Optimal approximation for unconstrained non-submodular minimization

Marwa El Halabi MARWASH@MIT.EDU
Stefanie Jegelka STEFJE@MIT.EDU
Massachusetts Institute of Technology

Abstract

Submodular function minimization is a well studied problem; existing algorithms solve it exactly or up to arbitrary accuracy. However, in many applications, the objective function is not exactly submodular. No theoretical guarantees exist in this case. While submodular minimization algorithms rely on intricate connections between submodularity and convexity, we show that these relations can be extended sufficiently to obtain approximation guarantees for non-submodular minimization. In particular, we prove how a projected subgradient method can perform well even for certain non-submodular functions. This includes important examples, such as objectives for structured sparse learning and variance reduction in Bayesian optimization. We also extend this result to noisy function evaluations. Our algorithm works in the value oracle model. We prove that in this model, the approximation result we obtain is the best possible with a subexponential number of queries.

1. Introduction

Many machine learning problems can be formulated as minimizing a *set function* H. In general, this problem is NP-hard, but it can be solved efficiently with additional structure. An important example is when H is *submodular*, i.e., it satisfies the diminishing returns (DR) property $H(A \cup \{i\}) - H(A) \geq H(B \cup \{i\}) - H(B)$, for all $A \subseteq B, i \in V \setminus B$. Several algorithms minimize submodular functions in polynomial time, either exactly or within arbitrary accuracy [1, 3, 10, 20, 25, 30, 38]. Submodularity is a natural model for a variety of applications. However, in many applications, such as structured sparse learning and Bayesian optimization, the objective function is not exactly submodular. Instead, it satisfies a weaker form of the diminishing returns property. An important class of such functions are α-*weakly DR-submodular* functions, introduced in [31]. The parameter α characterizes how close the function is to being submodular (see Section 2 for a precise definition). Furthermore, in many cases, only noisy evaluations of the objective are available. A natural question then arises: *Can submodular minimization algorithms extend to such non-submodular noisy functions?*

Non-submodular *maximization*, under various notions of approximate submodularity, has recently received a lot of attention [5, 11, 22–24, 28, 37, 39]. In contrast, only few studies consider non-submodular *minimization* [4, 34, 40, 43]. In this paper, we initiate the study of the *unconstrained* non-submodular minimization problem

$$\min_{S \subseteq V} \; H(S) := F(S) - G(S), \tag{1}$$

where F and G are normalized ($F(\emptyset) = G(\emptyset) = 0$) monotone (non-decreasing or non-increasing) functions, F is α-*weakly DR-submodular*, and G is β-*weakly DR-supermodular*, i.e., $-G$ is β^{-1}-

weakly DR-submodular. The definitions of weak DR-sub-/supermodularity only hold for monotone functions, and thus do not directly apply to H. This setting covers several important applications, including structured sparse learning and Bayesian optimization. In fact, we show that any set function H can be decomposed into functions F and G that satisfy these assumptions, albeit with properties leading to weaker approximations when the function is far from being submodular.

A key strategy for minimizing submodular functions exploits their tractable tight convex extension, which enables convex optimization algorithms. In general, such a tractable tight convex extension is impossible. Yet, in this paper, we show that for approximately submodular functions, we may approximate the subgradients of their intractable tight convex extension, and use them in a projected subgradient method (PGM), to obtain an approximate solution to Problem (1). This insight broadly expands the scope of submodular minimization techniques.

Contributions We provide the *first* approximation guarantee for unconstrained non-submodular minimization: PGM achieves a tight approximation of $H(S) \leq F(S^*)/\alpha - \beta G(S^*) + \epsilon$. We extend this result to the case where only a noisy oracle of H is accessible. We prove that this guarantee is optimal in the value oracle model. We apply our results to structured sparse learning, and structured batch Bayesian optimization, implying the *first* approximation guarantees for these problems.

2. Preliminaries

Let V be the ground set of size d. Given a set function F, the *marginal gain* of adding an element i to a set $A \subseteq V$ is $F(i|A) = F(A \cup \{i\}) - F(A)$. F is non-decreasing (non-increasing) if $F(A) \leq F(B)$ ($F(A) \geq F(B)$) for all $A \subseteq B$. F is *submodular* if $F(i|A) \geq F(i|B)$ for all $A \subseteq B, i \in V \setminus B$, *supermodular* if $F(i|A) \leq F(i|B)$, and *modular* if both hold. Relaxing these inequalities leads to the notions of weak DR-sub-/supermodularity introduced in [31] and [5].

Definition 1 *A set function F is* α-weakly DR-submodular, *with $\alpha > 0$, if*

$$F(i|A) \geq \alpha F(i|B), \forall A \subseteq B, i \in V \setminus B$$

Similarly, F is β-weakly DR-supermodular, with $\beta > 0$, if

$$F(i|B) \geq \beta F(i|A), \forall A \subseteq B, i \in V \setminus B.$$

We say that F is (α, β)-weakly DR-modular *if it satisfies both properties.*

If F is non-decreasing, then $\alpha, \beta \in (0, 1]$, and if it is non-increasing, then $\alpha, \beta \geq 1$. F is submodular (supermodular) iff $\alpha = 1$ ($\beta = 1$) and modular iff both $\alpha = \beta = 1$.

Minimizing a submodular set function F is equivalent to minimizing a non-smooth convex function, obtained by considering a *continuous extension* of F, from vertices of the hypercube $\{0, 1\}^d$ to the full hypercube $[0, 1]^d$. This extension, called the *Lovász extension* [32], is defined for all $\boldsymbol{s} \in \mathbb{R}^d$ as $f_L(\boldsymbol{s}) = \sum_{k=1}^d s_{j_k} F(j_k|S_{k-1})$, where $s_{j_1} \geq \cdots \geq s_{j_d}$ and $S_k = \{j_1, \cdots, j_k\}$, and is convex if and only if F is submodular. When F is submodular, minimizing f_L or F is equivalent. Moreover, a subgradient $\boldsymbol{\kappa}$ of f_L at any $\boldsymbol{s} \in \mathbb{R}^d$ can be computed efficiently by sorting the entries of $\boldsymbol{s}$ in decreasing order and taking $\kappa_{j_k} = F(j_k|S_{k-1})$ for all $k \in V$ [14].

This relation between submodularity and convexity allows for generic convex optimization algorithms to be used for minimizing F. However, it has been unclear how these relations are affected if the function is only approximately submodular. In this paper, we establish a similar relation between approximate submodularity and approximate convexity.

3. Main Results

We consider first the case where F and G are non-decreasing. We later extend our results to non-increasing functions. We assume a *value oracle* access to H; i.e., there is an oracle that, given a set $S \subseteq V$, returns the value $H(S)$. Interestingly, any set function can be decomposed in the form assumed in Problem (1), as the following proposition shows.

Proposition 2 *Given a set function H, and $\alpha, \beta \in (0, 1]$ such that $\alpha\beta < 1$, there exists a non-decreasing α-weakly DR-submodular function F and a non-decreasing (α, β)-weakly DR-modular function G such that $H(S) = F(S) - G(S)$ for all $S \subseteq V$.*

Computing such a decomposition is NP-hard in general, but is *not* required to run PGM. When H is far from being submodular, it may not be possible to decompose H as to obtain a non-trivial guarantee. However, many important non-submodular functions do admit a decomposition which leads to non-trivial bounds. We call such functions approximately submodular.

3.1 Continuous relaxations When H is not submodular, the connections between its Lovász extension and tight convex relaxation for exact minimization, outlined in Section 2, break down. However, Problem (1) can still be converted to a non-smooth convex optimization problem, via a different convex extension. Given a set function H, its *convex closure* h^- is the point-wise largest convex function from $[0, 1]^d$ to $\mathbb{R}$ that always lower bounds H. The following equivalence holds [13, Prop. 3.23]:

$$\min_{S \subseteq V} H(S) = \min_{\boldsymbol{s} \in [0,1]^d} h^-(\boldsymbol{s}). \tag{2}$$

Unfortunately, evaluating and optimizing the convex closure of a general set function is NP-hard [42]. The key property that makes Problem (2) efficient to solve when H is submodular is that its convex closure then coincides with its Lovász extension, i.e., $h^- = h_L$. This property no longer holds if H is only approximately submodular. But, in this case, a weaker key property holds: we show in Lemma 3 that the Lovász extension approximates h^-, and that the same vectors that served as its subgradients in the submodular case can still serve as approximate subgradients to h^-.

Lemma 3 *Given a vector $\boldsymbol{s} \in [0, 1]^d$ such that $s_{j_1} \geq \cdots \geq s_{j_d}$, we define $\boldsymbol{\kappa}$ such that $\kappa_{j_k} = H(j_k | S_{k-1})$ where $S_k = \{j_1, \cdots, j_k\}$. Then, $h_L(\boldsymbol{s}) = \boldsymbol{\kappa}^\top \boldsymbol{s} \geq h^-(\boldsymbol{s})$, $\boldsymbol{\kappa}(A) \leq \frac{1}{\alpha} F(A) - \beta G(A)$ for all $A \subseteq V$, and $\boldsymbol{\kappa}^\top \boldsymbol{s}' \leq \frac{1}{\alpha} f^-(\boldsymbol{s}') + \beta(-g)^-(\boldsymbol{s}')$ for all $\boldsymbol{s}' \in [0, 1]^d$.*

We can view the vector $\boldsymbol{\kappa}$ in Lemma 3 as an approximate subgradient of h^- at $\boldsymbol{s}$ in the following sense: $\frac{1}{\alpha} f^-(\boldsymbol{s}') + \beta(-g)^-(\boldsymbol{s}') \geq h^-(\boldsymbol{s}) + \langle \boldsymbol{\kappa}, \boldsymbol{s}' - \boldsymbol{s} \rangle, \forall \boldsymbol{s}' \in [0, 1]^d$. Lemma 3 also implies that h_L approximates h^- in the following sense: $h^-(\boldsymbol{s}) \leq h_L(\boldsymbol{s}) \leq \frac{1}{\alpha} f^-(\boldsymbol{s}) + \beta(-g)^-(\boldsymbol{s}), \forall \boldsymbol{s} \in [0, 1]^d$. We can thus say that h_L is approximately convex in this case. This key insight allows us to approximately minimize h^- using simple convex optimization algorithms.

