Greedy Algorithms for Optimization on General Atomic Sets

Thomas Zhang[1] and Sahand Negahban[2]

[1]Yale College
[2]Yale University, Department of Statistics & Data Science

Contents

Abstract

We consider the problem of finding a sparse solution to maximizing a restricted strongly concave, restricted smooth function over an atomic set. Specifically, we extend the work of Elenberg et al. (2018) and show that restricted strongly concave, restricted smooth objectives are *approximately* submodular–a.k.a. weakly submodular. Our analysis extends to general atomic sets, which include most of the standard examples such as the canonical basis vectors and rank-one matrices, as well as some non-trivial examples where results of this kind are poorly studied. We leverage weak submodularity to obtain strong multiplicative approximation guarantees for two flavors of greedy algorithms, which by design construct sparse atomic solutions. We conclude with a discussion of the applications of these algorithms.

1 Background and Definitions

Sparsity in its many forms is central to a variety of problems across data analysis, compressed sensing, and high-dimensional statistics. The notion of sparsity differs from problem to problem: in support recovery, one seeks sparsity in the support of a vector; in PCA, one seeks sparsity in the spectrum of a matrix. Atomic sets elegantly capture these different notions of "sparsity".

Definition 1.1 (Atomic Sets [12]) *An atomic set is a (possibly uncountable) set of vectors* $\mathcal{A} = \{v_i\} \subseteq V$ *that is symmetric: if* $v \in \mathcal{A}$ *then* $-v \in \mathcal{A}$*. We note that the convex hull* $\boldsymbol{conv}(\mathcal{A})$ *contains*

0 *and is a polytope when* $\mathcal{A}$ *is finite.* $\boldsymbol{conv}(\mathcal{A})$ *induces a norm from its gauge function that we call the "atomic norm" induced by* $\mathcal{A}$*:*

$$\|x\|_{\mathcal{A}} := \inf\{t > 0 : x \in t \cdot \boldsymbol{conv}(\mathcal{A})\}.$$

Familiar examples of atomic sets include the aforementioned coordinate basis vectors $\{e_i\} \subset \mathbb{R}^n$ and rank-one matrices $\{uv^\top\} \subset \mathbb{R}^{n\times d}$. These examples provide a nice intuition to why certain atomic sets induce sparsity: in the coordinate basis vector case, the atomic unit ball is precisely the ℓ^1 unit ball, which is a polytope, yielding vertex solutions–corresponding to individual atoms–when maximizing/minimizing convex/concave functions. This motivates the study of algorithms for atomic norm regularization [12]. However, in this paper we are concerned with greedy algorithms that *construct* sparse atomic solutions. The problem we consider in this paper is the following "sparse atomic" maximization:

$$\begin{aligned} &\max\ f\left(\sum_{i=1}^{k} c_i v_i\right) \\ &\text{s.t. } c_i \in \mathbb{R},\ v_i \in \mathcal{A},\ i = 1,\dots,k. \end{aligned}$$

We now assume that f is restricted strongly concave and restricted strongly smooth, which are defined as follows.

Definition 1.2 (Restricted Strong Concavity, Restricted Smoothness [11, 10]) *A function* $f : V \to \mathbb{R}$ *is restricted strongly concave with parameter* m_Ω *and restricted smooth with parameter* M_Ω *if for all* $x, y \in \Omega \subset V$*,*

$$-\frac{m_\Omega}{2}\|y - x\|^2 \geq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \geq -\frac{M_\Omega}{2}\|y - x\|^2.$$

We remark that if $\Omega' \subseteq \Omega$, then by first principles

$$M_{\Omega'} \leq M_\Omega, \quad m_{\Omega'} \geq m_\Omega. \tag{1.1}$$

In earlier work by Elenberg et al. [3], it has been shown that restricted strongly concave functions are "weakly" submodular in the sense that greedy algorithms can also attain multiplicative bounds on the optimal solution. Before we define submodularity ratio, we must first construct a set function g from f to which a submodularity ratio can be ascribed.

Definition 1.3 *Given atomic set* $\mathcal{A}$ *and an indexing, let* $L = \{u_i\} \subset \mathcal{A}$*, we slightly abuse notation and define the set function*

$$g(L) = \max_{c_i \in \mathbb{R}}\ f\left(\sum_{u_i \in L} c_i u_i\right) - f(0),$$

such that $g(\emptyset) = 0$ *and the sparse atomic maximization problem can be reformulated as*

$$\max_{|L| \leq r} g(L).$$

Definition 1.4 (Submodularity Ratio [1]) *Let* $S, L \subset [p]$ *be two disjoint index sets, and* $g : [p] \to \mathbb{R}$*. The submodularity ratio of* L *with respect to* S *is defined as*

$$\gamma_{L,S} := \frac{\sum_{i\in S}\left(g\left(L \cup \{i\}\right) - g(L)\right)}{g(L \cup S) - g(L)}.$$

Essentially, we are measuring the diminishing marginal returns property. Note that $\gamma_{L,S} \geq 1$ if f is submodular. We can also define the submodularity ratio of a set $U \subseteq [p]$ with respect to a number k:

$$\gamma_{U,k} := \min_{\substack{L,S: L \cap S = \emptyset; \\ L \subseteq U \\ |S| = k}} \gamma_{L,S}.$$

Given a set of atoms $A = \{v_1, \ldots, v_k\}$ and a function $f(\cdot)$, we denote the optimal linear combination $B^{(A)}$:

$$f(B^{(A)}) := \max_{c_i \in \mathbb{R}} f\left(\sum_{i=1}^{k} c_i v_i\right) \quad \text{such that} \quad f(B^{(A)}) = g(A).$$

Khanna et al. [7, 8] have shown that greedy algorithms attain constant-factor approximations of the optimal solution within r iterations for restricted strongly convex, restricted smooth functions over sparse vectors and low-rank matrices. Our first aim is to show that these algorithms and approximation ratios can be extended to general atomic sets that are "rich enough" in a quantifiable sense.

Condition 1.5 (Orthogonality Condition) *An atomic set $\mathcal{A} \subset V$ satisfies the orthogonality condition if, picking an arbitrary atom $v \in \mathcal{A}$, one can complete an orthogonal basis for the ambient vector space V by successively picking elements in $\mathcal{A}$. In particular, this condition is satisfied by unions of orthornormal bases and low-rank tensors.*

However, there are many cases where atomic sets are not quite dense enough to contain precisely orthogonal bases. For example, in the atomic set formed by Gaussians with unit variance, one can never find two atoms that are precisely orthogonal due to the tails, but one can get arbitrarily close. Methods for dealing with this problem might involve truncating the tails. We want to justify these methods by proving that the submodularity ratio still approximately holds for atomic sets where we can guarantee picking "almost" orthogonal vectors. We quantify this near-orthogonality using the notion of "incoherence" [2].

Definition 1.6 (Coherence Parameter [2, 14]) *Given a set of vectors $\{v_i\}_{i \in [n]}$, the incoherence parameter $\varepsilon \geq 0$ measures the maximum absolute inner product between any two vectors v_i, v_j:*

$$\varepsilon := \max_{i \neq j} |\langle v_i, v_j \rangle| .$$

Note that when ε is close to 0, the vectors v_i are close to orthogonal in a geometric sense. We call such a set of vectors "incoherent". We thus have an analogous "incoherence" condition:

Condition 1.7 (Incoherence Condition) *An atomic set $\mathcal{A} \subset V$ satisfies the ε-incoherence condition if, picking an arbitrary atom $v_1 \in \mathcal{A}$, one can complete a basis $\{v_i\}$ for the ambient vector space V by successively picking elements in $\mathcal{A}$, such that the incoherence parameter of $\{v_i\}$ is no more than ε. If $\varepsilon = 0$, the Orthogonality Condition 1.5 is recovered.*

A subset of atoms $S = \{v_i\} \subset \mathcal{A}$ satisfies the ε-incoherence condition with respect to another set of atoms L if:

1. *S is ε-incoherent*
2. *every $v_i \in S$ is ε-incoherent with respect to L.*

It is also helpful to measure how "coherent" one set of vectors is with respect to another.

