a|S - b)$ and $\tilde{g}_{\phi_j}(b|S - b)$ using N random samples;
      if $\tilde{g}_{\phi_j}(a|S - b) > \tilde{g}_{\phi_j}(b|S - b) + \Delta f(S)$ **then**
        $S \leftarrow S - b + a$;
        $i \leftarrow i + 1$;
        isLocalOpt $\leftarrow$ **false**;
        break
  until isLocalOpt *or* $i \geq M$;
  $S_{j+1} \leftarrow S$;
  if $f(S_{j+1}) > f(S_{\max})$ **then** $S_{\max} \leftarrow S_{j+1}$;
return $S_{\max}$

---

where the second inequality follows from the lower bound on g_{ϕ_j} in Lemma (14). Rearranging and using the upper bound on $g_{\phi_j}(S)$ from Lemma 14, together with the definition of δ and Δ, we obtain:

$$g_{\phi_j}(S - b + a) \geq \frac{\Delta f(S) + (1 - \delta)g_{\phi_j}(S)}{1 + \delta} \geq \frac{\frac{\varepsilon}{k}\frac{1}{h(\phi_j) \cdot H_k} + 1 - \delta}{1 + \delta} g_{\phi_j}(S)$$

$$\geq \frac{\frac{\varepsilon}{k}\frac{1}{h(4) \cdot H_k} + 1 - \delta}{1 + \delta} g_{\phi_j}(S) = \frac{1 + 3\delta}{1 + \delta} g_{\phi_j}(S) \geq (1 + \delta)g_{\phi_j}(S), \quad (21)$$

where the last inequality follows from $\frac{1+3x}{1+x} \geq \frac{(1+x)^2}{1+x}$ for all $0 \leq x \leq 1$.

Now suppose that $f(S_0) \geq ((1+\gamma^{-1})^{-2} - \varepsilon')f(O)$, which we have shown also occurs with high probability $1 - o(1)$. Then, since $\gamma \geq \frac{1}{7}$ and $\varepsilon' = \min(\frac{1}{128}, \varepsilon)$, we have $f(S_0) \geq \frac{1}{128}f(O)$. Suppose that $i = M$ when the algorithm is considering some guess ϕ_l. We consider how the current value of $g_{\phi_j}(S)$ changes throughout Algorithm 3, both as improvements are made and as j increases. As shown in (21), each of our M improvements increases this value by a factor of $(1 + \delta)$. Moreover, as shown in Lemma 15,

$$g_{\phi_j}(S) = g_{(1-\varepsilon)\phi_{j-1}}(S) \geq e^{-\phi_{j-1}\varepsilon}g_{\phi_{j-1}}(S) \geq e^{-4\varepsilon}g_{\phi_{j-1}}(S),$$

for any set S. Thus, each time j is incremented, the value $g_{\phi_j}(S)$ decreases by a factor of at most $e^{4\varepsilon}$. Since we made M improvements, we then have:

$$g_{\phi_l}(S_{\ell+1}) \geq (1+\delta)^M e^{-4l\varepsilon} g_{\phi_0}(S_0) \geq (1+\delta)^M e^{-4l\varepsilon} \gamma f(S_0) \geq (1+\delta)^M e^{-4l\varepsilon} \frac{1}{7} \frac{1}{128} f(O) \,,$$

where the second inequality follows from the lower bound on g given in Lemma 14, and the second from $\gamma \geq \frac{1}{7}$. The upper bound on g given by Lemma 14 implies that: $g_{\phi_l}(S_{l+1}) \leq h(\phi_l) H_k f(S_{l+1}) \leq h(4) H_k f(S_{l+1})$. Thus,

$$f(S_{l+1}) \geq (1+\delta)^M e^{-4l\varepsilon} \frac{1}{7 \cdot 128 \cdot h(4) \cdot H_k} f(O) \,.$$