3.2 Algorithm Equipped with the approximate subgradients of h^-, we can now apply an approximate projected subgradient method (PGM). Starting from an arbitraty $\boldsymbol{s}^1 \in [0, 1]^d$, PGM iteratively updates $\boldsymbol{s}^{t+1} = \Pi_{[0,1]^d}(\boldsymbol{s}^t - \eta \boldsymbol{\kappa}^t)$, where $\boldsymbol{\kappa}^t$ is the approximate subgradient at $\boldsymbol{s}^t$ from Lemma 3, and $\Pi_{[0,1]^d}$ is the projection onto $[0, 1]^d$. We set the step size to $\eta = \frac{R}{L\sqrt{T}}$, where $L = F(V) + G(V)$ is the Lipschitz constant, i.e., $\|\boldsymbol{\kappa}^t\|_2 \leq L$ for all t, and $R = 2\sqrt{d}$ is the domain radius $\|\boldsymbol{s}^1 - \boldsymbol{s}^*\|_2 \leq R$.

Importantly, the algorithm does not need to know the α and β parameters, which can be hard to compute in practice. In fact, the iterates taken are exactly the same as in the submodular case.

Theorem 4 *After T iterations of PGM, $\hat{\boldsymbol{s}} \in \arg\min_{t\in\{1,\cdots,T\}} h_L(\boldsymbol{s}^t)$ satisfies:*

$$h^-(\hat{\boldsymbol{s}}) \;\leq\; h_L(\hat{\boldsymbol{s}}) \;\leq\; \tfrac{1}{\alpha} f^-(\boldsymbol{s}^*) + \beta(-g)^-(\boldsymbol{s}^*) + \tfrac{RL}{\sqrt{T}},$$

where $\boldsymbol{s}^* \in \arg\min_{\boldsymbol{s}\in[0,1]^d} h^-(\boldsymbol{s})$. Let $\hat{S}_k = \{j_1, \cdots, j_k\}$ such that $\hat{s}_{j_1} \geq \cdots \geq \hat{s}_{j_d}$, and $\hat{S}_0 = \emptyset$. Then $\hat{S} \in \arg\min_{k\in\{0,\cdots,d\}} H(\hat{S}_k)$ satisfies

$$H(\hat{S}) \leq \tfrac{1}{\alpha} F(S^*) - \beta G(S^*) + \tfrac{RL}{\sqrt{T}},$$

where $S^* \in \arg\min_{S\subseteq V} H(S)$. This bound is tight even if F and G are weakly DR-modular.

To obtain a set that satisfies $H(\hat{S}) \leq F(S^*)/\alpha - \beta G(S^*) + \epsilon$, we thus need $O(dL^2/\epsilon^2)$ iterations of PGM, where the time per iteration is $O(d\log d + d\,\mathrm{EO})$, with EO the evaluation oracle time.
If F is regarded as a cost and G as a revenue, this guarantee states that the returned solution achieves at least a fraction β of the revenue of the optimal solution, by paying at most a $1/\alpha$-multiple of the cost. The quality of this guarantee depends on F, G and their parameters α, β; it becomes vacuous when $F(S^*)/\alpha \geq \beta G(S^*)$. If F is submodular and G is supermodular, Problem (1) reduces to submodular minimization and Theorem 4 recovers the guarantee $H(\hat{S}) \leq H(S^*) + RL/\sqrt{T}$. This result also extends to the case where F and G are non-increasing functions with $F(V) = G(V) = 0$. Applying PGD to $H(V \setminus S)$ then yields $H(S) \leq \alpha F(S^*) - G(S^*)/\beta + RL/\sqrt{T}$.

3.3 Extension to noisy evaluations To the best of our knowledge, *minimizing* noisy oracles of submodular functions was only studied in [6]. We address a more general setup where the underlying function H is not necessarily submodular. We assume again that F and G are normalized and non-decreasing. The result easily extend to non-increasing functions by minimizing $H(V \setminus S)$.

Proposition 5 *Assume we have an approximate oracle $\tilde{H}$ with input parameters $\epsilon, \delta \in (0,1)$, such that for every $S \subseteq V$, $|\tilde{H}(S) - H(S)| \leq \epsilon$ with probability $1-\delta$. We run PGM with $\tilde{H}$ for T iterations. Let $\hat{\boldsymbol{s}} = \arg\min_{t\in\{1,\cdots,T\}} \tilde{h}_L(\boldsymbol{s}^t)$, and $\hat{S}_k = \{j_1, \cdots, j_k\}$ such that $\hat{s}_{j_1} \geq \cdots \geq \hat{s}_{j_d}$. Then $\hat{S} \in \arg\min_{k\in\{0,\cdots,d\}} \tilde{H}(\hat{S}_k)$ satisfies*

$$H(\hat{S}) \leq \tfrac{1}{\alpha} F(S^*) - \beta G(S^*) + \epsilon',$$

with probability $1-\delta'$, by choosing $\epsilon = \frac{\epsilon'}{8d}$, $\delta = \frac{\delta' \epsilon'^2}{32 d^2}$ and $T = (4\sqrt{d}L/\epsilon')^2$ iterations.

Blais et al [6] consider the same setup for the special case of submodular H, and use the cutting plane method of [30]. Their runtime has better dependence $O(\log(1/\epsilon'))$ on the error ϵ', but worse dependence $O(d^3)$ on the dimension $d = |V|$, and their result needs oracle accuracy $\epsilon = O(\epsilon'^2/d^5)$.

3.4 Inapproximability Result Problem (1) is equivalent to general set function minimization by Proposition 2. Then solving Problem (1) with any multiplicative factor approximation is NP-Hard [26, 41]. Moreover, in the value oracle model, it is not possible to obtain any multiplicative constant factor approximation, using a subexponential number of queries [26]. It is thus necessary to consider bicriteria-like approximation guarantees as we do in Theorem 4. We prove now that this approximation guarantee is optimal in the value oracle model.

Theorem 6 *For any $\alpha, \beta \in (0, 1]$ such that $\alpha\beta < 1, d > 2$ and $\delta > 0$, there are instances of Problem (1) such that no (deterministic or randomized) algorithm, using less than exponentially many queries, can always find a solution $S \subseteq V$ of expected value at most $\frac{1}{\alpha}F(S^*) - \beta G(S^*) - \delta$.*

3.5 Applications

We discuss two application examples that benefit from the theory in this work.

Structured sparse learning: Structured sparse learning aims to estimate a *sparse* parameter vector whose support is known to have a particular *structure*, such as group-sparsity, clustering, tree-structure, or diversity [29, 33]. Such problems can be formulated as $\min_{\boldsymbol{x}\in\mathbb{R}^d} \ell(\boldsymbol{x}) + \lambda F(\text{supp}(\boldsymbol{x}))$, where $\text{supp}(\boldsymbol{x}) = \{i \in V | x_i \neq 0\}$, ℓ is a convex loss function and F is a set function favoring the desirable supports. One may write this problem as $\min_{S\subseteq V} \lambda F(S) - G^\ell(S)$, where $G^\ell(S) = \ell(0) - \min_{\text{supp}(\boldsymbol{x})\subseteq S} \ell(\boldsymbol{x})$ is a normalized non-decreasing set function. Recently, it was shown that if ℓ has restricted smoothness and strong convexity, G^ℓ is weakly modular [8, 18, 36]; a notion of approximate modularity which is weaker than weak DR-modularity. This allowed for approximation guarantees of greedy algorithms to be applied to the constrained variant of this problem, but only for the special case of sparsity constraint [11, 18], and for some near-modular constraints [37]. In applications, however, the structure of interest is often better modeled by a non-modular regularizer F, which may be submodular [2] or non-submodular [15, 16]. Weak modularity of G^ℓ is not enough for our results to apply, but, if the loss function ℓ is smooth, strongly convex, and is generated from random data, then we show that G^ℓ is also (α_G, β_G)-weakly DR-modular, for some $\alpha_G, \beta_G > 0$ that depend on the conditioning of ℓ.

Our results thus apply whenever F is weakly DR-submodular. Examples include submodular regularizers [2], but also non-submodular ones such as the range function [16], which favors interval supports, with applications in time-series and cancer diagnosis [35], and the cost function considered in [37], which favors the selection of sparse and cheap features, with applications in healthcare.

Structured batch Bayesian optimization: The goal in batch Bayesian optimization is to optimize an unknown expensive-to-evaluate noisy function with as few batches of function evaluations as possible [12, 21]. For example, evaluations can correspond to performing expensive experiments. The evaluation points are chosen to maximize an acquisition function subject to a cardinality constraint. Several acquisition functions have been proposed for this purpose, amongst others the *variance reduction* function [7, 27]. This function is used to maximally reduce the variance of the posterior distribution over potential maximizers of the unknown function. Often, the unknown function is modeled by a Gaussian process. In this case, we show that the variance reduction function is normalized non-decreasing (β, β)-weakly DR-modular with $\beta = \frac{\lambda^2_{\min}(\boldsymbol{K})}{\lambda_{\max}(\boldsymbol{K})(\lambda_{\min}(\boldsymbol{K})+\sigma^2)}$, where $\boldsymbol{K}$ is the positive definite kernel matrix, and $\lambda_{\max}(\boldsymbol{K}), \lambda_{\min}(\boldsymbol{K})$ are its largest and smallest eigenvalues. It can thus be maximized with a greedy algorithm to a β-approximation [40].

This problem may also be phrased as an instance of Problem (1), with G being the variance reduction function, and $F(S) = \lambda|S|$ an item-wise cost. This formulation easily allows to include nonlinear costs with (weak) decrease in marginal costs (economies of scale). For example, in the sensor placement application, the cost of placing a sensor in a hazardous environment may diminish if other sensors are also placed in similar environments. Unlike previous works, the approximation guarantee in Theorem 4 still applies to such cost functions, while maintaining the β-approximation with respect to G.

References

[1] Brian Axelrod, Yang P Liu, and Aaron Sidford. Near-optimal approximate discrete and continuous submodular function minimization. *arXiv preprint arXiv:1909.00171*, 2019.

[2] F. Bach. Structured sparsity-inducing norms through submodular functions. In *NIPS*, pages 118–126, 2010.

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

[4] Wenruo Bai, Rishabh Iyer, Kai Wei, and Jeff Bilmes. Algorithms for optimizing the ratio of submodular functions. In *International Conference on Machine Learning*, pages 2751–2759, 2016.