Definition 1.8 (Closeness) *We define two sets of vectors $\{v_i\}$ and $\{q_i\}$ to have closeness parameter of ε, $0 \leq \varepsilon \leq 1$ if*

$$1 - \varepsilon = \min_i \frac{|\langle v_i, q_i \rangle|}{\|v_i\| \, \|q_i\|}.$$

In other words, two sets of vectors are close if the inner products between vectors with the same indices are close to 1 (assuming that the indices are assigned appropriately).

Our Contributions: In this paper, we make three main contributions. First, we use a multitude of linear algebraic tools to prove that restricted strongly concave, restricted smooth functions over atomic sets satisfying the ε-Incoherence Condition 1.7 are weakly submodular for sufficiently small ε. Second, we extend a greedy scheme proposed by Khanna et al. [7] to the general atomic setting and show that it attains constant-factor approximations of the optimal r-sparse solution in no more than r iterations. This is a significant leg up on many linear convergence results of this type [15, 9, 4, 12], where the speed of the linear convergence is dependent on the ambient space dimension. Our analysis therefore extends to atomic sets living in arbitrary (separable) Hilbert spaces. In our third and final major contribution, we observe that the existing greedy schemes may not actually be fast for general large-scale optimization due to the many re-optimization steps, which is unfortunate since large-scale problems are the main motivation for these greedy algorithms. We remedy this by introducing a novel, non-iterative algorithm Oblivious that returns an r-sparse solution that attains a constant-factor approximation ratio to the optimal.

Aside from the computational motivations, extending weak submodularity bounds and algorithm formulation to admit incoherence has many interesting theoretical implications. As aforementioned, admitting incoherence allows us to consider atomic sets that contain probability distributions. Secondly, incoherence implies that uncountable atomic sets, for example the rank-one matrices, can actually be discretized into a finite search space without losing too much in the way of algorithmic guarantees.

2 A Few Technical Lemmas

In this section, we introduce and prove a few important lemmas regarding incoherence and closeness using fundamental linear algebraic tools.

Lemma 2.1 *Let $U = \{u_i\}$ and $V = \{v_i\}$ be sets of unit-length vectors of the same cardinality.*

1. *If $\|u_i - v_i\| < \varepsilon$ for all i, then U and V are at least $\varepsilon^2/2$-close.*
2. *If U and V are ε-close, and U is further assumed to be a set of orthogonal vectors, then for sufficiently small ε, V is at most $5\sqrt{\varepsilon}$-incoherent.*

Proof of Lemma 2.1: 1. follows from the following calculation

$$\begin{aligned} \|u_i - v_i\|^2 &= \|u_i\|^2 - 2\langle u_i, v_i \rangle + \|v_i\|^2 \\ &= 2 - 2\langle u_i, v_i \rangle \\ &< \varepsilon^2 \\ \implies \langle u_i, v_i \rangle &> 1 - \frac{\varepsilon^2}{2}. \end{aligned}$$

Reversing 1, we can prove 2: U and V are ε-close implies $\|u_i - v_i\| < \sqrt{2\varepsilon}$ for each i. Thus given arbitrary v_i and v_j, $i \neq j$, we can use triangle inequality to get the following

$$\begin{aligned}
\|v_i - v_j\| &= \|v_i - u_i + u_i - u_j + u_j - v_j\| \\
&\leq \|v_i - u_i\| + \|u_i - u_j\| + \|u_j - v_j\| \\
&< 2\sqrt{2\varepsilon} + \sqrt{2} \\
\implies 2\langle v_i, v_j\rangle &> -(8\varepsilon + 8\sqrt{\varepsilon}) \\
\langle v_i, v_j\rangle &> -5\sqrt{\varepsilon} \quad \text{for } 0 < \varepsilon < \frac{1}{16}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\|v_i + v_j\| &= \|v_i - u_i + u_i + u_j - u_j + v_j\| \\
&\leq \|v_i - u_i\| + \|u_i + u_j\| + \|u_j - v_j\| \\
&< 2\sqrt{2\varepsilon} + \sqrt{2} \\
\implies 2\langle v_i, v_j\rangle &< 8\varepsilon + 8\sqrt{\varepsilon} \\
\langle v_i, v_j\rangle &< 5\sqrt{\varepsilon} \quad \text{for } 0 < \varepsilon < \frac{1}{16} \\
\implies |\langle v_i, v_j\rangle| &< 5\sqrt{\varepsilon} \quad \text{for all } i.
\end{aligned}$$

■

This lemma tells us that to control the closeness of two sets of vectors, controlling the inner product $\langle u_i, v_i\rangle$ is equivalent to controlling the Euclidean norm of the error vectors $u_i - v_i$. Furthermore, as one would expect, sufficient closeness to an orthogonal set implies that the original set must have been incoherent to begin with. We lastly remark that for the orders of ε we require, ε always satisfies $0 < \varepsilon < \frac{1}{16}$.

Lemma 2.2 (Straightening Incoherent Vectors via SVD) *Given linearly independent vectors $S = \{v_1, \ldots, v_s\}$ that are ε-incoherent, there exists a set of orthonormal vectors $Q = \{q_1, \ldots, q_s\}$ that is δ-close to S, where $\delta < \frac{(s-1)^2}{2}\varepsilon^2$. Furthermore, $\boldsymbol{span}(S) = \boldsymbol{span}(Q)$.*

Proof of Lemma 2.2: for convenience, we will treat v_i as finite-dimensional column vectors and use matrix notation. We will formalize this to abstract real Hilbert spaces at the end. Consider the matrix $A = [v_1 \cdots v_s]$. We then consider its singular value decomposition

$$A = U\Sigma V^\top.$$

We claim the columns of matrix $A' = UV^\top = [q_1 \cdots q_s]$ satisfy the properties we want. Though it is not necessary to the proof, we note that A' is closest matrix with orthonormal columns to A in the Frobenius norm [5]. Observe that by construction the column spaces of A and A' both lie in the span of the first s columns of U, and since A and A' are of the same rank, this implies that $\mathbf{span}(S) = \mathbf{span}(Q)$.

To bound $\|v_i - q_i\|$, we use the variational characterization of eigenvalues. We observe that since A' is an orthogonal matrix, its non-zero singular values are $\sigma_k(A') = 1$. Furthermore, since v_i are all unit vectors, we know $\sigma_{\max}(A) = \lambda_{\max}(A^\top A) \geq 1$, which we verify by observing $\lambda_{\max}(A^\top A) \geq e_1^\top A^\top A e_1 = \|v_1\|^2 = 1$. Thus, we have the following series of inequalities

$$\sigma_{\min}(A) - 1 \leq \sigma_{\min}(A - A') \leq \sqrt{x^\top (A - A')^\top (A - A')x} \leq \sigma_{\max}(A - A') = \sigma_{\max}(A) - 1,$$

for all x with $\|x\| = 1$. In particular, setting $x = e_i$, we get

$$\sigma_{\min}(A) - 1 \leq \|v_i - q_i\| \leq \sigma_{\max}(A) - 1. \tag{2.1}$$

Thus, it suffices to bound the singular values of A. Note that since the columns of A are unit vectors, the matrix $A^\top A$ is of the form

$$A^\top A = \begin{bmatrix} 1 & v_1^\top v_2 & \cdots & v_1^\top v_s \\ v_2^\top v_1 & 1 & \cdots & v_2^\top v_s \\ \vdots & \vdots & \ddots & \vdots \\ v_s^\top v_1 & v_s^\top v_2 & \cdots & 1 \end{bmatrix}.$$

By the Gershgorin disc theorem [5], every eigenvalue of $A^\top A$ lies in the union of the Gershgorin discs.

$$\begin{aligned} \lambda_k(A^\top A) &\in \bigcup_{i=1}^{s} \left\{ z \in \mathbb{C} : \left| z - (A^\top A)_{ii} \right| \leq \sum_{j \neq i} \left| (A^\top A)_{ij} \right| \right\} \\ \implies \lambda_k(A^\top A) &\in \bigcup_{i=1}^{s} \left\{ z \in \mathbb{C} : |z - 1| < \sum_{j \neq i} \varepsilon = (s-1)\varepsilon \right\}, \end{aligned}$$

where the last line comes for the ε-incoherence of v_i. Since $A^\top A$ is symmetric, all its eigenvalues are real, and thus we have the following bound

$$1 - (s-1)\varepsilon < \lambda_k(A^\top A) < 1 + (s-1)\varepsilon.$$

For sufficiently small ε, this demonstrates that $A^\top A$ is positive definite. By the definition of singular values, we now have

$$\begin{aligned} \sigma_k(A)^2 &= \lambda_k(A^\top A) > 0 \\ \sigma_k(A) - 1 &= \frac{\lambda_k(A^\top A) - 1}{\sigma_k(A) + 1} \\ &\leq \frac{(s-1)\varepsilon}{\sigma_k(A) + 1} \\ &< (s-1)\varepsilon. \end{aligned}$$

Specifically, this gets us $\sigma_{\max}(A) - 1 < (s-1)\varepsilon$, which plugging back into inequality (2.1) gets us the bound

$$\|v_i - q_i\| < (s-1)\varepsilon \quad \text{for } i = 1, \ldots, s.$$

Lemma 2.1 implies that the vector sets S and Q are at least $\frac{(s-1)^2}{2}\varepsilon^2$-close.