Since $l \leq L$ and $M = \log_{1+\delta}(e^{4L\varepsilon} 7 \cdot 128 \cdot h(4) \cdot H_k)$, the set $S_{\max}$ returned by the algorithm thus has $f(S_{\max}) \geq f(S_{l+1}) \geq f(O)$, as claimed. ∎

We are now ready to prove our main claim, from Section 5, restated here for convenience:

Theorem 11 *Let $\mathcal{M} = (X, \mathcal{I})$ be a matroid, $f : 2^X \to \mathbb{R}_+$ be a (γ, β)-weakly submodular function, and $\varepsilon > 0$. Then, there is a randomized algorithm that with probability $1 - o(1)$ returns a set S satisfying $f(S) \geq \left(\frac{\gamma^2(1 - e^{-\phi(\gamma,\beta)})}{\phi(\gamma,\beta)} - \mathcal{O}(\varepsilon) \right) f(O)$ for any solution $O \in \mathcal{I}$, where $\phi(\gamma, \beta) = \gamma^2 + \beta(1 - \gamma)$. The algorithm runs in time $\tilde{\mathcal{O}}(nk^4 \varepsilon^{-3})$.*

Proof We have shown that $f(S_0) \geq \left((1 + \gamma^{-1})^{-2} - \varepsilon' \right) f(O)$ with probability $1 - o(1)$, where $\varepsilon' = \min(\varepsilon, \frac{1}{128})$. If $\gamma < 1/7$ or $\phi(\gamma, \beta) \notin [3/4, 4]$, then Lemma 17 implies that $(1 + \gamma^{-1})^{-2} > \frac{\gamma^2(1 - e^{-\phi(\gamma,\beta)})}{\phi(\gamma,\beta)}$, and so the claim follows as $f(S_{\max}) \geq f(S_0)$. Thus, we suppose that $\gamma \geq 1/7$ and $\phi(\gamma, \beta) \in [3/4, 4]$. Then, if Algorithm 3 makes M improvements, Lemma 18 implies that the set returned by the algorithm is optimal with probability at least $1 - o(1)$.

In the remaining case, we have $\phi(\gamma, \beta) \in [3/4, 4]$, $\gamma \geq 1/7$, and each set S_{j+1} produced by the algorithm must have $\tilde{g}_{\phi_j}(o_l|S_{j+1} - s_l) \leq \tilde{g}_{\tilde{\phi}_j}(s_l|S_{j+1} - s_l) + \Delta f(S)$ for every $s_l \in S$ and $o_l \in O$. Since $\gamma \geq 1/7$ and the algorithm makes at most M improvements, with probability $1 - o(1)$, we have $\left| \tilde{g}_{\phi_j}(e|A) - g_{\phi_j}(e|A) \right| \leq \gamma \delta f(A + e)$ for all guesses ϕ_j and e, A considered by the algorithm. Thus,

$$\begin{aligned}
g_{\phi_j}&(S_{j+1} - s_l + o_l) - g_{\phi_j}(S_{j+1}) \\
&= g_{\phi_j}(o_l|S_{j+1} - s_l) - g_{\phi_j}(s_l|S_{j+1} - s_l) \\
&\leq \tilde{g}_{\phi_j}(o_l|S_{j+1} - s_l) + \delta\gamma f(S_{j+1} - s_l + o_l) - \tilde{g}_{\phi_j}(s_l|S_{j+1} - s_l) + \delta\gamma f(S_{j+1}) \\
&\leq \Delta f(S_{j+1}) + \delta\gamma f(S_{j+1}) + \delta\gamma f(S_{j+1} - s_l + o_l) \\
&\leq (\Delta + 2\delta)f(O) \\
&= O\left(\tfrac{\varepsilon}{k}\right) \cdot f(O).
\end{aligned}$$