[5] Andrew An Bian, Joachim M Buhmann, Andreas Krause, and Sebastian Tschiatschek. Guarantees for greedy maximization of non-submodular functions with applications. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 498–507. JMLR. org, 2017.

[6] Eric Blais, Clément L Canonne, Talya Eden, Amit Levi, and Dana Ron. Tolerant junta testing and the connection to submodular optimization and function isomorphism. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2113–2132. Society for Industrial and Applied Mathematics, 2018.

[7] Ilija Bogunovic, Jonathan Scarlett, Andreas Krause, and Volkan Cevher. Truncated variance reduction: A unified approach to bayesian optimization and level-set estimation. In *Advances in neural information processing systems*, pages 1507–1515, 2016.

[8] Ilija Bogunovic, Junyao Zhao, and Volkan Cevher. Robust maximization of non-submodular objectives. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 890–899. PMLR, 09–11 Apr 2018. URL http://proceedings.mlr.press/v84/bogunovic18a.html.

[9] Sébastien Bubeck. Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980*, 15, 2014.

[10] Deeparnab Chakrabarty, Yin Tat Lee, Aaron Sidford, and Sam Chiu-wai Wong. Subquadratic submodular function minimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 1220–1231, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4528-6. doi: 10.1145/3055399.3055419. URL http://doi.acm.org/10.1145/3055399.3055419.

[11] A. Das and D. Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv preprint arXiv:1102.3975*, 2011.

[12] Thomas Desautels, Andreas Krause, and Joel W. Burdick. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *Journal of Machine*

Learning Research, 15:4053–4103, 2014. URL http://jmlr.org/papers/v15/desautels14a.html.

[13] Shaddin Dughmi. Submodular functions: Extensions, distributions, and algorithms. a survey. *arXiv preprint arXiv:0912.0322*, 2009.

[14] Jack Edmonds. Submodular functions, matroids, and certain polyhedra. In *Combinatorial Optimization–Eureka, You Shrink!*, pages 11–26. Springer, 2003.

[15] M. El Halabi and V. Cevher. A totally unimodular view of structured sparsity. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, pp. 223-231*, 2015.

[16] M. El Halabi, F. Bach, and V Cevher. Combinatorial penalties: Structure preserved by convex relaxations. *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, 2018.

[17] Marwa El Halabi. *Learning with Structured Sparsity: From Discrete to Convex and Back*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2018.

[18] 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.

[19] Uriel Feige, Vahab S Mirrokni, and Jan Vondrák. Maximizing non-monotone submodular functions. *SIAM Journal on Computing*, 40(4):1133–1153, 2011.

[20] S. Fujishige and S. Isotani. A submodular function minimization algorithm based on the minimum-norm base. *Pacific Journal of Optimization*, 7(1):3–17, 2011.

[21] Javier González, Zhenwen Dai, Philipp Hennig, and Neil Lawrence. Batch bayesian optimization via local penalization. In *Artificial intelligence and statistics*, pages 648–657, 2016.

[22] Avinatan Hassidim and Yaron Singer. Submodular optimization under noise. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1069–1122, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR. URL http://proceedings.mlr.press/v65/hassidim17a.html.

[23] Avinatan Hassidim and Yaron Singer. Optimization for approximate submodularity. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 394–405. Curran Associates Inc., 2018.

[24] Thibaut Horel and Yaron Singer. Maximization of approximately submodular functions. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3045–3053. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6236-maximization-of-approximately-submodular-functions.pdf.

[25] Satoru Iwata and James B Orlin. A simple combinatorial algorithm for submodular function minimization. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 1230–1237. Society for Industrial and Applied Mathematics, 2009.

[26] Rishabh Iyer and Jeff Bilmes. Algorithms for approximate minimization of the difference between submodular functions, with applications. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, pages 407–417, Arlington, Virginia, United States, 2012. AUAI Press. ISBN 978-0-9749039-8-9. URL http://dl.acm.org/citation.cfm?id=3020652.3020697.

[27] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(Feb):235–284, 2008.

[28] Alan Kuhnle, J David Smith, Victoria G Crawford, and My T Thai. Fast maximization of non-submodular, monotonic functions on the integer lattice. *arXiv preprint arXiv:1805.06990*, 2018.

[29] Anastasios Kyrillidis, Luca Baldassarre, Marwa El Halabi, Quoc Tran-Dinh, and Volkan Cevher. Structured sparsity: Discrete and convex approaches. In *Compressed Sensing and its Applications*, pages 341–387. Springer, 2015.

[30] Yin Tat Lee, Aaron Sidford, and Sam Chiu-wai Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1049–1065. IEEE, 2015.

[31] Benny Lehmann, Daniel Lehmann, and Noam Nisan. Combinatorial auctions with decreasing marginal utilities. *Games and Economic Behavior*, 55(2):270–296, 2006.

[32] L. Lovász. Submodular functions and convexity. In *Mathematical Programming The State of the Art*, pages 235–257. Springer, 1983.

[33] Guillaume Obozinski and Francis Bach. A unified perspective on convex structured sparsity: Hierarchical, symmetric, submodular norms and beyond. 2016. URL https://hal-enpc.archives-ouvertes.fr/hal-01412385.

[34] Chao Qian, Jing-Cheng Shi, Yang Yu, Ke Tang, and Zhi-Hua Zhou. Optimizing ratio of monotone set functions. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 2606–2612. AAAI Press, 2017. ISBN 978-0-9992411-0-3. URL http://dl.acm.org/citation.cfm?id=3172077.3172251.

[35] F. Rapaport, E. Barillot, and J.P. Vert. Classification of arraycgh data using fused svm. *Bioinformatics*, 24(13):i375–i382, 2008.

[36] Shinsaku Sakaue. Weakly modular maximization: Applications, hardness, tractability, and efficient algorithms. *arXiv preprint arXiv:1805.11251*, 2018.

[37] Shinsaku Sakaue. Greedy and iht algorithms for non-convex optimization with monotone costs of non-zeros. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of*

Machine Learning Research, volume 89 of *Proceedings of Machine Learning Research*, pages 206–215. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/sakaue19a.html.

[38] Alexander Schrijver. A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *Journal of Combinatorial Theory, Series B*, 80(2):346–355, 2000.

[39] Adish Singla, Sebastian Tschiatschek, and Andreas Krause. Noisy submodular maximization via adaptive sampling with applications to crowdsourced image collection summarization. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

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

[41] Luca Trevisan. Inapproximability of combinatorial optimization problems. *arXiv preprint cs/0409043*, 2004.

[42] Jan Vondrák. *Submodularity in combinatorial optimization*. PhD thesis, Charles University, 2007.

[43] Yi-Jing Wang, Da-Chuan Xu, Yan-Jun Jiang, and Dong-Mei Zhang. Minimizing ratio of monotone non-submodular functions. *Journal of the Operations Research Society of China*, Mar 2019. ISSN 2194-6698. doi: 10.1007/s40305-019-00244-1. URL https://doi.org/10.1007/s40305-019-00244-1.

Appendix A. Appendix

A.1 Proofs of main results

Proposition 7 *Given a set function H, and $\alpha, \beta \in (0, 1]$ such that $\alpha\beta < 1$, there exists a non-decreasing α-weakly DR-submodular function F and a non-decreasing (α, β)-weakly DR-modular function G such that $H(S) = F(S) - G(S)$ for all $S \subseteq V$.*

Proof This decomposition builds on the decomposition of H into the difference of two non-decreasing submodular functions [26]. We start by choosing any function G' which is non-decreasing (α, β)-weakly DR-modular, and is strictly α-weakly DR-submodular, i.e., $\epsilon_{G'} = \min_{i \in V, A \subset B \subseteq V \setminus i} G'(i|A) - \alpha G'(i|B) > 0$. It is always possible to find such a function: We provide an example in Proposition 8 for $\alpha = 1$. For $\alpha < 1$, we can simply use $G'(S) = |S|$. It is not possible though to choose G' such that $\alpha = \beta = 1$ (this would imply $G'(i|B) \geq G'(i|A) > G'(i|B)$). We construct F and G based on G'.
Let $\epsilon_H = \min_{i \in V, A \subseteq B \subseteq V \setminus i} H(i|A) - \alpha H(i|B) < 0$ the violation of α-weak DR-submodularity of H; we may use a lower bound $\epsilon'_H \leq \epsilon_H$. We define $F'(S) = H(S) + \frac{|\epsilon'_H|}{\epsilon_{G'}} G'(S)$, then $F'(i|S) \geq \alpha F'(i|T), \forall i \in V, S \subset T \subseteq V \setminus i$, but not necessarily for $S = T$ since F' is not necessarily non-decreasing. To correct for that, let $V^- = \{i : F'(i|V \setminus i) < 0\}$ and define $F(S) = F'(S) - \sum_{i \in S \cap V^-} F'(i|V \setminus i)$. For all $i \in V, S \subseteq V \setminus i$, if $i \notin V^-$ then $F(i|S) = F'(i|S) \geq \alpha F'(i|V \setminus i) \geq 0$, otherwise $F(i|S) = F'(i|S) - F'(i|V \setminus i) \geq (\alpha - 1)F'(i|V \setminus i) \geq 0$ for $S \neq V \setminus i$ and $F(i|V \setminus i) = 0$. F is thus non-decreasing α-weakly DR-submodular. We also define $G(S) = \frac{|\epsilon'_H|}{\epsilon_{G'}} G'(S) - \sum_{i \in S \cap V^-} F'(i|V \setminus i)$, then $H(S) = F(S) - G(S)$ and G is non-decreasing (α, β)-weakly DR-modular. ■

Proposition 8 *Given $\beta \in (0, 1)$, let $G'(S) = g(|S|)$ where $g(x) = \frac{1}{2}ax^2 + (1 - \frac{1}{2}a)x$ with $a = \frac{\beta - 1}{d - 1}$. Then G' is non-decreasing $(1, \beta)$-weakly DR-modular, and is strictly submodular, with $\epsilon_{G'} = \min_{i \in V, A \subset B \subseteq V \setminus i} G'(i|A) - G'(i|B) = -a > 0$.*