Extending this proof to Hilbert spaces is simple. The key is that since we started with finitely many vectors S, A has a finite-dimensional range space, and thus is a compact operator. The spectral theorem for self-adjoint compact operators may be applied on $A^\top A$, which also has finite-dimensional range space. Thus, our arguments using the Gershgorin disc theorem carry through. Since we only consider finite vector sets, and we never use any argument that assumes a finite-dimensional ambient space, this extension to vectors in Hilbert spaces will work for all the lemmas and theorems to come. ■

Lemma 2.3 (Straightening Incoherent Vectors via Gram-Schmidt) *Given a set of vectors $S = \{v_1, \ldots, v_k\}$ that has a sufficiently small incoherence parameter ε, applying the Gram-Schmidt process to S yields a set of orthogonal vectors $Q = \{q_i\}$ that has closeness parameter $\delta < k^2\varepsilon$.*

Proof of Lemma 2.3: define the matrix $A = [v_1 \cdots v_k]$ as usual. Recall the unnormalized Gram-Schmidt process on S:

$$\begin{aligned} u_1 &= v_1 \\ &\vdots \\ u_k &= v_k - \sum_{i=1}^{k-1} \frac{\langle u_i, v_k \rangle}{\langle u_i, u_i \rangle} u_i. \end{aligned}$$

Define q_i to be normalized: $u_i / \|u_i\|$. Let us also define

$$D_i = \det\left((A^\top A)_{[i]}\right),$$

where $(A^\top A)_{[i]}$ is the i-th leading principal submatrix of $A^\top A$. We have the following well-known expression for $\|u_i\|$:

$$\begin{aligned} \|u_i\| &= \sqrt{\frac{D_i}{D_{i-1}}} \\ &\geq \sqrt{D_i} \\ &\geq \sqrt{D_k}, \end{aligned}$$

where first inequality follows from an application of Hadamard's inequality, which states: given positive-definite $M \in M_k$,

$$\det(M) \leq \prod_{i=1}^{k} M_{ii}.$$

Setting $M = A^\top A$, we have $M_{ii} = 1$, and thus

$$D_{i-1} \leq 1.$$

To show $\sqrt{D_i} \geq \sqrt{D_k}$, we use a lemma concerning Schur complements, which states: given a non-singular symmetric matrix A partitioned into

$$A = \begin{bmatrix} A' & B \\ B^\top & C \end{bmatrix},$$

then

$$\det(A) = \det(A') \det(C - B^\top A'^{-1} B).$$

We apply this on $(A^\top A)_{[k]} = A^\top A = \begin{bmatrix} (A^\top A)_{[k-1]} & b_k \\ b_k^\top & 1 \end{bmatrix}$, $b_k = [v_k^\top v_1 \ \cdots \ v_k^\top v_{k-1}]^\top$:

$$\begin{aligned} D_k &= \det(A^\top A) \\ &= \det((A^\top A)_{[k-1]}) \left(1 - b_k^\top (A^\top A)_{[k-1]}^{-1} b_k\right) \\ &\leq D_{k-1}. \end{aligned}$$

Since $A^\top A$ is positive definite, its leading principal submatrices are positive definite and thus $0 < b_k^\top (A^\top A)^{-1}_{[k-1]} b_k < 1$. Thus, we have shown $\sqrt{D_i} \geq \sqrt{D_k}$. To estimate the value of D_k, we recall that the determinant of a matrix is the product of its eigenvalues. From the proof of Lemma 2.2 we have the following lower bound on the eigenvalues of $A^\top A$:

$$1 - (k-1)\varepsilon < \lambda_{\min}(A^\top A).$$

Thus,

$$\begin{aligned} D_k &= \prod_{i=1}^{k} \lambda_i(A^\top A) \\ &> (1-(k-1)\varepsilon)^k \\ &> \left(1 - \frac{\omega}{k}\right)^k \quad \text{setting } \omega = k^2\varepsilon \\ &> 1 - \omega \\ \implies \|u_i\|^2 &> 1 - \omega. \end{aligned}$$

On the other hand, observe that from the Gram-Schmidt process,

$$\begin{aligned} \|u_i\|^2 &= \langle v_i, u_i \rangle, \\ \|u_i - v_i\|^2 &= \sum_{j=1}^{i-1} \langle d_j, v_i \rangle^2 \\ &= \|u_i\|^2 - 2\langle u_i, v_i \rangle + 1 \\ &= 1 - \|u_i\|^2, \\ \|q_i - v_i\|^2 &= 2 - 2\langle q_i, v_i \rangle \\ &= 2 - 2\frac{\langle u_i, v_i \rangle}{\|u_i\|} \\ &= 2(1 - \|u_i\|) \\ &< 2(1 - \sqrt{1-\omega}) \\ &= 2\left(\frac{\omega}{1+\sqrt{1-\omega}}\right) \\ &\leq 2\omega. \end{aligned}$$

By Lemma 2.1, Q and S are at least ω-close. Therefore, there exists a $\delta < \omega = k^2\varepsilon$ such that Q and S are δ-close. ■

Lemma 2.4 (Error from Straightening) *Given a set of unit-length vectors $S = \{v_i\}$ and an ε-close orthogonal set of vectors $Q = \{q_i\}$ where $|S| = |Q| = k$ and $\boldsymbol{span}(S) = \boldsymbol{span}(Q) = W \subseteq V$, then given some vector $u \in V$, we have the following lower and upper bounds*

$$(1-\delta)\sum_{i=1}^{k} \langle u, q_i \rangle^2 \leq \sum_{i=1}^{k} \langle u, v_i \rangle^2 \leq (1+\delta)\sum_{i=1}^{k} \langle u, q_i \rangle^2$$

$$(1-\delta)\|P_W(u)\|^2 \leq \sum_{i=1}^{k} \langle u, v_i \rangle^2 \leq (1+\delta)\|P_W(u)\|^2$$

for some $0 \leq \delta < 5(k-1)\sqrt{\varepsilon} < 5k\sqrt{\varepsilon}$*, where* $P_W(u)$ *denotes the orthogonal projection of* u *onto* W*.*

This lemma tells us that if the spans of a set of vectors and its rounded orthogonal set are identical, such as the case for any set of vectors and its Gram-Schmidt orthogonalized basis, then we can replace the expression in the middle of the inequality with a much simpler projection term while only accruing a small multiplicative error.

Proof of Lemma 2.4: for convenience, we treat vectors v_i as coordinate vectors and use matrix notation.