Consider the smallest j such that $\phi_{j+1} \triangleq 4(1 - \varepsilon)^{j+1} < \phi(\gamma, \beta)$. Then, $\phi_{j+1} < \phi(\gamma, \beta) \leq \phi_{j+1}/(1 - \varepsilon) \triangleq \phi_j$. Let $\tilde{\beta} = \frac{\phi_j - \gamma^2}{1 - \gamma}$. Then, $\phi(\gamma, \tilde{\beta}) = \gamma^2 + \frac{\phi_j - \gamma^2}{1 - \gamma}(1 - \gamma) = \phi_j$ and $\tilde{\beta} \geq \frac{\phi(\gamma,\beta) - \gamma^2}{1 - \gamma} = \beta$, so f is also $(\gamma, \tilde{\beta})$-weakly submodular. Theorem 9 then implies

$$f(S_{j+1}) \geq \frac{\gamma^2}{h(\phi_j)} f(O) + \sum_{i=1}^{k} \left[g_{\phi_j}(S) - g_{\phi_j}(S - s_l + o_l) \right] \geq \left(\frac{\gamma^2}{h(\phi_j)} - \mathcal{O}(\varepsilon) \right) f(O)$$

By Lemma 15 part 2, our choice of j, and $\phi_j \leq 4$,

$$h(\phi_j) \leq e^{\phi_j \varepsilon} h((1-\varepsilon)\phi_j) \leq e^{\phi_j \varepsilon} h(\phi(\gamma, \beta)) \leq e^{4\varepsilon} h(\phi(\gamma, \beta))$$

Thus, $f(S_{j+1}) \geq \left(\frac{\gamma^2}{h(\phi(\gamma, \beta))} - \mathcal{O}(\varepsilon) \right) f(O) = \left(\frac{\gamma^2(1 - e^{-\phi(\gamma, \beta)})}{\phi(\gamma, \beta)} - \mathcal{O}(\varepsilon) \right) f(O)$.

The running time of the algorithm is dominated by the number of value oracle queries made to f. The initialization requires running RESIDUALRANDOMGREEDY $\tilde{\mathcal{O}}(\varepsilon^{-2})$ times, each of which requires $\mathcal{O}(nk)$ value queries. The remaining execution makes at most $M = \tilde{\mathcal{O}}(\varepsilon^{-1}k)$ local search improvements, each requiring at most $Nnk = \tilde{\mathcal{O}}(nk^3 \varepsilon^{-2})$ value queries to find. Altogether the running time is thus at most $\tilde{\mathcal{O}}(nk^4 \varepsilon^{-3})$. ∎

Appendix C. A-optimal design for Bayesian linear regression

In Bayesian linear regression, we suppose data is generated by a linear model $\mathbf{y} = X^T \boldsymbol{\theta} + \boldsymbol{\varepsilon}$, where $\mathbf{y} \in \mathbb{R}^n, X \in \mathbb{R}^{p \times n}$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I)$, where I is the identity matrix. Here, $X = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix}$ with $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of data, and $\mathbf{y}$ is a vector corresponding observations for the response variable. The variable $\mathbf{e}$ represents Gaussian noise with 0 mean and variance σ^2. When the number of columns n (i.e., the number of potential observations) is very large, *experimental design* focuses on selecting a small subset $S \subset \{1, 2, \ldots, n\}$ of columns of X to maximally reduce the variance of the estimator $\boldsymbol{\theta}$.

Let $X_S, \mathbf{y}_S$ be the matrix X (the vector $\mathbf{y}$ respectively) restricted to columns (rows respectively) indexed by S. From classical statistical theory, the optimal choice of parameters for any such S is given by $\hat{\boldsymbol{\theta}}_S = (X_S^T X_S)^{-1} X_S \mathbf{y}_S$ and satisfies $\text{Var}(\hat{\boldsymbol{\theta}}_S) = \sigma^2 (X_S^T X_S)^{-1}$. Because the variance of $\hat{\boldsymbol{\theta}}_S$ is a matrix, there is not a universal function which one tries to minimize to find the appropriate set S. Instead, there are multiple objective functions depending on the context leading to different optimality criteria.