Proof g is a concave function, since $a < 0$, hence $G'(S)$ is submodular. It also follows that

$$\begin{aligned}
\max_{i \in V, S \subseteq T \subseteq V \setminus i} \frac{G'(i|S)}{G'(i|T)} &= \max_{i \in V, S \subseteq T \subseteq V \setminus i} \frac{G'(i)}{G'(i|V \setminus i)} \\
&= \frac{\frac{1}{2}a + (1 - \frac{1}{2}a)}{\frac{1}{2}a(d^2 - (d-1)^2) + (1 - \frac{1}{2}a)(d - (d-1))} \\
&= \frac{1}{\frac{1}{2}a(2d - 2) + 1} \\
&= \frac{1}{\beta}.
\end{aligned}$$

We also have

$$\begin{aligned}
\epsilon_{G'} &= \min_{i \in V, S \subset T \subseteq V \setminus i} G'(i|S) - G'(i|T) \\
&= \min_{T \subset V} 2g(|T|) - g(|T|-1) - g(|T|+1) \\
&= \min_{T \subset V} \tfrac{1}{2}a(2|T|^2 - (|T|-1)^2 - (|T|+1)^2) + (1 - \tfrac{1}{2}a)(2|T| - (|T|-1) - (|T|+1)) \\
&= -a.
\end{aligned}$$

■

Lemma 9 *Given a vector $\boldsymbol{s} \in [0,1]^d$ such that $s_{j_1} \geq \cdots \geq s_{j_d}$, we define $\boldsymbol{\kappa}$ such that $\kappa_{j_k} = H(j_k|S_{k-1})$ where $S_k = \{j_1, \cdots, j_k\}$. Then, $h_L(\boldsymbol{s}) = \boldsymbol{\kappa}^\top \boldsymbol{s} \geq h^-(\boldsymbol{s})$, $\boldsymbol{\kappa}(A) \leq \frac{1}{\alpha}F(A) - \beta G(A)$ for all $A \subseteq V$, and $\boldsymbol{\kappa}^\top \boldsymbol{s}' \leq \frac{1}{\alpha} f^-(\boldsymbol{s}') + \beta(-g)^-(\boldsymbol{s}')$ for all $\boldsymbol{s}' \in [0,1]^d$.*

Proof We use the following formulation of the convex closure [17, Def. 20]:

$$h^-(\boldsymbol{s}) = \max_{\boldsymbol{\kappa} \in \mathbb{R}^d, \rho \in \mathbb{R}} \{\boldsymbol{\kappa}^\top \boldsymbol{s} + \rho : \boldsymbol{\kappa}(A) + \rho \leq H(A), \forall A \subseteq V\},$$

where $\boldsymbol{\kappa}(A) = \sum_{i \in A} x_i$. Given any feasible point $(\boldsymbol{\kappa}', \rho')$ in the definition of h^-, i.e., $\boldsymbol{\kappa}(A) + \rho' \leq H(A), \forall A \subseteq V$, we have:

$$\begin{aligned}
\boldsymbol{\kappa}^\top \boldsymbol{s} - (\boldsymbol{\kappa}'^\top \boldsymbol{s} + \rho') &= \sum_{k=1}^d s_{j_k}(H(j_k|S_{k-1}) - \kappa'_{j_k}) - \rho' \\
&= \sum_{k=1}^{d-1} (s_{j_k} - s_{j_{k+1}}) \big(H(S_k) - \boldsymbol{\kappa}'(S_k)\big) + s_{j_d} \big(H(V) - \boldsymbol{\kappa}'(V)\big) - \rho' \\
&\geq \left(\sum_{k=1}^{d-1} (s_{j_k} - s_{j_{k+1}}) + s_{j_d}\right) \rho' - \rho' \\
&= (s_{j_1} - 1)\rho' \geq 0
\end{aligned}$$

Hence $\boldsymbol{\kappa}^\top \boldsymbol{s} \geq h^-(\boldsymbol{s})$. The last inequality holds by noting that $\rho' \leq 0$ since $\boldsymbol{\kappa}(\emptyset) + \rho' \leq H(\emptyset) = 0$. The upper bound on $\boldsymbol{\kappa}(A)$ for any $A \subseteq V$ follows from the definition of weak DR-submodularity.

$$\begin{aligned}
\boldsymbol{\kappa}(A) &= \sum_{j_k \in A} H(j_k|S_{k-1}) \\
&\leq \sum_{j_k \in A} \frac{1}{\alpha} F(j_k|A \cap S_{k-1}) - \beta G(j_k|A \cap S_{k-1}) \\
&= \sum_{k=1}^d \frac{1}{\alpha}(F(A \cap S_k) - F(A \cap S_{k-1})) - \beta(G(A \cap S_k) - G(A \cap S_{k-1})) \\
&= \frac{F(A)}{\alpha} - \beta G(A)
\end{aligned}$$

Note that $\boldsymbol{\kappa}$ can be written as $\boldsymbol{\kappa} = \boldsymbol{\kappa}^F - \boldsymbol{\kappa}^G$ where $\boldsymbol{\kappa}^F_{j_k} = F(j_k|S_{k-1})$ and $\boldsymbol{\kappa}^G_{j_k} = G(j_k|S_{k-1})$. We have $\boldsymbol{\kappa}^F(A) \leq \frac{F(A)}{\alpha}$ and $\boldsymbol{\kappa}^G(A) \leq \beta G(A), \forall A \subseteq V$. Hence $(\alpha\boldsymbol{\kappa}^F, \mathbf{0})$ and $(\frac{1}{\beta}\boldsymbol{\kappa}^G, \mathbf{0})$ are feasible points in the definitions of f^- and $(-g)^-$. The bound on $\boldsymbol{\kappa}^\top \boldsymbol{s}'$ for any $\boldsymbol{s}' \in [0,1]^d$ then follows directly from the definitions of f^- and $(-g)^-$. ■

Theorem 4 *After T iterations of PGM, $\hat{\mathbf{s}} \in \arg\min_{t \in \{1,\cdots,T\}} h_L(\boldsymbol{s}^t)$ satisfies:*

$$h^-(\hat{\mathbf{s}}) \;\leq\; h_L(\hat{\mathbf{s}}) \;\leq\; \tfrac{1}{\alpha} f^-(\boldsymbol{s}^*) + \beta(-g)^-(\boldsymbol{s}^*) + \tfrac{RL}{\sqrt{T}},$$

where $\boldsymbol{s}^* \in \arg\min_{\boldsymbol{s}\in[0,1]^d} h^-(\boldsymbol{s})$. Let $\hat{S}_k = \{j_1, \cdots, j_k\}$ such that $\hat{s}_{j_1} \geq \cdots \geq \hat{s}_{j_d}$, and $\hat{S}_0 = \emptyset$. Then $\hat{S} \in \arg\min_{k\in\{0,\cdots,d\}} H(\hat{S}_k)$ satisfies

$$H(\hat{S}) \leq \tfrac{1}{\alpha} F(S^*) - \beta G(S^*) + \tfrac{RL}{\sqrt{T}},$$

where $S^* \in \arg\min_{S\subseteq V} H(S)$. This bound is tight even if F and G are weakly DR-modular.

Proof We prove first the bound on $h^-(\hat{\mathbf{s}})$ and $h_L(\hat{\mathbf{s}})$. Let $\boldsymbol{z}^{t+1} = \boldsymbol{s}^t - \eta\boldsymbol{\kappa}^t$, then note that $\|\boldsymbol{s}^{t+1} - \boldsymbol{s}^*\|_2 \leq \|\boldsymbol{z}^{t+1} - \boldsymbol{s}^*\|_2$ due to the properties of projection (see for e.g., [9, Lemma 3.1]), it follows then

$$\begin{aligned}
\langle \boldsymbol{\kappa}^t, \boldsymbol{s}^t - \boldsymbol{s}^* \rangle &= \frac{1}{\eta}\langle \boldsymbol{s}^t - \boldsymbol{z}^{t+1}, \boldsymbol{s}^t - \boldsymbol{s}^* \rangle \\
&= \frac{1}{2\eta}(\|\boldsymbol{s}^t - \boldsymbol{z}^{t+1}\|_2^2 + \|\boldsymbol{s}^t - \boldsymbol{s}^*\|_2^2 - \|\boldsymbol{z}^{t+1} - \boldsymbol{s}^*\|_2^2) \\
&= \frac{1}{2\eta}(\|\boldsymbol{s}^t - \boldsymbol{s}^*\|_2^2 - \|\boldsymbol{z}^{t+1} - \boldsymbol{s}^*\|_2^2) + \frac{\eta}{2}\|\boldsymbol{\kappa}^t\|_2^2 \\
&\leq \frac{1}{2\eta}(\|\boldsymbol{s}^t - \boldsymbol{s}^*\|_2^2 - \|\boldsymbol{s}^{t+1} - \boldsymbol{s}^*\|_2^2) + \frac{\eta}{2}\|\boldsymbol{\kappa}^t\|_2^2
\end{aligned}$$

Summing over t we get

$$\sum_{t=1}^{T} \langle \boldsymbol{\kappa}^t, \boldsymbol{s}^t - \boldsymbol{s}^* \rangle \leq T\frac{R^2}{2\eta} + \frac{\eta T L^2}{2}$$

Since F is α-weakly DR submodular and $-G$ is $\frac{1}{\beta}$-weakly DR submodular, we have by lemma 3 for all t, $(\boldsymbol{\kappa}^t)^\top \boldsymbol{s}^* \leq \frac{f^-(\boldsymbol{s}^*)}{\alpha} + \beta(-g)^-(\boldsymbol{s}^*)$ and $(\boldsymbol{\kappa}^t)^\top \boldsymbol{s}^t = h_L(\boldsymbol{s}^t) \geq h^-(\boldsymbol{s}^t)$. Plugging in the value of η, we thus obtain

$$\min_t h^-(\boldsymbol{s}^t) \leq \min_t h_L(\boldsymbol{s}^t) \leq \frac{f^-(\boldsymbol{s}^*)}{\alpha} + \beta(-g)^-(\boldsymbol{s}^*) + \frac{RL}{\sqrt{T}}. \tag{3}$$

By definition of the Lovász extension, we have:

$$h_L(\hat{s}) = \sum_{k=1}^{d}(\hat{s}_{j_k} - \hat{s}_{j_{k+1}})H(\hat{S}_k) + \hat{s}_{j_d}H(V) \geq \min_{k\in\{0,\cdots,d\}} H(\hat{S}_k).$$