Let $A = [v_1 \cdots v_k]$ and $A' = [q_1 \cdots q_k]$. Observe that

$$u^\top A A^\top u = \sum_{i=1}^{k} \langle u, v_i \rangle^2 , \quad u^\top A' A'^\top u = \sum_{i=1}^{k} \langle u, q_i \rangle^2 .$$

We observe that if $u \perp W$, then the inequality is automatically fulfilled. Let us assume that $u \in W$. Therefore, $u^\top A' A'^\top u = \|u\|^2$. It thus suffices to bound the non-zero eigenvalues of $AA^\top$:

$$(1-\delta)\|u\|^2 \leq u^\top A A^\top u \leq (1+\delta)\|u\|^2 .$$

Recall that the non-zero eigenvalues of $AA^\top$ are identical to those of $A^\top A$. If we can establish that S is ω-incoherent, then following the proof of Lemma 2.2 we will find

$$(1-(k-1)\omega)\|u\|^2 \leq u^\top A A^\top u \leq (1+(k-1)\omega)\|u\|^2 .$$

Thus it remains to find the incoherence parameter ω for S given that S and Q are ε-close. By Lemma 2.1, S is $5\sqrt{\varepsilon}$-incoherent. Thus we have the inequality

$$(1-5(k-1)\sqrt{\varepsilon})\|u\|^2 \leq u^\top A A^\top u \leq (1+5(k-1)\sqrt{\varepsilon})\|u\|^2 .$$

Therefore, for some $0 \leq \delta < 5(k-1)\sqrt{\varepsilon}$, we have

$$(1-\delta)\sum_{i=1}^{k} \langle u, q_i \rangle^2 \leq \sum_{i=1}^{k} \langle u, v_i \rangle^2 \leq (1+\delta)\sum_{i=1}^{k} \langle u, q_i \rangle^2 .$$

The formalization of this proof to Hilbert spaces follows from the same argument as in Lemma 2.2. ■

3 Main Theorem: Submodularity Ratio Bound

Theorem 3.1 (Submodularity Ratio Bound) *Let* $\mathcal{A}$ *be an atomic set that lives in real inner product space* V*. Let* $L = \{u_i\}$ *be a set of* k *atoms and* $S = \{v_j\}$ *be a set of* s *atoms such that the atoms in* S *are pairwise* ε*-incoherent, and are* ε*-incoherent with the atoms in* L*. If* f *is* m_n*-strongly concave over all linear combinations of* n *atoms, and* $\tilde{M}_1$*-smooth over the set*

$$\tilde{\Omega} := \left\{ \left(\sum_{i=1}^{p} c_i u_i, \sum_{j=1}^{q} d_i v_i \right) : |\{u_i\} \triangle \{v_i\}| \leq 1 \right\},$$

(where $\triangle$ denotes symmetric difference) in other words the set of all pairs of vectors that admit an atomic representation that differ at most by 1 atom. We then have the following submodularity ratio for f:

$$\gamma_{L,S} := \frac{\sum_{j\in S}\left(g(L\cup\{j\})-g(L)\right)}{g\left(L\cup S\right)-g(L)} \geq \frac{1-5s\sqrt{\varepsilon}}{1+3\sqrt{s}(k+s)\sqrt{\varepsilon}}\frac{m_{k+s}}{\tilde{M}_1}.$$

This theorem is analogous to the one provided by Khanna et al. [7], except we replace rank-one matrices with atoms and require the ε-Incoherence Condition 1.7. The original proof structure is retained but many more technical arguments must be made that make extensive use of the lemmas we have seen. For convenience, we further assume that the atoms in L and S are normalized: $\|u_i\| = \|v_j\| = 1$.

Proof: essentially this boils down to lower bounding the numerator and upper bounding the denominator. We observe that

$$g(L\cup\{j\}) - g(L) = f\left(B^{(L\cup\{j\})}\right) - f\left(B^{(L)}\right),$$

and thus it suffices to deal with $f(\cdot)$. We denote $B_j,\ j\in S$ as the orthogonal projection of B onto the atom v_j:

$$B_j = \langle B, v_j\rangle\, v_j.$$

By the optimality of $B^{(L\cup\{j\})}$ and restricted strong concavity, we now have the following series of inequalities

$$\begin{aligned}
f\left(B^{(L\cup\{j\})}\right) - f\left(B^{(L)}\right) &\geq f\left(B^{(L)} + \alpha_j B_j^{(L\cup S)}\right) - f\left(B^{(L)}\right) \\
&\geq \left\langle \nabla f\left(B^{(L)}\right), \left(B^{(L)} + \alpha_j B_j^{(L\cup S)}\right) - B^{(L)}\right\rangle \\
&\quad - \frac{\tilde{M}_1}{2}\left\|\left(B^{(L)} + \alpha_j B_j^{(L\cup S)}\right) - B^{(L)}\right\|^2 \\
&= \left\langle \nabla f\left(B^{(L)}\right), \alpha_j B_j^{(L\cup S)}\right\rangle - \alpha_j^2\frac{\tilde{M}_1}{2}\left\|B_j^{(L\cup S)}\right\|^2,
\end{aligned}$$

where α_j is an arbitrary constant. Let us set $\alpha_j = \frac{\left\langle \nabla f\left(B^{(L)}\right), B_j^{(L\cup S)}\right\rangle}{\tilde{M}_1\left\|B_j^{(L\cup S)}\right\|^2}$:

$$\left\langle \nabla f\left(B^{(L)}\right), \alpha_j B_j^{(L\cup S)}\right\rangle - \alpha_j^2\frac{\tilde{M}_1}{2}\left\|B_j^{(L\cup S)}\right\|^2 = \frac{1}{2}\frac{\left\langle \nabla f\left(B^{(L)}\right), B_j^{(L\cup S)}\right\rangle^2}{\tilde{M}_1\left\|B_j^{(L\cup S)}\right\|^2}.$$

Therefore, we have established that

$$\begin{aligned}
f\left(B^{(L\cup\{j\})}\right) - f\left(B^{(L)}\right) &\geq \frac{1}{2}\frac{\left\langle \nabla f\left(B^{(L)}\right), B_j^{(L\cup S)}\right\rangle^2}{\tilde{M}_1\left\|B_j^{(L\cup S)}\right\|^2} \\
\implies \sum_{j\in S} f\left(B^{(L\cup\{j\})}\right) - f\left(B^{(L)}\right) &\geq \sum_{j\in S}\frac{1}{2}\frac{\left\langle \nabla f\left(B^{(L)}\right), B_j^{(L\cup S)}\right\rangle^2}{\tilde{M}_1\left\|B_j^{(L\cup S)}\right\|^2} \\
&= \frac{1}{2\tilde{M}_1}\sum_{j\in S}\frac{\left\langle \nabla f\left(B^{(L)}\right), B_j^{(L\cup S)}\right\rangle^2}{\left\|B_j^{(L\cup S)}\right\|^2}.
\end{aligned}$$

Using Lemma 2.2 to find an orthonormal basis $Q = \{q_i\}$ for $\mathbf{span}\,(S)$ and Lemma 2.4, we get

$$\frac{1}{2\tilde{M}_1}\sum_{j\in S}\frac{\left\langle \nabla f\left(B^{(L)}\right), B_j^{(L\cup S)}\right\rangle^2}{\left\|B_j^{(L\cup S)}\right\|^2} \geq \frac{1-5s\sqrt{\varepsilon}}{2\tilde{M}_1}\left\|P_S\left(\nabla f\left(B^{(L)}\right)\right)\right\|^2$$

where $P_S(X)$ denotes the projection of X onto the $\mathbf{span}\,(S)$. Setting $\delta = 1-5s\sqrt{\varepsilon}$, we have the following lower bound on the numerator

$$\sum_{j\in S} f\left(B^{(L\cup\{j\})}\right) - f\left(B^{(L)}\right) \geq \frac{1-\delta}{2\tilde{M}_1}\left\|P_S\left(\nabla f\left(B^{(L)}\right)\right)\right\|^2.$$

Before moving on to upper bounding the denominator, we address the case when $\nabla f\left(B^{(L)}\right) \perp \mathbf{span}\,(S)$, in which case the norm of the projection onto $\mathbf{span}\,(S)$ is 0 and we might get an unproductive submodularity ratio of 0. We argue that this cannot happen using the optimality of $B^{(L)}$. By definition of $B^{(L)}$ it is the optimal linear combination of the atoms in L, and thus by the first-order optimality condition over linear subspaces we have

$$\left\langle \nabla f\left(B^{(L)}\right), X\right\rangle = 0, \quad X \in \mathbf{span}\,(L).$$

In other words, $\nabla f\left(B^{(L)}\right) \perp \mathbf{span}\,(L)$. If further $\nabla f\left(B^{(L)}\right) \perp \mathbf{span}\,(S)$, then we get a trivial submodularity ratio of 1, which is greater than 0. This addresses this case of $\nabla f\left(B^{(L)}\right) \perp \mathbf{span}\,(S)$. Thus it suffices to consider the case when $\left\|P_S\left(\nabla f\left(B^{(L)}\right)\right)\right\| > 0$. We now move onto bounding the denominator.