As in Krause et al. (2008); Bian et al. (2017); Harshaw et al. (2019), we consider the *A-optimal* design objective. We suppose our prior probability distribution has $\boldsymbol{\theta} \sim \mathcal{N}(0, \Lambda)$. We start by stating a standard result from Bayesian linear regression.

Lemma 19 *Given the previous assumption, and the prior on $\boldsymbol{\theta} \sim \mathcal{N}(0, \Lambda)$, The posterior distribution of $\boldsymbol{\theta}$ follows a normal distribution $p(\boldsymbol{\theta}|\mathbf{y}_S) \sim \mathcal{N}(M_S^{-1} X_S \mathbf{y}_S, M_S^{-1})$, where $M_S^{-1} = \left(\sigma^{-2} X_S X_S^T + \Lambda^{-1} \right)^{-1}$.*

In A-optimal design, our objective function seeks to reduce the variance of the posterior distribution of $\boldsymbol{\theta}$ by reducing the trace of M_S^{-1}, i.e., the sum of the variance of the regression coefficients. Mathematically, we seek to maximize the following objective function

$$F(S) = \text{tr}(\Lambda) - \text{tr}(M_S^{-1}) = \text{tr}(\Lambda) - \text{tr}((\Lambda^{-1} + \sigma^{-2} X_S X_S^T)^{-1}). \tag{22}$$

The function F is not submodular as shown in Krause et al. (2008). The current tightest estimation of the lower weak-submodular ratio of F is due to Harshaw et al. Harshaw et al. (2019). They show that $\gamma \geq (1 + \frac{s^2}{\sigma^2} \lambda_{\max}(\Lambda))^{-1}$, where $s = \max_{i \in [n]} \|\mathbf{x}_i\|$. Here we give a bound on the upper weak-submodularity ratio β.

Theorem 20 *Assume a prior distribution $\boldsymbol{\theta} \sim \mathcal{N}(0, \Lambda)$, and let $s = \max_{i \in [n]} \|\mathbf{x}_i\|$. The function F is $(1/c, c)$-weakly submodular with $c = 1 + \frac{s^2}{\sigma^2} \lambda_{\max}(\Lambda)$.*

Observe that like for the R^2 objective, our upper bound for β is the the inverse of the lower bound for γ.

Proof The lower bound on γ is shown is Harshaw et al. (2019). It remains to prove the upper bound on β. Let B be some set of observations and $A \subseteq B$ with $k = |A|$ and for convenience, define $T = B \setminus A$. By the Sherman-Morrisson-Woodbury formula (see Lemma 22), we have

$$
\begin{aligned}
F(B) - F(A) &= \operatorname{tr}(M_A^{-1}) - \operatorname{tr}(M_B^{-1}) \\
&= \operatorname{tr}\big((M_B - \sigma^{-2} X_T X_T^T)^{-1}\big) - \operatorname{tr}(M_B^{-1}) \\
&= \operatorname{tr}\big(M_B^{-1} + M_B^{-1} X_T (\sigma^2 I - X_T^T M_B^{-1} X_T)^{-1} X_T^T M_B^{-1}\big) - \operatorname{tr}(M_B^{-1}) \\
&= \operatorname{tr}\big(M_B^{-1} X_T (\sigma^2 I - X_T^T M_B^{-1} X_T)^{-1} X_T^T M_B^{-1}\big) \\
&= \operatorname{tr}\big((\sigma^2 I - X_T^T M_B^{-1} X_T)^{-1} X_T^T M_B^{-2} X_T\big).
\end{aligned}
\tag{23}
$$

The third equality uses the linearity of the trace while the last equality uses the cyclic property of the trace. We use the previous equation to derive an upper and lower bound for the numerator and denominator of the submodularity ratio respectively. Applying (23) with $A = B \setminus \{i\}$ (and so $T = \{i\}$) we obtain

$$
F(B) - F(B - i) = \frac{\operatorname{tr}(\mathbf{x}_i^T M_B^{-2} \mathbf{x}_i)}{\sigma^2 - \mathbf{x}_i^T M_B^{-1} \mathbf{x}_i}.
$$