The bound on $H(\hat{S})$ then follows from Eq. (3), and the extension property $f^-(s^*) = F(S^*), (-g)^-(s^*) = -G(S^*)$.
To show that this approximation is tight, we construct the following example. Let $H(S) = F(S) - G(S)$ such that $F(S) = |S| + \frac{d}{\beta} - 1$ if $1 \in S$, $F(S) = \alpha|S|$ otherwise and $G(S) = |S| + \frac{d}{\beta} - 1$ if $1 \in S$, $G(S) = \frac{1}{\beta}|S|$ otherwise, for some $\alpha, \beta \in (0, 1)$. Then, F is monotone $(\alpha, 1)$-weakly DR-modular and G is monotone (1,β)-weakly DR-modular. The solution obtained by PGM have value $H(\hat{S}) = 0$, while the optimal solution $S^* = V \setminus \{1\}$ have value $H(S^*) = (\alpha - \frac{1}{\beta})(n-1) < 0$. Hence, $H(\hat{S}) = \frac{F(S^*)}{\alpha} - \beta G(S^*) = \frac{\alpha(n-1)}{\alpha} - \beta\frac{(n-1)}{\beta}$.
It's easy to see that both F and G are monotone functions. For all $A \subseteq B, i \in V \setminus B$, we have

$$\frac{F(i|A)}{F(i|B)} = \begin{cases} 1 & \text{if } 1 \in A \text{ or } 1 \notin B \\ \alpha & \text{if } 1 \notin A, 1 \in B \\ \frac{\frac{d}{\beta} + (1-\alpha)|A|}{\frac{d}{\beta} + (1-\alpha)|B|} & \text{if } i = 1 \end{cases}$$

Note that $\frac{d}{\beta} + (1-\alpha)|A| \geq \frac{d}{\beta} \geq \alpha(\frac{d}{\beta} + (1-\alpha)|B|)$, hence $\alpha \leq \frac{F(i|A)}{F(i|B)} \leq 1$, which proves that F is supermodular and α-weakly DR-submodular.
Similarly we have

$$\frac{F(i|A)}{F(i|B)} = \begin{cases} 1 & \text{if } 1 \in A \text{ or } 1 \notin B \\ \frac{1}{\beta} & \text{if } 1 \notin A, 1 \in B \\ \frac{\frac{d}{\beta} + (1-\frac{1}{\beta})|A|}{\frac{d}{\beta} + (1-\frac{1}{\beta})|B|} & \text{if } i = 1 \end{cases}$$

Note that $\frac{d}{\beta} + (1 - \frac{1}{\beta})|A| \leq \frac{d}{\beta} \leq \frac{1}{\beta}(\frac{d}{\beta} + (1 - \frac{1}{\beta})|B|)$, hence $1 \leq \frac{F(i|A)}{F(i|B)} \leq \frac{1}{\beta}$, which proves that G is monotone submodular and β-weakly DR-supermodular.
It remains to show that the solution obtained by PGM and thresholding have value $H(\hat{S}) = 0$. We can assume w.l.o.g that the starting point $\boldsymbol{s}^1$ is such that the largest element is $j_1 = 1$ (otherwise we can modify the example to have whatever is the largest element as the "bad element"). Note that $H(1) = H(j_k|S_k) = 0$ for all $k \in [d]$, hence $\boldsymbol{\kappa}^1 = \mathbf{0}$ and $\boldsymbol{s}^t = \boldsymbol{s}^1$ and $\boldsymbol{\kappa}^t = \mathbf{0}$ for all $t \in \{1, \cdots, T\}$. Thresholding $\boldsymbol{s}^1$ would thus yield $H(\hat{S}) = 0$, with $\hat{S} = \emptyset$ or any other set such that $1 \in \hat{S}$. ■

Proposition 10 *Assume we have an approximate oracle $\tilde{H}$ with input parameters $\epsilon, \delta \in (0, 1)$, such that for every $S \subseteq V$, $|\tilde{H}(S) - H(S)| \leq \epsilon$ with probability $1 - \delta$. We run PGM with $\tilde{H}$ for T iterations. Let $\hat{\boldsymbol{s}} = \arg\min_{t \in \{1, \cdots, T\}} \tilde{h}_L(\boldsymbol{s}^t)$, and $\hat{S}_k = \{j_1, \cdots, j_k\}$ such that $\hat{s}_{j_1} \geq \cdots \geq \hat{s}_{j_d}$. Then $\hat{S} \in \arg\min_{k \in \{0, \cdots, d\}} \tilde{H}(\hat{S}_k)$ satisfies*

$$H(\hat{S}) \leq \tfrac{1}{\alpha} F(S^*) - \beta G(S^*) + \epsilon',$$

with probability $1 - \delta'$, by choosing $\epsilon = \frac{\epsilon'}{8d}$, $\delta = \frac{\delta' \epsilon'^2}{32 d^2}$ and $T = (4\sqrt{d}L/\epsilon')^2$ iterations.

Proof Let $\boldsymbol{\kappa}$ be defined as $\kappa_{j_k} = H(j_k|S_{k-1})$ and $\tilde{\boldsymbol{\kappa}}$ as $\tilde{\kappa}_{j_k} = \tilde{H}(j_k|S_{k-1})$. For all $k \in V$, we have $|\tilde{\kappa}_{j_k} - \kappa_{j_k}| \leq 2\epsilon$ with probability $1 - 2d\delta$ (by a union bound). Hence, for every $S \subseteq V$, we have

$|\tilde{\boldsymbol{\kappa}}(S) - \boldsymbol{\kappa}(S)| \leq 2\epsilon|S|$. Plugging this into the proof of Theorem 4 directly yields

$$H(\hat{S}) \leq \tfrac{F(S^*)}{\alpha} - \beta G(S^*) + 2\epsilon(|S^*| + 1) + \frac{RL}{\sqrt{T}}$$

The proposition follows by setting ϵ, δ and T to the chosen values. ■

Theorem 6 *For any $\alpha, \beta \in (0, 1]$ such that $\alpha\beta < 1, d > 2$ and $\delta > 0$, there are instances of Problem (1) such that no (deterministic or randomized) algorithm, using less than exponentially many queries, can always find a solution $S \subseteq V$ of expected value at most $\frac{1}{\alpha}F(S^*) - \beta G(S^*) - \delta$.*

Proof We use a similar proof technique to [19]. Let C, D be two sets that partition the ground set $V = C \cup D$ such that $|C| = |D| = d/2$. We construct a normalized set function H whose values depend only on $k(S) = |S \cap C|$ and $\ell(S) = |S \cap D|$. In particular, we define

$$H(S) = \begin{cases} 0 & \text{if } |k(S) - \ell(S)| \leq \epsilon d \\ \frac{2\alpha\delta}{2-d} & \text{otherwise} \end{cases},$$

for some $\epsilon \in [1/d, 1/2)$. By Proposition 2, given a non-decreasing (α, β)-weakly DR-modular function G', we can write $H(S) = F(S) - G(S)$, where $F(S) = H(S) + \frac{|\epsilon_H|}{\epsilon_{G'}}G'(S)$ is normalized non-decreasing α-weakly DR-submodular, and $G(S) = \frac{|\epsilon_H|}{\epsilon_{G'}}G'(S)$ is normalized non-decreasing (α, β)-weakly DR-modular. Note that $V^- = \emptyset$ in this case, since $H(i|V \setminus i) = 0$. We choose $G'(S) = |S|$ if $\alpha < 1$, then $\epsilon_{G'} = \min_{i \in V, S \subset T \subseteq T \setminus i} G(i|S) - \alpha G(i|T) = 1 - \alpha > 0$. If $\alpha = 1$, we use the $(1, \beta)$-weakly DR-modular function defined in Proposition 8, then $\epsilon_{G'} = \frac{1-\beta}{d-1} > 0$.
Let the partition (C, D) be random and unknown to the algorithm. We argue that, with high probability, any given query S will be "balanced", i.e., $|k(S) - \ell(S)| \leq \epsilon d$. Hence no deterministic algorithm can distinguish between H and the constant zero function. Given a fixed $S \subseteq V$, let $X_i = 1$ if $i \in C$ and 0 otherwise, for all $i \in S$, then $\mu = \mathbb{E}[\sum_{i \in S} X_i] = \sum_{i \in S} \frac{|C|}{d} = \frac{|S|}{2}$. Then by a Chernoff's bound we have $Pr(|k(S) - \ell(S)| > \epsilon d) \leq 2\exp(-\frac{\epsilon^2 d}{4})$. Hence, given a sequence of $e^{\frac{\epsilon^2 d}{8}}$ many queries, the probability that each query S is balanced, and thus have value $H(S) = 0$, is still at least $1 - 2e^{-\frac{\epsilon^2 d}{8}}$. On the other hand, we have $H(S^*) = \frac{2\alpha\delta}{2-d} < 0$, achieved at $S^* = C$ or D. Moreover, note that $\epsilon_H = \min_{i \in V, S \subseteq T \subseteq T \setminus i} H(i|S) - \alpha H(i|T) = (1 + \alpha)H(S^*)$. Hence

$$\tfrac{1}{\alpha}F(S^*) - \beta G(S^*) - \delta = \tfrac{1}{\alpha}H(S^*)\left(1 - (1 - \alpha\beta)(1 + \alpha)\frac{G'(S^*)}{\epsilon_{G'}}\right) - \delta < 0,$$

since $\frac{G'(S^*)}{\epsilon_{G'}} = \frac{d}{2(1-\alpha)}$ if $\alpha < 1$, and $\frac{G'(S^*)}{\epsilon_{G'}} \geq \frac{3d(d-1)}{8(1-\beta)}$, if $\alpha = 1$.
Therefore, with high probability, the algorithm cannot find a set with value $H(S) \leq \frac{1}{\alpha}F(S^*) - \beta G(S^*) - \delta$. This also holds for a randomized algorithm, by averaging over its random choices. ■

A.2 Proofs for structured sparse learning application We prove that the auxiliary function G^ℓ in structured sparse learning problems is weakly DR-modular for all sets of cardinality k, when ℓ has ν-restricted smoothness (RSM) and μ-restricted strong convexity (RSC).
Let's recall the definition of RSC/RSM.