By the restricted strong concavity by f:

$$f\left(B^{(L)}\right) - f\left(B^{(L\cup S)}\right) + \left\langle \nabla f\left(B^{(L)}\right), B^{(L\cup S)} - B^{(L)}\right\rangle - \frac{m_{k+s}}{2}\left\|B^{(L\cup S)} - B^{(L)}\right\|^2 \geq 0.$$

Therefore, we can use this to bound the denominator

$$\begin{aligned} f\left(B^{(L\cup S)}\right) - f\left(B^{(L)}\right) &\leq \left\langle \nabla f\left(B^{(L)}\right), B^{(L\cup S)} - B^{(L)}\right\rangle - \frac{m_{k+s}}{2}\left\|B^{(L\cup S)} - B^{(L)}\right\|^2 \\ &\leq \max_{X\in\mathbf{span}(L\cup S)}\left\langle \nabla f\left(B^{(L)}\right), X - B^{(L)}\right\rangle - \frac{m_{k+s}}{2}\left\|X - B^{(L)}\right\|^2. \end{aligned}$$

Now we use the fact that $B^{(L)}$ is the optimal solution of f over $\mathbf{span}\,(L)$. From our earlier discussion about the orthogonality of $\nabla f\left(B^{(L)}\right)$ to $\mathbf{span}\,(L)$ we have

$$\left\langle \nabla f\left(B^{(L)}\right), P_L(X) - B^{(L)}\right\rangle = 0 \quad \text{for all } X,$$

since $P_L(X) - B^{(L)} \in \mathrm{span}\,(L)$. We want to know what the optimal X looks like. Let us define $\mathbb{S} = \mathbf{span}\,(L)^{\perp}$. We observe that X must be of the form

$$X = B^{(L)} + X_{\mathbb{S}},$$

where $X_{\mathbb{S}}$ is in $\mathbb{S}$. This must be true because the L component of X doesn't change the inner product, but we want to minimize the second term $\frac{m_{k+s}}{2}\left\|X - B^{(L)}\right\|^2$. We want X_L to cancel out

$B^{(L)}$, since

$$\begin{aligned}\left\|X - B^{(L)}\right\|^2 &= \left\langle X - B^{(L)}, X - B^{(L)} \right\rangle \\ &= \left\langle X_{\mathbb{S}} + (X_L - B^{(L)}), X_{\mathbb{S}} + (X_L - B^{(L)}) \right\rangle \\ &= \|X_{\mathbb{S}}\|^2 + 2\left\langle X_{\mathbb{S}}, X_L - B^{(L)} \right\rangle + \left\|X_L - B^{(L)}\right\|^2 \\ &= \|X_{\mathbb{S}}\|^2 + \left\|X_L - B^{(L)}\right\|^2.\end{aligned}$$

This leaves us to determine $X_{\mathbb{S}}$. Notice that our problem has been reduced to

$$\begin{aligned}&\max_{X_{\mathbb{S}}} \left\langle \nabla f\left(B^{(L)}\right), X_{\mathbb{S}} \right\rangle - \frac{m_{k+s}}{2}\|X_{\mathbb{S}}\|^2 \\ =\ &\max_{X_{\mathbb{S}}} \left\langle P_{\mathbb{S}}\left(\nabla f\left(B^{(L)}\right)\right), X_{\mathbb{S}} \right\rangle - \frac{m_{k+s}}{2}\|X_{\mathbb{S}}\|^2.\end{aligned}$$

We observe that the optimal $X_{\mathbb{S}}$ will be collinear with $P_{\mathbb{S}}\left(\nabla f\left(B^{(L)}\right)\right)$, in which case this reduces to a calculus problem:

$$\max_c\ c\left\|P_{\mathbb{S}}\left(\nabla f\left(B^{(L)}\right)\right)\right\|^2 - c^2\frac{m_{k+s}}{2}\left\|P_{\mathbb{S}}\left(\nabla f\left(B^{(L)}\right)\right)\right\|^2.$$

It is simple to verify that the maximizer is

$$X^* = c^* P_{\mathbb{S}}\left(\nabla f\left(B^{(L)}\right)\right) = \frac{1}{m_{k+s}} P_{\mathbb{S}}\left(\nabla f\left(B^{(L)}\right)\right),$$

which yields

$$\begin{aligned}f\left(B^{(L\cup S)}\right) - f\left(B^{(L)}\right) &\le \left\langle \nabla f\left(B^{(L)}\right), \frac{1}{m_{k+s}} P_{\mathbb{S}}\left(\nabla f\left(B^{(L)}\right)\right) \right\rangle - \frac{m_{k+s}}{2}\left\|\frac{1}{m_{k+s}} P_{\mathbb{S}}\left(\nabla f\left(B^{(L)}\right)\right)\right\|^2 \\ &\le \frac{1}{2m_{k+s}}\left\|P_{\mathbb{S}}\left(\nabla f\left(B^{(L)}\right)\right)\right\|^2.\end{aligned}$$

It remains to relate $\left\|P_{\mathbb{S}}\left(\nabla f\left(B^{(L)}\right)\right)\right\|$ and $\left\|P_S\left(\nabla f\left(B^{(L)}\right)\right)\right\|$, so that we can cancel terms on the numerator and denominator.

Let $Q_L = \{q_i\}_{i=1}^k$ be any orthogonal basis for $\mathbf{span}(L)$, and $Q_S = \{p_j\}_{j=1}^s$ be the orthogonal basis for $\mathbf{span}(S)$ obtained from Lemma 2.2. Observe that $Q_L \cup S$ is an ε-incoherent basis for $\mathbf{span}(L \cup S)$. Applying Gram-Schmidt on $Q_L \cup S$ starting at q_1, we get a full orthogonal basis for $\mathbf{span}(L \cup S)$: $Q = \{q_i\}_{i=1}^{k+s}$. Note that $\{q_i\}_{i=k+1}^{k+s}$ forms an orthogonal basis for $\mathbb{S}$. Define the matrices

$$\begin{aligned}A_L &= [q_1 \cdots q_k] \\ A_{\mathbb{S}} &= [q_{k+1} \cdots q_{k+s}] \\ A_S &= [p_1 \cdots p_s].\end{aligned}$$

We then express $P_S\left(\nabla f\left(B^{(L)}\right)\right) = A_S x$ for some $x \in \mathbb{R}^s$ such that

$$\|x\| = \left\|P_S\left(\nabla f\left(B^{(L)}\right)\right)\right\|.$$

We therefore have

$$\begin{aligned}
\left\|P_{\mathbb{S}}\left(\nabla f\left(B^{(L)}\right)\right)\right\| &= \left\|A_{\mathbb{S}}^{\top}\nabla f\left(B^{(L)}\right)\right\| \\
&= \left\|A_{\mathbb{S}}^{\top}A_S x\right\| \\
&\leq \left\|A_{\mathbb{S}}^{\top}(A_S - A_{\mathbb{S}})x\right\| + \|x\| \\
&\leq \left(\left\|A_{\mathbb{S}}^{\top}(A_S - A_{\mathbb{S}})\right\|_2 + 1\right)\|x\| \\
&= \left(\|A_S - A_{\mathbb{S}}\|_2 + 1\right)\left\|P_S\left(\nabla f\left(B^{(L)}\right)\right)\right\|,
\end{aligned}$$

where $\|X\|_2$ indicates the operator norm of X. We can upper bound $\|A_S - A_{\mathbb{S}}\|_2$. Note that the columns of $A_S - A_{\mathbb{S}}$ are $p_i - q_{k+i}$:

$$\begin{aligned}
\|A_S - A_{\mathbb{S}}\|_2 &= \max_{\|x\|=1}\left\|(A_S - A_{\mathbb{S}})^{\top}x\right\| \\
&= \max_{\|x\|=1}\left\|\begin{bmatrix}(p_1 - q_{k+1})^{\top}x\\ \vdots \\ (p_s - q_{k+s})^{\top}x\end{bmatrix}\right\| \\
&\leq \max_{\|x\|=1}\left\|\begin{bmatrix}\|p_1 - q_{k+1}\|\,\|x\|\\ \vdots \\ \|p_s - q_{k+s}\|\,\|x\|\end{bmatrix}\right\| \\
&= \sqrt{\sum_{i=1}^{s}(\|p_i - q_{k+i}\|)^2} \\
&\leq \sqrt{\sum_{i=1}^{s}(\|p_i - v_i\| + \|v_i - q_{k+i}\|)^2}.
\end{aligned}$$