Let $\preceq$ be the Loewner ordering of positive semidefinite matrices, where $A \preceq B$ if and only if $B - A \succeq 0$. First, observe that $\Lambda^{-1} \preceq M_R$ for any set R, which implies that $\Lambda \succeq M_R^{-1}$. Using a second time the Sherman-Morrison-Woodbury formula (Lemma 22) together with the previous observation, we get

$$
\begin{aligned}
\big(\sigma^2 - \mathbf{x}_i^T M_B^{-1} \mathbf{x}_i\big)^{-1} &= \sigma^{-2} + \sigma^{-4} \mathbf{x}_i^T \big(M_B - \sigma^{-2} \mathbf{x}_i \mathbf{x}_i^T\big)^{-1} \mathbf{x}_i, \\
&= \sigma^{-2} + \sigma^{-4} \mathbf{x}_i^T M_{B \setminus \{i\}}^{-1} \mathbf{x}_i, \\
&\leq \sigma^{-2} + \sigma^{-4} \mathbf{x}_i^T \Lambda \mathbf{x}_i, \\
&\leq \sigma^{-2} + \sigma^{-4} \lambda_{\max}(\Lambda) s^2,
\end{aligned}
$$

where $s = \max_i \|\mathbf{x}_i\|_2$ and the last inequality follows by the Courant-Fischer min-max theorem. Summing over all $i \in T = B \setminus A$ and using the linearity of the trace, we have

$$
\begin{aligned}
\sum_{i \in T} F(i|B - i) = \sum_{i \in T} \frac{\operatorname{tr}(\mathbf{x}_i^T M_B^{-2} \mathbf{x}_i)}{\sigma^2 - x_i^T M_B^{-1} x_i} &\leq \big(\sigma^{-2} + s^2 \sigma^{-4} \lambda_{\max}(\Lambda)\big) \sum_{i \in T} \operatorname{tr}(\mathbf{x}_i^T M_B^{-2} \mathbf{x}_i) \\
&= \big(\sigma^{-2} + s^2 \sigma^{-4} \lambda_{\max}(\Lambda)\big) \operatorname{tr}(X_T^T M_B^{-2} X_T).
\end{aligned}
\tag{24}
$$

Returning to the expression of $F(B) - F(A)$, we note that M_B is positive definite, which implies that M_B^{-1} is positive definite. This in turn implies that $-X_T^T M_B^{-1} X_T \preceq 0$ and so $\sigma^2 I - X_T^T M_B^{-1} X_T \preceq \sigma^2 I$. Thus, $(\sigma^2 I - X_T^T M_B^{-1} X_T)^{-1} \succeq \sigma^{-2} I \succ 0$. Therefore,

$$
\operatorname{tr}((\sigma^2 I - X_T^T M_B^{-1} X_T)^{-1} X_T^T M_B^{-2} X_T) \geq \operatorname{tr}(\sigma^{-2} X_T^T M_B^{-2} X_T) = \sigma^{-2} \operatorname{tr}(X_T^T M_B^{-2} X_T).
$$

Combining this with the bound (24), we have:

$$\frac{\sum_{i \in T} F(i|B-i)}{F(B) - F(A)} \leq \frac{\left(\sigma^{-2} + \sigma^{-4}\lambda_{\max}(\Lambda) \cdot s^2\right) \operatorname{tr}\left(X_T^T M_B^{-2} X_T\right)}{\sigma^{-2} \operatorname{tr}\left(X_T^T M_B^{-2} X_T\right)} \leq 1 + \frac{s^2}{\sigma^2}\lambda_{\max}(\Lambda).$$

Recalling that $T = B \setminus A$, this completes the proof. ∎

Appendix D. How large can β be?