Definition 11 (RSM/RSC) *Given a differentiable function $\ell : \mathbb{R}^d \to \mathbb{R}$ and $\Omega \subset \mathbb{R}^d \times \mathbb{R}^d$, ℓ is μ_Ω-RSC and ν_Ω-RSM if $\frac{\mu_\Omega}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \leq \ell(\boldsymbol{y}) - \ell(\boldsymbol{x}) - \langle \nabla\ell(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle \leq \frac{\nu_\Omega}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2, \ \forall(\boldsymbol{x},\boldsymbol{y}) \in \Omega$.*

If ℓ is RSC/RSM on $\Omega = \{(\boldsymbol{x},\boldsymbol{y}) : \|\boldsymbol{x}\|_0 \leq k_1, \|\boldsymbol{y}\|_0 \leq k_1, \|\boldsymbol{x}-\boldsymbol{y}\|_0 \leq k_2\}$, we denote by $\mu_{k_1,k_2}, \nu_{k_1,k_2}$ the corresponding RSC and RSM parameters. For simplicity, we also define $\mu_k := \mu_{k,k}, \nu_k := \mu_{k,k}$.
Before we can prove Proposition 14, we need two key lemmas. Lemma 12 restates a result from [18], which relates the marginal gain of G^ℓ to the marginal decrease in ℓ. In Lemma 13, we argue that for a class of loss functions, namely RSC/RSM functions of the form $\ell(\boldsymbol{x}) = L(\boldsymbol{x}) - \boldsymbol{z}^\top\boldsymbol{x}$, where $\boldsymbol{z}$ is a random vector, the corresponding minimizer has full support with probability one. Proposition 14 then follows from these two lemmas by noting that ℓ thus have non-zero marginal decrease, with respect to any $i \in V$, with probability one.

Lemma 12 *Given $G^\ell(S) = \ell(0) - \min_{\text{supp}(\boldsymbol{x})\subseteq S} \ell(\boldsymbol{x})$, then for any disjoint sets $A, B \subseteq V$ and a corresponding minimizer $\boldsymbol{x}^A := \arg\min_{\text{supp}(\boldsymbol{x})\subseteq A} \ell(\boldsymbol{x})$, if ℓ is $\mu_{|A\cup B|}$-RSC and $\nu_{|A|,|B|}$-RSM, we have:*

$$\frac{\|[\nabla\ell(\boldsymbol{x}^A)]_B\|_2^2}{2\nu_{|A\cup B|,|B|}} \leq G^\ell(B|A) \leq \frac{\|[\nabla\ell(\boldsymbol{x}^A)]_B\|_2^2}{2\mu_{|A\cup B|}}$$

Lemma 13 *If $\boldsymbol{x}^\star$ is the minimizer of $\min_{x\in\mathbb{R}^d} L(\boldsymbol{x}) - \boldsymbol{z}^\top\boldsymbol{x}$, where L is a strongly-convex and smooth loss function, and $\boldsymbol{z} \in \mathbb{R}^d$ has a continuous density w.r.t to the Lebesgue measure, then $\boldsymbol{x}^\star$ has full support with probability one.*

Proof This follows directly from [16, Theorem 1] by taking $\Phi(x) = 0$. We include the proof here for completeness.

Since L is strongly-convex, given $\boldsymbol{z}$ the corresponding minimizer $\boldsymbol{x}^\star$ is unique, then the function $E(\boldsymbol{z}) := \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} L(\boldsymbol{x}) - \boldsymbol{z}^T\boldsymbol{x}$ is well defined. Given fixed $i \in V$, we show that the set $S_i := \{\boldsymbol{z} : [E(\boldsymbol{z})]_i = 0\}$ has measure zero. Then, taking the union of the finitely many sets $S_i, \forall i \in V$, all of zero measure, we have $P(\exists \boldsymbol{z} \in \mathbb{R}^d, \exists i \in V, \text{ s.t., } [E(\boldsymbol{z})]_i = 0) = 0$.

To show that the set S_i has measure zero, let $\boldsymbol{z}_1, \boldsymbol{z}_2 \in S_i$ and denote by $\mu > 0$ the strong convexity constant of L. We have by optimality conditions:

$$\Big((\boldsymbol{z}_1 - \nabla L(E(\boldsymbol{z}_1))) - (\boldsymbol{z}_2 - \nabla L(E(\boldsymbol{z}_2)))\Big)^\top \Big(E(\boldsymbol{z}_1) - E(\boldsymbol{z}_2)\Big) = 0$$

Hence,

$$\begin{aligned}
(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top(E(\boldsymbol{z}_1) - E(\boldsymbol{z}_2)) &\geq \big(\nabla L(E(\boldsymbol{z}_1)) - \nabla L(E(\boldsymbol{z}_2))\big)^\top \big(E(\boldsymbol{z}_1) - E(\boldsymbol{z}_2)\big) \\
(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top(E(\boldsymbol{z}_1) - E(\boldsymbol{z}_2)) &\geq \mu\|E(\boldsymbol{z}_1) - E(\boldsymbol{z}_2)\|_2^2 \\
\frac{1}{\mu}\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2 &\geq \|E(\boldsymbol{z}_1) - E(\boldsymbol{z}_2)\|_2
\end{aligned}$$

Thus E is a deterministic Lipschitz-continuous function of $\boldsymbol{z}$. By optimality conditions $\boldsymbol{z} = \nabla L(E(\boldsymbol{z}))$, then $z_i = \nabla L(E(\boldsymbol{z}_{V \setminus i}))_i$. Thus z_i is a Lipschitz-continuous function of $\boldsymbol{z}_{V \setminus i}$, which can only happen with zero measure. ■

Proposition 14 *Let $\ell(\boldsymbol{x}) = L(\boldsymbol{x}) - \boldsymbol{z}^\top \boldsymbol{x}$, where L is $\mu_{|U|}$-RSC and $\nu_{|U|}$-RSM for some $U \subseteq V$ and $\boldsymbol{z} \in \mathbb{R}^d$ has a continuous density w.r.t the Lebesgue measure. Then there exist $\alpha_G, \beta_G > 0$ such that G^ℓ is (α_G, β_G)-weakly DR modular on U (i.e., Def. 1 restricted to sets $A \subseteq B \subseteq U$).*

Proof Given $S \subseteq U, i \in U \setminus S$, let $\boldsymbol{x}^S := \arg\min_{\text{supp}(\boldsymbol{x}) \subseteq S} \ell(\boldsymbol{x})$, then by Lemma 12 and $\nu_{|S|+1,1} \leq \nu_{|S|+1}$ we have:

$$\frac{[\nabla \ell(\boldsymbol{x}^S)]_i^2}{2\nu_{|S|+1}} \leq G^\ell(i|S) \leq \frac{[\nabla \ell(\boldsymbol{x}^S)]_i^2}{2\mu_{|S|+1}}$$

We argue that $[\nabla \ell(x^S)]_i^2 \neq 0$ with probability one. For that, we define $\ell_S(\boldsymbol{u}) := \ell(\boldsymbol{x})$, where $[\boldsymbol{x}]_S = \boldsymbol{u}, [\boldsymbol{x}]_{V \setminus S} = 0, \forall \boldsymbol{u} \in \mathbb{R}^{|S|}$, then ℓ_S is $\mu_{|S|}$-strongly convex and $\nu_{|S|}$-smooth on $\mathbb{R}^{|S|}$. Hence, by lemma 13, the minimizer $\boldsymbol{u}^\star$ of ℓ_S has full support with probability one, and thus $\text{supp}(\boldsymbol{x}^S) = S$ also with probability one. By the same argument, we have $\text{supp}(\boldsymbol{x}^{S \cup \{i\}}) = S \cup \{i\}$. We can thus deduce that $[\nabla \ell(\boldsymbol{x}^S)]_i^2 \neq 0$, since otherwise $G^\ell(i|S) = 0$, which implies that $\boldsymbol{x}^{S \cup \{i\}} = x^S$ (minimizer is unique) and $\text{supp}(\boldsymbol{x}^{S \cup \{i\}}) = S$, which happens with probability zero.
For all $S \subseteq T \subseteq U, i \in U \setminus T$, the following bounds hold:

$$\frac{\mu_{|T|+1}[\nabla \ell(\boldsymbol{x}^S)]_i^2}{\nu_{|S|+1}[\nabla \ell(\boldsymbol{x}^T)]_i^2} \leq \frac{G^\ell(i|S)}{G^\ell(i|T)} \leq \frac{\nu_{|T|+1}[\nabla \ell(\boldsymbol{x}^S)]_i^2}{\mu_{|S|+1}[\nabla \ell(\boldsymbol{x}^T)]_i^2}$$

G^ℓ is then (α_G, β_G)-weakly DR-modular with $\alpha_G := \min_{S \subseteq T \subseteq U, i \in U \setminus T} \frac{\mu_{|T|+1}[\nabla \ell(\boldsymbol{x}^S)]_i^2}{\nu_{|S|+1}[\nabla \ell(\boldsymbol{x}^T)]_i^2} > 0$ and $\beta_G := \min_{S \subseteq T \subseteq U, i \in U \setminus T} \frac{\mu_{|S|+1}[\nabla \ell(\boldsymbol{x}^T)]_i^2}{\nu_{|T|+1}[\nabla \ell(\boldsymbol{x}^S)]_i^2} > 0$. ■

A.3 Proofs for structured batch Bayesian optimization application Let f be modeled by a Gaussian process with zero mean and kernel function $k(\boldsymbol{x}, \boldsymbol{x}')$, and we observe noisy evaluations $y = f(\boldsymbol{x}) + \epsilon$ of the function, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Given a set $\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_d\}$ of potential maximizers of f, each $\boldsymbol{x}_i \in \mathbb{R}^n$, and a set $S \subseteq V$, let $\boldsymbol{y}_S = [y_i]_{i \in S}$ be the corresponding observations at points $x_i, i \in S$. The posterior distribution of f given $\boldsymbol{y}_S$ is again a Gaussian process, with posterior covariance $k_S(\boldsymbol{x}, \boldsymbol{x}')$, and variance $\sigma_S^2(\boldsymbol{x})$:

$$\begin{aligned} k_S(\boldsymbol{x}, \boldsymbol{x}') &= k(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{k}_S(\boldsymbol{x})^\top (\boldsymbol{K}_S + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{k}_S(\boldsymbol{x}'), \\ \sigma_S^2(\boldsymbol{x}) &= k_S(\boldsymbol{x}, \boldsymbol{x}), \end{aligned}$$

where $\boldsymbol{k}_S = [k(\boldsymbol{x}_i, \boldsymbol{x})]_{i \in S}$, and $\boldsymbol{K}_S = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j \in S}$ is the corresponding submatrix of the positive definite kernel matrix $\boldsymbol{K}$. The variance reduction function is defined as:

$$G(S) = \sum_{i \in V} \sigma^2(\boldsymbol{x}_i) - \sigma_S^2(\boldsymbol{x}_i),$$

where $\sigma^2(\boldsymbol{x}_i) = k(\boldsymbol{x}_i, \boldsymbol{x}_i)$. We show that G is weakly DR-modular. To do that, we first show that the objective in noisy column subset selection problems is weakly DR-modular, generalizing the result of [40]. We then show that the variance reduction function can be written as a noisy column subset selection objective.