From Lemma 2.2,

$$\|p_i - v_i\| < (s-1)\varepsilon, \quad i = 1, \ldots, s,$$

From Lemma 2.3,

$$\|v_i - q_{k+i}\| < 2(k+s)\sqrt{\varepsilon}, \quad i = 1, \ldots, s.$$

Thus we have the following upper bound on the spectral norm

$$\begin{aligned}
\|A_S - A_{\mathbb{S}}\|_2 &< \sqrt{\sum_{i=1}^{s}\left((s-1)\varepsilon + 2(k+s)\sqrt{\varepsilon}\right)^2} \\
&< \sqrt{\sum_{i=1}^{s}\left(3(k+s)\sqrt{\varepsilon}\right)^2} \\
&= 3\sqrt{s}(k+s)\sqrt{\varepsilon}.
\end{aligned}$$

Setting $\tau = 3\sqrt{s}(k+s)\sqrt{\varepsilon}$, we have

$$\left\|P_{\mathbb{S}}\left(\nabla f\left(B^{(L)}\right)\right)\right\| < (1+\tau)\left\|P_S\left(\nabla f\left(B^{(L)}\right)\right)\right\|.$$

Returning to upper bounding the denominator of the submodularity ratio, we now have

$$\begin{aligned} f\left(B^{(L\cup S)}\right) - f\left(B^{(L)}\right) &\leq \frac{1}{2m_{k+s}} \left\| P_{\mathbb{S}}\left(\nabla f\left(B^{(L)}\right)\right)\right\|^2 \\ &< \frac{1+\tau}{2m_{k+s}} \left\| P_{S}\left(\nabla f\left(B^{(L)}\right)\right)\right\|^2. \end{aligned}$$

Putting the inequalities for the numerator and denominator together, we get

$$\begin{aligned} \gamma_{L,S} &:= \frac{\sum_{i\in S}\left(g(L\cup\{j\}) - g(L)\right)}{g\left(L\cup S\right) - g(L)} \\ &= \frac{\sum_{i\in S}\left(f(B^{(L\cup\{j\})}) - f(B^{(L)})\right)}{f\left(B^{(L\cup S)}\right) - f(B^{(L)})} \\ &\geq \frac{1-\delta}{1+\tau}\frac{m_{k+s}}{\tilde{M}_1} \\ &= \frac{1-5s\sqrt{\varepsilon}}{1+3\sqrt{s}(k+s)\sqrt{\varepsilon}}\frac{m_{k+s}}{\tilde{M}_1} \end{aligned}$$

■

We remark that to achieve a submodularity ratio close to $\frac{m_{k+s}}{\tilde{M}_1}$, ε should be on the order $0 \leq \varepsilon < \frac{1}{(k+s)^4}$. While this might look like a small quantity, we will later see that for our purposes $k, s \leq r$, where r is the sparsity constraint on the problem. Therefore, the incoherence parameter only scales with the sparsity of the problem, rather than the dimension of the ambient space. As previously discussed, when an atomic set fulfills the Orthogonality Condition 1.5, $\varepsilon = 0$ can be precisely attained, which results in a submodularity ratio that is precisely $\frac{m_{k+s}}{\tilde{M}_1}$, which matches the submodularity ratio Khanna et al. [7, 8] derived for rank-one matrices and coordinate basis vectors.

4 Greedy Algorithms and Approximation Guarantees

We now want to show that we can leverage weak submodularity to obtain approximation guarantees on simple greedy algorithms that depend only on the underlying atomic sparsity of the problem. We recall that the problem we are trying to solve is

$$\begin{aligned} &\max\ g(L) \\ &\text{s.t. } |L| \leq r, \end{aligned}$$

where L are sets of atoms $\{v_i\}$. We now assume access to the following oracles.

1. `OMPSel`(β, L) (Orthogonal Matching Pursuit Selector) is a linear oracle that given precision parameter β and atomic set L returns atom $\hat{v} \in \mathcal{A}$ such that

$$\left\langle \nabla f(B^{(L)}), \hat{v} \right\rangle \geq \beta \max_{v\perp L} \left\langle \nabla f(B^{(L)}), v \right\rangle, \tag{4.1}$$

which is well-studied in the context of dictionary learning, sparse recovery, and low-rank optimization [7, 9, 14, 4]. This particular oracle is nice because it can be easily computed in certain cases of interest. For example, in the low-rank matrix case, `OMPSel` can be computed by finding the left and right singular vectors corresponding to the largest singular value of $\nabla f\left(B^{(L)}\right)$: $\left\|\nabla f\left(B^{(L)}\right)\right\|_2$ [13].

2. More generally under the ε-Incoherence Condition 1.7, we replace `OMPSel` with a general Linear Maximization Oracle $\texttt{LMO}(\beta, u, L)$ that returns atom $\hat{v}$ such that

$$\langle u, \hat{v} \rangle \geq \beta \max_{v \perp_\varepsilon L} \langle u, v \rangle , \tag{4.2}$$

where $v \perp_\varepsilon L$ denotes atoms $v \in \mathcal{A}$ that are ε-incoherent with L.

3. We further assume that given a set of atoms $L = \{v_i\}$ and a function f, we can find

$$B^{(L)} = \arg\max_{v_i \in L} f\left(\sum_i c_i v_i\right).$$

Since f is restricted strongly concave in our discussion, this is a convex programming problem.

We propose two distinct greedy schemes for maximizing a restricted strongly concave function.

4.1 Iterative Atom Selection

The first greedy scheme is inspired by the simple algorithm for feature selection and low-rank optimization proposed in Khanna et al. [8, 7]. With the above assumptions on oracles, the approximation guarantees derived in their work can be translated for general sparse atomic optimization. We use the following algorithm known as GECO (Greedy Efficient Component Optimization) [13], where the parameters are $\text{GECO}(\mathcal{A}, k, \beta)$: $\mathcal{A}$ denotes relevant parameters for the atomic set, k is the sparsity parameter, and β is the precision parameter for the linear oracle in the algorithm.

Algorithm 1 $\text{GECO}(\mathcal{A}, k, \beta)$

$L_0 \leftarrow \emptyset$
for $t = 1, \ldots, k$ **do**
  Compute $B^{(L_{t-1})}$, $\nabla f\left(B^{(L_{t-1})}\right)$
  $v_t \leftarrow \texttt{OMPSel}(\beta, L_{t-1})$
  $\left(v_t \leftarrow \texttt{LMO}(\beta, \nabla f\left(B^{(L_{t-1})}\right), L_{t-1})\right)$
  $L_t \leftarrow L_{t-1} \cup \{v_t\}$
end for
return $L_k, B^{(L_k)}, f(L_k)$

Theorem 4.1 (Approximation Guarantee for GECO) *Let atomic set $\mathcal{A}$ satisfy incoherence conditions described in Theorem 3.1. Let L_k be the greedy solution obtained by GECO after k iterations, and let L^* be the optimal atomic set of size at most r satisfying the rank constraint of the problem. If f is m_{k+r} restricted strong concave and $\tilde{M}_1$ restricted smooth on the sets defined in Theorem 3.1. Then*

$$g(L_k) \geq \left(1 - \frac{1}{e^{kw}}\right) g(L^*),$$

where $w = \left(\beta^2 \eta \frac{m_{k+r}}{\tilde{M}_1}\right) \frac{1}{r}$, $\eta = \frac{1}{1+3\sqrt{r(k+r)}\sqrt{\varepsilon}}$.