We have shown that the R^2 objective (Section 3) and the A-optimal design objective for Bayesian linear regression (Section C) are $(c, 1/c)$-weakly submodular for some parameter c. A natural question to ask is whether, given $\gamma > 0$, there is a small non-trivial bound for β independent of the size of the ground set. Here we show that this is not true in general, by proving the following claim stated in Section 5:

Theorem 12 *For any $\gamma > 0$ and $k > 0$ there exists a function on a ground set of size k that is γ-weakly submodular from below but not β-weakly submodular from above for any $\beta < \binom{k-\gamma}{k-1} = \Theta(k^{1-\gamma})$.*

The intuition behind the construction is simple. We build a set function recursively with lower submodularity ratio exactly γ. The recurrence relation holds until the $(k-1)^{\text{th}}$ marginal, which allows us to have a large value for the final marginal and thus increase β.

Proof of Theorem 12 We start by constructing a monotone set function f on a ground set of k elements. The elements are indistinguishable, meaning that for any given set S, two elements $e, e' \in X \setminus S$ have the same marginal contribution. Therefore, because elements are indistinguishable, the value of a set is a function of its size. Let x_i be the value of any set of size $i = 0, 1, \ldots, k$. Additionally, let $x_0 = f(\emptyset) = 0$ and $x_k = 1$. We define x_i inductively with the following recurrence for $i = 0, 1, \ldots, k-2$:

$$x_{i+1} = \frac{k-i-\gamma}{k-i} \cdot x_i + \frac{\gamma}{k-i} \quad \text{or equivalently} \quad x_{i+1} - x_i = \frac{\gamma}{k-i}(1 - x_i). \quad (25)$$

It can easily be shown (by induction) that the described sequence is valid, i.e. it is monotone and each $x_i \in [0, 1]$. Additionally, we note that the sequence satisfies:

$$1 - x_{i+1} = 1 - \left(\frac{k-i-\gamma}{k-i} \cdot x_i + \frac{\gamma}{k-i}\right) = \left(1 - \frac{\gamma}{k-i}\right)(1 - x_i), \quad (26)$$

for all $i = 0, 1, \ldots, k-2$.

First, we show that f has a lower submodularity ratio at most γ. We prove that for any B and $A \subset B$ such that $|B| = j$ and $|A| = i$:

$$\frac{\sum_{e \in B \setminus A} f(e|A)}{f(B) - f(A)} = \frac{(j-i)(x_{i+1} - x_i)}{x_j - x_i} \geq \gamma. \quad (27)$$

First, we consider the case in which $j = k$. If $i = k-1$, then the left-hand side of (27) is 1. If $i \leq k-2$, then applying the identity (25), and recalling that $x_k = 1$ gives:

$$\frac{(k-i)(x_{i+1} - x_i)}{x_k - x_i} = \frac{(k-i) \cdot \frac{\gamma}{k-i}(1 - x_i)}{1 - x_i} = \gamma.$$

for any $i = 1, \ldots, k-1$.

Next, we consider the case in which $j \leq k-1$ and so $i \leq k-2$. Then, by employing recursively the identity (26) we obtain

$$x_j - x_i = (1 - x_i) - (1 - x_j) = (1 - x_i)\left(1 - \prod_{\ell=i}^{j-1}\left(1 - \frac{\gamma}{k-\ell}\right)\right). \tag{28}$$

Since $\gamma \in [0, 1]$, we can use the generalized Bernoulli $(1 - 1/n)^x \leq 1 - x/n$ for $x \in [0, 1]$ to bound each term in the product above. This gives:

$$\prod_{\ell=i}^{j-1}\left(1 - \frac{\gamma}{k-\ell}\right) \geq \prod_{\ell=i}^{j-1}\left(1 - \frac{1}{k-\ell}\right)^\gamma = \left(\frac{k-j}{k-i}\right)^\gamma = \left(1 - \frac{j-i}{k-i}\right)^\gamma \geq 1 - \left(\frac{j-i}{k-i}\right),$$

where in the last inequality we used the fact that $(1 - x)^\gamma \geq 1 - x$ for $x \in [0, 1]$ and $\gamma \in [0, 1]$. Thus we have:

$$\frac{(j-i)(x_{i+1} - x_i)}{x_j - x_i} = \frac{(j-i)\frac{\gamma}{k-i}(1 - x_i)}{(1 - x_i)\left(1 - \prod_{\ell=i}^{j-1}\left(1 - \frac{\gamma}{k-\ell}\right)\right)} \geq \frac{(j-i)\frac{\gamma}{k-i}(1 - x_i)}{(1 - x_i)\frac{j-i}{k-i}} = \gamma,$$

where we have used (25) and (28) in the first equation. Combining these cases, we find f is γ-weakly submodular from below.

To complete the proof we now show that f is not β-weakly submodular from above for any $\beta < \binom{k-\gamma}{k-1}$. Here we consider the case in which $A = \emptyset$ and $B = X$ and show that:

$$\frac{\sum_{e \in B} f(e|B - e)}{f(B) - f(\emptyset)} = \frac{k(x_k - x_{k-1})}{x_k - x_0} = \binom{k-\gamma}{k-1}.$$

Recall that $x_k = 1$, and $x_0 = 0$, which implies that the denominator is equal to 1. Recursively applying the identity (26) gives

$$k(1 - x_{k-1}) = k \cdot \prod_{\ell=0}^{k-2}\left(1 - \frac{\gamma}{k-\ell}\right) = k\prod_{\ell=2}^{k}\left(1 - \frac{\gamma}{\ell}\right)$$

$$= \frac{\prod_{\ell=2}^{k}(\ell - \gamma)}{\prod_{\ell=1}^{k-1}\ell} = \frac{\prod_{\ell=1}^{k-1}(\ell + 1 - \gamma)}{\prod_{\ell=1}^{k-1}\ell} = \binom{k-\gamma}{k-1},$$

as required. ∎

Appendix E. Basic results from linear algebra and probability

We make use of the following basic results related to matrix inverses (see e.g. (Rasmussen and Williams, 2005, Section A.3))

Lemma 21 (Block Matrix Inverse) *Let B and A be matrices of dimension $k \times k$ and $h \times h$, respectively, and let U and V be matrices of dimension $k \times h$ and $h \times k$ respectively. Then,*

$$\begin{pmatrix} B & U \\ V & A \end{pmatrix}^{-1} = \begin{pmatrix} B^{-1} + B^{-1}USVB^{-1} & -B^{-1}US \\ -SVB^{-1} & S \end{pmatrix}$$

where $S = (A - VB^{-1}U)^{-1}$ is the Schur complement of B.

Lemma 22 (Sherman-Morrisson-Woodbury formula) *Let A, U, C, V be matrices of conformable sizes. Then,*
$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

We also make use of the following bound on eigenvalues of normalized covariance matrices shown in Lemmas 13, 14, and 15 of Das and Kempe (2018):

Lemma 23 *Let $\mathcal{L}$ and $\mathcal{S} = \{X_1, X_2, \ldots, X_n\}$ be two disjoint sets of zero-mean random variables each of which has variance at most 1. Let C be the covariance matrix of the set $\mathcal{L} \cup \mathcal{S}$. Let C_ρ be the covariance matrix of the set $\{\mathrm{Res}(X_1, \mathcal{L}), \ldots, \mathrm{Res}(X_n, \mathcal{L})\}$ after normalization of the random variables to have unit variance. Then $\lambda_{\min}(C_\rho) \geq \lambda_{\min}(C)$.*

Finally, we use the following form of the Chernoff bound given in (Alon and Spencer, 2016, Theorem A.1.16):

Lemma 24 (Chernoff Bound) *Let X_i, $1 \leq i \leq n$ be mutually independent random variables with $\mathbb{E}[X_i] = 0$ and $|X_i| \leq 1$ for all i. Set $S = X_1 + \cdots + X_n$. Then for any a,*

$$\Pr[S > a] < e^{-a^2/2n}$$