We start by giving explicit expressions for the marginals of the objective in noisy column subset selection problems.

Proposition 15 *Let $\ell(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{\sigma^2}{2}\|\boldsymbol{x}\|^2$ for some $\sigma \geq 0$ and $G(S) = \ell(0) - \min_{\mathrm{supp}(\boldsymbol{x})\subseteq S} \ell(\boldsymbol{x})$, then*

$$G(i|S) = [\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_i^2 \, \phi(S,i) = \Big(\frac{\boldsymbol{y}^\top R^S(\boldsymbol{a}_i)}{2\sqrt{\phi(S,i)}}\Big)^2,$$

where $\boldsymbol{a}_i$ is the ith column of $\boldsymbol{A}$, $\boldsymbol{x}^S(\boldsymbol{a}_i) := \arg\min_{\mathrm{supp}(\boldsymbol{x})\subseteq S} \frac{1}{2}\|\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{\sigma^2}{2}\|\boldsymbol{x}\|^2$ is the vector of optimal regression coefficients, $R^S(\boldsymbol{a}_i) = \boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x}^S(\boldsymbol{a}_i)$ the corresponding residual, and $\phi(S,i) = \frac{1}{2}\|R^S(\boldsymbol{a}_i)\|^2 + \frac{\sigma^2}{2}\|\boldsymbol{x}^S(\boldsymbol{a}_i)\|^2 + \frac{\sigma^2}{2}$.

Proof Given $i \in V, S \subseteq V \setminus i$, let $\boldsymbol{x}^S(\boldsymbol{y}) := \arg\min_{\mathrm{supp}(\boldsymbol{x})\subseteq S} \ell(\boldsymbol{x})$, and $R^S(\boldsymbol{y}) = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^S(\boldsymbol{y})$ the corresponding residual. Let $\gamma = [\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_i$, then we can write

$$\begin{aligned}
\boldsymbol{y} &= \boldsymbol{A}\boldsymbol{x}^{S\cup i}(\boldsymbol{y}) + R^{S\cup i}(\boldsymbol{y}) \\
&= \boldsymbol{A}_S[\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_S + \boldsymbol{a}_i\gamma + R^{S\cup i}(\boldsymbol{y}) \\
&= \boldsymbol{A}_S([\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_S + \gamma\boldsymbol{x}^S(\boldsymbol{a}_i)) + \gamma R^S(\boldsymbol{a}_i) + R^{S\cup i}(\boldsymbol{y})
\end{aligned}$$

By optimality conditions we have $-\boldsymbol{A}_{S\cup i}^\top R^{S\cup i}(\boldsymbol{y}) + \sigma^2\boldsymbol{x}^{S\cup i}(\boldsymbol{y}) = 0$ and $-\boldsymbol{A}_S^\top R^S(\boldsymbol{a}_i) + \sigma^2\boldsymbol{x}^S(\boldsymbol{a}_i) = 0$. Let $\hat{\boldsymbol{x}}^S(\boldsymbol{y}) = [\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_S + \gamma\boldsymbol{x}^S(\boldsymbol{a}_i)$, then $\hat{\boldsymbol{x}}^S(\boldsymbol{y})$ satisfies the constraint $\mathrm{supp}(\hat{\boldsymbol{x}}^S(\boldsymbol{y})) = S$ and the optimality condition $\boldsymbol{A}_S^\top(\boldsymbol{A}_S\hat{\boldsymbol{x}}^S(\boldsymbol{y}) - \boldsymbol{y}) + \sigma^2\hat{\boldsymbol{x}}^S(\boldsymbol{y}) = 0$. We can see this by plugging in the expression for $\boldsymbol{y}$ and using the optimality conditions on $\boldsymbol{x}^{S\cup i}(\boldsymbol{y})$ and $\boldsymbol{x}^S(\boldsymbol{a}_i)$.

$$\boldsymbol{A}_S^\top(\boldsymbol{A}_S\hat{\boldsymbol{x}}^S(\boldsymbol{y}) - \boldsymbol{y}) + \sigma^2\hat{\boldsymbol{x}}^S(\boldsymbol{y}) = -\boldsymbol{A}_S^\top(\gamma R^S(\boldsymbol{a}_i) + R^{S\cup i}(\boldsymbol{y})) + \sigma^2([\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_S + \gamma\boldsymbol{x}^S(\boldsymbol{a}_i)) = 0$$

Hence $\hat{\boldsymbol{x}}^S(\boldsymbol{y}) = \boldsymbol{x}^S(\boldsymbol{y})$. By the optimality condition on $\boldsymbol{x}^{S\cup i}(\boldsymbol{y})$, we also have

$$\begin{aligned}
R^S(\boldsymbol{a}_i)^\top R^{S\cup i}(\boldsymbol{y}) &= \boldsymbol{a}_i^\top R^{S\cup i}(\boldsymbol{y}) - \boldsymbol{x}^S(\boldsymbol{a}_i)^\top\boldsymbol{A}_S^\top R^{S\cup i}(\boldsymbol{y}) \\
&= \sigma^2\gamma - \sigma^2\boldsymbol{x}^S(\boldsymbol{a}_i)^\top[\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_S
\end{aligned}$$

The marginals are thus given by

$$\begin{aligned}
G(i|S) &= \ell(\boldsymbol{x}^S(\boldsymbol{y})) - \ell(\boldsymbol{x}^{S\cup i}(\boldsymbol{y})) \\
&= \frac{1}{2}\|\gamma R^S(\boldsymbol{a}_i) + R^{S\cup i}(\boldsymbol{y})\|_2^2 + \frac{\sigma^2}{2}\|[\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_S + \gamma\boldsymbol{x}^S(\boldsymbol{a}_i)\|^2 - \frac{1}{2}\|R^{S\cup i}(\boldsymbol{y})\|_2^2 - \frac{\sigma^2}{2}\|\boldsymbol{x}^{S\cup i}(\boldsymbol{y})\|^2 \\
&= \frac{\gamma^2}{2}\|R^S(\boldsymbol{a}_i)\|^2 + \sigma^2\gamma^2 - \sigma^2\gamma\boldsymbol{x}^S(\boldsymbol{a}_i)^\top[\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_S + \frac{\sigma^2}{2}\gamma^2\|\boldsymbol{x}^S(\boldsymbol{a}_i)\|^2 + \sigma^2\gamma\boldsymbol{x}^S(\boldsymbol{a}_i)^\top[\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_S - \frac{\sigma^2}{2}\gamma^2 \\
&= \gamma^2(\frac{1}{2}\|R^S(\boldsymbol{a}_i)\|^2 + \frac{\sigma^2}{2}\|\boldsymbol{x}^S(\boldsymbol{a}_i)\|^2 + \frac{\sigma^2}{2})
\end{aligned}$$

Hence $G(i|S) = [\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_i^2 \phi(S,i)$.
By the optimality condition on $\boldsymbol{x}^S(\boldsymbol{a}_i)$ we also have:

$$\begin{aligned}
\frac{1}{2}\boldsymbol{y}^\top R^S(\boldsymbol{a}_i) &= \frac{1}{2}(R^{S\cup i}(\boldsymbol{y}) + \boldsymbol{A}_{S\cup i}\boldsymbol{x}^{S\cup i}(\boldsymbol{y}))^\top R^S(\boldsymbol{a}_i) \\
&= \frac{1}{2}\Big(\sigma^2\gamma - \sigma^2 \boldsymbol{x}^S(\boldsymbol{a}_i)^\top [\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_S + [\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_S^\top \boldsymbol{A}_S^\top R^S(\boldsymbol{a}_i) + [\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_i^\top \boldsymbol{a}_i^\top R^S(\boldsymbol{a}_i)\Big) \\
&= \frac{1}{2}\Big(\sigma^2\gamma - \sigma^2 \boldsymbol{x}^S(\boldsymbol{a}_i)^\top [\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_S + \sigma^2[\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_S^\top \boldsymbol{x}^S(\boldsymbol{a}_i) + [\boldsymbol{x}^{S\cup i}(\boldsymbol{y})]_i (R^S(\boldsymbol{a}_i) + \boldsymbol{A}_S\boldsymbol{x}^S(\boldsymbol{a}_i))^\top R^S(\boldsymbol{a}_i \\
&= \frac{1}{2}\Big(\sigma^2\gamma + \gamma\|R^S(\boldsymbol{a}_i)\|_2^2 + \gamma\sigma^2\|\boldsymbol{x}^S(\boldsymbol{a}_i)\|_2^2\Big)
\end{aligned}$$

Hence $\Big(\frac{\boldsymbol{y}^\top R^S(\boldsymbol{a}_i)}{2\sqrt{\phi(S,i)}}\Big)^2 = \gamma^2\phi(S,i) = G(i|S)$. ■

Proposition 16 *Given a positive-definite matrix $\boldsymbol{A}$, let $\boldsymbol{a}_i$ be the ith column of $\boldsymbol{A}$, and $\ell_i(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{\sigma^2}{2}\|\boldsymbol{x}\|^2$ for all $i \in V$, for some $\sigma \geq 0$. Then the function $G(S) = \sum_{i\in V}\ell(0) - \min_{\mathrm{supp}(\boldsymbol{x})\subseteq S}\ell_i(\boldsymbol{x})$ is a non-decreasing (β,β)-weakly DR-modular function, with $\beta = (\frac{\lambda_{\min}(\boldsymbol{A})}{\sigma^2+\lambda_{\min}(\boldsymbol{A})})^2\frac{1}{\kappa^2(\boldsymbol{A})}$, where $\kappa(\boldsymbol{A}) = \lambda_{\max}(\boldsymbol{A})/\lambda_{\min}(\boldsymbol{A})$ is the condition number of $\boldsymbol{A}$.*