Note that the above theorem implies that we can attain a constant factor approximation of the true optimal value after iterating the algorithm r times, where r is the underlying rank of the optimal solution. To the authors' knowledge, this matches existing results for sparse and low-rank matrix

recovery [3, 8, 7, 4], and is the first of its type in general sparse atomic recovery. Results for sparse atomic recovery have shown linear convergence of greedy algorithms, except the constant factor approximation depends on the dimension of the ambient space rather than the sparsity.

Toward proving the above theorem, we introduce some notation and prove a key lemma. Define $D(t) = g(L_t) - g(L_{t-1}),\ D(0) := 0$ as the improvement after the t-th iteration of the algorithm. Let us also define $R(t) = g(L^*) - g(L_t)$ as the *remaining* distance from the optimal value after the t-th iteration. The following key lemma relates the weak submodularity constant from the previous section to the incremental improvement $D(t)$.

Lemma 4.2 *At iteration t, we have the following lower bound on the incremental gain*

$$D(t) \geq \beta^2 \eta \frac{m_{t+r}}{\tilde{M}_1} \frac{1}{r} R(t) \geq \beta^2 \eta \frac{m_{k+r}}{\tilde{M}_1} \frac{1}{r} R(t),$$

where the second inequality comes for free from the properties of weak submodularity (1.1).

Proof of Lemma 4.2: given L^*, which is an atomic set of size at most r, and $L_t =: L$, which has size t, we find the smallest ε-incoherent atomic set S that is also ε-incoherent with L. such that $\mathbf{span}\,(L \cup S) \supset \mathbf{span}\,(L^*)$. Observe that S can have at most r members. By Theorem 3.1, we have

$$\gamma_{L_t,S} := \frac{\sum_{i \in S} \left(g(L_t \cup \{i\}) - g(L_t)\right)}{g(L_t \cup S) - g(L_t)} \geq \frac{m_{t+r}}{\tilde{M}_1}.$$

We therefore have the following series of inequalities for $D(t)$:

$$\begin{aligned}
D(t) &= g\left(L \cup \{t\}\right) - g\left(L\right) \\
&= f\left(B^{(L \cup \{t\})}\right) - f\left(B^{(L)}\right) \\
&\geq f\left(B^{(L)} + \alpha v_t\right) - f\left(B^{(L)}\right) \quad \text{for any } \alpha \\
&\geq \left\langle \nabla f\left(B^{(L)}\right), \alpha v_t \right\rangle - \alpha^2 \frac{\tilde{M}_1}{2} \quad \text{by restricted smoothness.}
\end{aligned}$$

We now choose $\alpha = \frac{\beta}{\tilde{M}_1} \left\|\nabla f\left(B^{(L)}\right)\right\|_{\mathcal{A}}$, where $\|\cdot\|_{\mathcal{A}}$ denotes the atomic norm. For the coordinate basis vectors, this corresponds to $\|\cdot\|_\infty$; for rank-one matrices, this corresponds to $\|\cdot\|_2$, the operator norm. Since S contains at most r atoms, We have the following inequality

$$\left\|P_S\left(\nabla f\left(B^{(L)}\right)\right)\right\|^2 \leq r \left\|\nabla f\left(B^{(L)}\right)\right\|_{\mathcal{A}}^2.$$

by the assumption on `LMO`, v_t satisfies

$$\left\langle \nabla f\left(B^{(L)}\right), v_t \right\rangle \geq \beta \left\|\nabla f\left(B^{(L)}\right)\right\|_{\mathcal{A}}.$$

Going back to the sequence of lower bounds for $D(t)$, we now have

$$\begin{aligned}
&\left\langle \nabla f\left(B^{(L)}\right), \alpha v_t \right\rangle - \alpha^2 \frac{\tilde{M}_1}{2} \\
\geq\ & \frac{\beta^2}{2\tilde{M}_1} \left\| \nabla f\left(B^{(L)}\right) \right\|_{\mathcal{A}}^2 \\
\geq\ & \frac{\beta^2}{\tilde{M}_1} \frac{1}{r} \left\| P_S\left(\nabla f\left(B^{(L)}\right)\right) \right\|^2 \\
>\ & \beta^2 \eta \frac{m_{t+r}}{\tilde{M}_1} \frac{1}{r} \left(f\left(B^{(L\cup S)}\right) - f\left(B^{(L)}\right)\right) \quad \text{from Theorem 3.1} \\
\geq\ & \beta^2 \eta \frac{m_{t+r}}{\tilde{M}_1} \frac{1}{r} \left(f\left(B^{(L^*)}\right) - f\left(B^{(L)}\right)\right) \quad \text{since } \operatorname{span}(L^*) \subset \operatorname{span}(L \cup S) \\
=\ & \beta^2 \eta \frac{m_{t+r}}{\tilde{M}_1} \frac{1}{r} R(t) \geq \beta^2 \eta \frac{m_{k+r}}{\tilde{M}_1} \frac{1}{r} R(t). \quad \blacksquare
\end{aligned}$$

Proof of Theorem 4.1: we observe that the lemma gives us

$$D(t+1) = R(t) - R(t+1) \geq \beta^2 \eta \frac{m_{k+r}}{\tilde{M}_1} \frac{1}{r} R(t) =: wR(t),$$

which gives us

$$\begin{aligned}
R(t+1) \leq (1-w)R(t) \leq (1-w)^{t+1} R(0) = (1-w/k)^{t+1} g(L^*), \\
R(t) = g(L^*) - g(L_t) \leq (1-w)^t R(0) = (1-w)^t g(L^*), \\
g(L_t) \geq (1-(1-w)^t) g(L^*), \\
g(L_k) \geq \left(1 - \frac{1}{e^{kw}}\right) g(L^*).
\end{aligned}$$

■

4.2 Oblivious Atom Selection

The main setback of GECO$(\mathcal{A}, k, \beta)$ is the fact that the weights of all previously selected atoms must be re-optimized at each iteration. This adds up to $O(r^2)$ re-weighting computations, which can be costly. Unfortunately, optimizing the weights of the chosen atoms is fundamental to proving our approximation guarantees, since our proof of the submodularity ratio stems from bounds on $f(B^{(L\cup\{i\})})$ and $f(B^{(L\cup S)})$, which we recall are the optimal values over linear combinations of the atom sets $L\cup\{i\}$ and $L\cup S$, respectively. Therefore, in order to minimize the number of re-weighting computations, we consider finding r atoms in one go without regard to the change to f inflicted by each individual atom, and then optimize the weights of the atoms as the last step. We prove that, provided the r atoms are chosen properly, this scheme also attains a constant-factor approximation ratio. The algorithm is as follows.

Algorithm 2 Oblivious$(\mathcal{A}, k, \beta)$

$L_0 \leftarrow \emptyset$
Compute $\nabla f(0)$
$L \leftarrow \text{AtomBasis}(\nabla f(0), L_0, \mathcal{A}, k)$
Compute $B^{(L)}$
return $L, B^{(L)}, f(L)$

where AtomBasis$(u, R, \mathcal{A}, k)$ is an algorithm that selects an approximation L of the best k atoms satisfying the ε-Incoherence Condition 1.7 with respect to R that maximize the projection of u onto $\mathbf{span}\,(L)$. Intuitively, AtomBasis finds an r-subspace spanned by ε-incoherent atoms that (approximately) best explains u.

Algorithm 3 AtomBasis$(u, R, \mathcal{A}, k)$

$L \leftarrow \emptyset$
for $i = 1, \ldots, k$ **do**
    $v_i \leftarrow \texttt{LMO}(\beta, u, R \cup L)$
    $L \leftarrow L \cup \{v_i\}$
end for
return L

We remark that AtomBasis is often simple to compute. In the case $\mathcal{A} = \{e_i\} \subset \mathbb{R}^n$, AtomBasis$(u, R, \mathcal{A}, k)$ can be naively implemented by finding the k largest entries of u in absolute value that are outside of R. In the case $\mathcal{A} = \left\{uv^\top\right\} \subset \mathbb{R}^{n \times d}$, AtomBasis$(u, R, \mathcal{A}, k)$ can be computed by finding the top k singular vector pairs under an orthogonality constraint, and SVD can be approximately computed quickly [13]. In general, AtomBasis boils down to linear maximizations over a convex set $\mathbf{conv}\,(\mathcal{A})$ with incoherence/orthogonality constraints.