Proof For all $j \in V$, let $G_j(S) := \ell(0) - \min_{\mathrm{supp}(\boldsymbol{x})\subseteq S}\ell_j(\boldsymbol{x})$, then we can write $G(S) = \sum_{j\in V} G_j(S)$. Given $i \in V, S \subseteq V \setminus i$, let $\boldsymbol{x}^S(\boldsymbol{a}_i) := \arg\min_{\mathrm{supp}(\boldsymbol{x})\subseteq S}\ell_i(\boldsymbol{x})$ be the optimal regression coefficients, $R^S(\boldsymbol{a}_i) = \boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x}^S(\boldsymbol{a}_i)$ the corresponding residual. By Proposition 15, we have for all $j \in V$:

$$G_j(i|S) = [\boldsymbol{x}^{S\cup i}(\boldsymbol{a}_j)]_i^2\ \phi(S,i) = \Big(\frac{\boldsymbol{a}_j^\top R^S(\boldsymbol{a}_i)}{2\sqrt{\phi(S,i)}}\Big)^2,$$

where $\phi(S,i) = \frac{1}{2}\|R^S(\boldsymbol{a}_i)\|^2 + \frac{\sigma^2}{2}\|\boldsymbol{x}^S(\boldsymbol{a}_i)\|^2 + \frac{\sigma^2}{2}$. Note that $\phi(S,i) > 0$ since $\boldsymbol{A}$ is positive definite (columns are linearly independent).
In the noiseless case $\sigma = 0$, we have $\boldsymbol{x}^{S\cup i}(\boldsymbol{a}_i) = \mathbb{1}_i$. In the noisy case $\sigma > 0$, we have by optimality conditions

$$\begin{aligned}
(\boldsymbol{A}_{S\cup i}^\top \boldsymbol{A}_{S\cup i} + \sigma^2 I)\boldsymbol{x}^{S\cup i}(\boldsymbol{a}_i) &= \boldsymbol{A}_{S\cup i}^\top \boldsymbol{a}_i \\
(\boldsymbol{A}_{S\cup i}^\top \boldsymbol{A}_{S\cup i} + \sigma^2 I)\boldsymbol{x}^{S\cup i}(\boldsymbol{a}_i) &= (\boldsymbol{A}_{S\cup i}^\top \boldsymbol{A}_{S\cup i} + \sigma^2 I)\mathbb{1}_i - \sigma^2\mathbb{1}_i \\
\boldsymbol{x}^{S\cup i}(\boldsymbol{a}_i) &= \mathbb{1}_i - \sigma^2(\boldsymbol{A}_{S\cup i}^\top \boldsymbol{A}_{S\cup i} + \sigma^2 I)^{-1}\mathbb{1}_i
\end{aligned}$$

Since $(\sigma^2 + \lambda_{\min}(\boldsymbol{A}))^{-1} I \succcurlyeq (\boldsymbol{A}_{S\cup i}^\top \boldsymbol{A}_{S\cup i} + \sigma^2 I)^{-1} \succcurlyeq (\sigma^2 + \lambda_{\max}(\boldsymbol{A}))^{-1} I$, we have

$$1 - \frac{\sigma^2}{\sigma^2 + \lambda_{\min}(\boldsymbol{A})} \leq [\boldsymbol{x}^{S\cup i}(\boldsymbol{a}_i)]_i \leq 1 - \frac{\sigma^2}{\sigma^2 + \lambda_{\max}(\boldsymbol{A})}.$$

We will construct two unit vectors $\boldsymbol{y}, \boldsymbol{z}$ such that $\frac{1}{2}(\frac{\lambda_{\min}(\boldsymbol{A})}{\sigma^2+\lambda_{\min}(\boldsymbol{A})})^2\|\boldsymbol{A}\boldsymbol{y}\|_2^2 \leq G(i|S) \leq \frac{1}{2}\|\boldsymbol{A}\boldsymbol{z}\|_2^2$.

Let $w_j = \frac{\boldsymbol{a}_j^\top R^S(\boldsymbol{a}_i)}{2\sqrt{\phi(S,i)}}, \forall j \in V$ and $\boldsymbol{z} = \boldsymbol{w}/\|\boldsymbol{w}\|_2$. Hence $\|\boldsymbol{z}\|_2 = 1$ and

$$\begin{aligned}\tfrac{1}{2}R^S(\boldsymbol{a}_i)^\top \boldsymbol{A}\boldsymbol{z} &= \tfrac{1}{2}\sum_{j\in V} R^S(\boldsymbol{a}_i)^\top \boldsymbol{a}_j \frac{w_j}{\|\boldsymbol{w}\|_2} \\ &= \sqrt{\phi(S,i)}\sum_{j \in V}\frac{w_j^2}{\|\boldsymbol{w}\|_2} \\ &= \sqrt{\phi(S,i)}\|\boldsymbol{w}\|_2.\end{aligned}$$

Note that $\|\boldsymbol{w}\|_2^2 = G(i|S)$ and $\phi(S,i) \geq \frac{1}{2}\|R^S(\boldsymbol{a}_i)\|^2$. Then by Cauchy-Schwartz inequality, we have:

$$\begin{aligned}G(i|S) &\leq \frac{\|R^S(\boldsymbol{a}_i)\|^2\|\boldsymbol{A}\boldsymbol{z}\|^2}{4\phi(S,i)} \\ &\leq \tfrac{1}{2}\|\boldsymbol{A}\boldsymbol{z}\|^2.\end{aligned}$$

Let $v_S = \boldsymbol{x}^S(\boldsymbol{a}_i), v_i = -1$ and zero elsewhere, and $y = v/\|v\|_2$. Hence $\|\boldsymbol{y}\|_2 = 1, \|v\|_2 \geq 1$ and

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{y}\|_2 &= \frac{\|R^S(\boldsymbol{a}_i)\|_2}{\|v\|_2} \\ &\leq \|R^S(\boldsymbol{a}_i)\|_2.\end{aligned}$$

Note that $G(i|S) \geq G_i(i|S) = (1 - \frac{\sigma^2}{\sigma^2+\lambda_{\min}(\boldsymbol{A})})^2\ \phi(S,i) \geq \frac{1}{2}(\frac{\lambda_{\min}(\boldsymbol{A})}{\sigma^2+\lambda_{\min}(\boldsymbol{A})})^2\|R^S(\boldsymbol{a}_i)\|_2^2 \geq \frac{1}{2}(\frac{\lambda_{\min}(\boldsymbol{A})}{\sigma^2+\lambda_{\min}(\boldsymbol{A})})^2\|\boldsymbol{A}\boldsymbol{y}\|_2^2$.
The proposition follows then from

$$\tfrac{1}{2}(\frac{\lambda_{\min}(\boldsymbol{A})}{\sigma^2+\lambda_{\min}(\boldsymbol{A})})^2\lambda^2_{\min}(\boldsymbol{A}) = \tfrac{1}{2}(\frac{\lambda_{\min}(\boldsymbol{A})}{\sigma^2+\lambda_{\min}(\boldsymbol{A})})^2\max_{\|\boldsymbol{y}\|_2=1}\|\boldsymbol{A}\boldsymbol{y}\|_2^2 \leq G(i|S) \leq \max_{\|\boldsymbol{z}\|_2=1}\tfrac{1}{2}\|\boldsymbol{A}\boldsymbol{z}\|_2^2 = \tfrac{1}{2}\lambda^2_{\max}(\boldsymbol{A}).$$

■

For the special case of $\sigma = 0$, we recover the result of [40].

Corollary 17 *Given a positive-definite kernel matrix $\boldsymbol{K}$, we define for any $i \in V$, $\ell_i(\boldsymbol{z}) = \|\boldsymbol{y} - \boldsymbol{K}^{1/2}\boldsymbol{z}\|_2^2 + \sigma^2\|\boldsymbol{z}\|_2^2$ with $\boldsymbol{y} = \boldsymbol{K}^{1/2}\mathbb{1}_i$, then we can write the variance reduction function $G(S) = \sum_{i\in V}\sigma^2(\boldsymbol{x}_i) - \sigma_S^2(\boldsymbol{x}_i) = \sum_{i\in V}\ell(\boldsymbol{0}) - \min_{\text{supp}(\boldsymbol{z})\subseteq S}\ell(\boldsymbol{z})$. Then G is a non-decreasing (β,β)-weakly DR-modular function, with $\beta = \frac{\lambda^2_{\min}(\boldsymbol{K})}{\lambda_{\max}(\boldsymbol{K})(\sigma^2+\lambda_{\min}(\boldsymbol{K}))}$, where $\lambda_{\max}(\boldsymbol{K})$ and $\lambda_{\min}(\boldsymbol{K})$ are the largest and smallest eigenvalues of $\boldsymbol{K}$.*

Proof For a fixed $i \in V, S \subseteq V \setminus i$, let $\boldsymbol{z}^S := \arg\min_{\text{supp}(x)\subseteq S}\ell_i(\boldsymbol{z})$. Then by optimality conditions $\boldsymbol{z}^S = (\boldsymbol{K}_S + \sigma^2\boldsymbol{I}_S)^{-1}\boldsymbol{k}_S(\boldsymbol{x}_i)$. Hence $\ell(\boldsymbol{z}^S) = \|\boldsymbol{y}\|_2^2 - 2\boldsymbol{y}^T\boldsymbol{K}_S^{1/2}\boldsymbol{z}^S + (\boldsymbol{z}^S)^\top(\boldsymbol{K}_S + \sigma^2\boldsymbol{I}_S)\boldsymbol{z}^S = \|\boldsymbol{y}\|_2^2 - \boldsymbol{k}_S(\boldsymbol{x}_i)(\boldsymbol{K}_S + \sigma^2\boldsymbol{I}_S)^{-1}\boldsymbol{k}_S(\boldsymbol{x}_i)$. It follows then that $\sigma^2(\boldsymbol{x}_i) - \sigma_S^2(\boldsymbol{x}_i) = \ell(\boldsymbol{0}) - \min_{\text{supp}(\boldsymbol{z})\subseteq S}\ell(\boldsymbol{z})$. The corollary then follows from Proposition 16. ■