Theorem 4.3 (Approximation Guarantee for Oblivious) *Let atomic set $\mathcal{A}$ satisfy the conditions described in Theorem 3.1. Let L_r be the solution obtained by Oblivious$(\mathcal{A}, r, \beta)$, and let L^* be the optimal atomic set of size at most r satisfying the rank constraint of the problem. Given f is m_r restricted strong concave and $\tilde{M}_1$ restricted smooth on the sets defined in Theorem 3.1, then*

$$g(L_r) \geq \beta^2 \eta \frac{m_r}{\tilde{M}_1} g(L^*),$$

where $\eta = \frac{1-5(r-1)\sqrt{\varepsilon}}{1+4r^{3/2}\sqrt{\varepsilon}}$.

Observe that if $\mathcal{A}$ satisfies the Orthogonality Condition 1.5, and LMO is perfectly precise, we get the bound $g(L_r) \geq \frac{m_r}{\tilde{M}_1} g(L^*)$. From the inequality $x \geq 1 - \frac{1}{e^x}$, we observe that the approximation guarantee for Oblivious is actually better than the one we derive for GECO, which is probably a byproduct of the proof techniques used. We make use of the following lemma to prove the theorem, which essentially states that AtomBasis picks atoms that allow us to both upper and lower bound the improvement $g(K) \to g(K \cup L)$ with respect to the projection of the gradient at $B^{(K)}$ onto $\mathbf{span}\,(L)$.

Lemma 4.4 *Let K be an arbitrary set of k atoms. Let L be the set of r atoms $\{v_i\}$ returned by AtomBasis$(u, K, \mathcal{A}, r)$, which satisfies the ε-Incoherence Condition 1.7. Then we have*

$$\frac{1-\tau}{1+\delta} \frac{1}{2\tilde{M}_1} \left\| P_L \left(\nabla f \left(B^{(K)} \right) \right) \right\|^2 \leq f(B^{(K \cup L)}) - f(B^{(K)}) \leq \frac{1+\xi}{2m_{k+r}} \left\| P_L \left(\nabla f(B^{(K)}) \right) \right\|^2$$

where $\delta = (r-1)\varepsilon,\ \tau = 5(r-1)\sqrt{\varepsilon},\ \xi = 3\sqrt{r}(k+r)\sqrt{\varepsilon}$.

Proof of Lemma 4.4: The bound on the denominator of the submodularity ratio derived in the proof of Theorem 3.1 gets us an immediate upper bound:

$$f(B^{(K\cup L)}) - f(B^{(K)}) \leq \frac{1+3\sqrt{r}(k+r)\sqrt{\varepsilon}}{2m_{k+r}} \left\| P_L\left(\nabla f(B^{(K)})\right)\right\|^2.$$

To establish the lower bound, we retrace and repurpose the first half of the proof of Theorem 3.1, where we lower bounded $\sum_{j\in S} f(B^{(L\cup\{j\})}) - f(B^{(L)})$. Recall that $B_j^{(K\cup L)} := \left\langle B^{(K\cup L)}, v_j\right\rangle v_j$, $v_j \in L$, such that

$$\begin{aligned} f\left(B^{(K\cup L)}\right) - f\left(B^{(K)}\right) &\geq f\left(B^{(K)} + \sum_{j\in L}\alpha_j B_j^{(K\cup L)}\right) - f\left(B^{(K)}\right) \\ &\geq \sum_{j\in L}\left\langle \nabla f\left(B^{(K)}\right), \alpha_j B_j^{(K\cup L)}\right\rangle - \frac{\tilde{M}_1}{2}\left\|\sum_{j\in L}\alpha_j B_j^{(K\cup L)}\right\|^2 \\ &= \sum_{j\in L}\left\langle \nabla f\left(B^{(K)}\right), \alpha_j B_j^{(K\cup L)}\right\rangle - \frac{\tilde{M}_1}{2}x^\top V^\top V x, \end{aligned}$$

where $V = [v_1 \cdots v_r]$, $x = \left[\alpha_1\left\|B_1^{(K\cup L)}\right\| \ \cdots\ \alpha_r\left\|B_r^{(K\cup L)}\right\|\right]^\top$. By the proof of Lemma 2.2, since L is ε-incoherent

$$V^\top V \preceq (1+\delta)I_r, \quad \delta = (r-1)\varepsilon$$

where $\preceq$ is the Loewner partial order. Thus, replacing $V^\top V$ with $(1+\delta)I_r$, we get

$$\begin{aligned} f\left(B^{(K\cup L)}\right) - f\left(B^{(K)}\right) &\geq \sum_{j\in L}\left\langle \nabla f\left(B^{(K)}\right), \alpha_j B_j^{(K\cup L)}\right\rangle - (1+\delta)\frac{\tilde{M}_1}{2}x^\top x \\ &= \sum_{j\in L}\left(\left\langle \nabla f\left(B^{(K)}\right), \alpha_j B_j^{(K\cup L)}\right\rangle - (1+\delta)\frac{\tilde{M}_1}{2}\alpha_j^2\left\|B_j^{(K\cup L)}\right\|^2\right). \end{aligned}$$

Setting $\alpha_j = \frac{\left\langle \nabla f\left(B^{(K)}\right), B_j^{(K\cup L)}\right\rangle}{(1+\delta)\tilde{M}_1\left\|B_j^{(K\cup L)}\right\|^2}$, we get

$$\begin{aligned} f\left(B^{(K\cup L)}\right) - f\left(B^{(K)}\right) &\geq \frac{1}{2(1+\delta)\tilde{M}_1}\sum_{j\in L}\frac{\left\langle \nabla f\left(B^{(K)}\right), B_j^{(K\cup L)}\right\rangle^2}{\left\|B_j^{(K\cup L)}\right\|^2} \\ &\geq \frac{1-\tau}{1+\delta}\frac{1}{2\tilde{M}_1}\left\|P_L\left(\nabla f\left(B^{(K)}\right)\right)\right\|^2, \quad \tau = 5(r-1)\sqrt{\varepsilon}, \end{aligned}$$

where the last inequality follows from Lemma 2.3. This completes the bound

$$\frac{1-\tau}{1+\delta}\frac{1}{2\tilde{M}_1}\left\|P_L\left(\nabla f\left(B^{(K)}\right)\right)\right\|^2 \leq f(B^{(K\cup L)}) - f(B^{(K)}) \leq \frac{1+\xi}{2m_{k+r}}\left\|P_L\left(\nabla f(B^{(K)})\right)\right\|^2$$

where $\delta = (r-1)\varepsilon,\ \tau = 5(r-1)\sqrt{\varepsilon},\ \xi = 3\sqrt{r}(k+r)\sqrt{\varepsilon}$. ■

Proof of Theorem 4.3: Referring to Lemma 4.4, set $K = L_0 = \emptyset$, and L is the set of r atoms $\{v_k\}$ returned by AtomBasis$(u, K, \mathcal{A}, r)$, which satisfy the ε-Incoherence Condition 1.7. We thus have the following bound on $f(B^{(L)}) - f(0) = g(L)$:

$$\frac{1-\tau}{1+\delta}\frac{1}{2\tilde{M}_1}\left\|P_L\left(\nabla f\left(B^{(K)}\right)\right)\right\|^2 \leq g(L) \leq \frac{1+\xi}{2m_r}\left\|P_L\left(\nabla f(B^{(K)})\right)\right\|^2,$$

where $\delta = (r-1)\varepsilon$, $\tau = 5(r-1)\sqrt{\varepsilon}$, $\xi = 3r^{3/2}\sqrt{\varepsilon}$. By the construction of the AtomBasis algorithm, the elements of L form an ε-incoherent basis for an r-dimensional space that maximizes the norm of $\nabla f(0)$ projected onto it, or at least gets a factor of β close to the maximmum accounting for the precision of the linear oracle. In other words

$$\|P_L\left(\nabla f(0)\right)\|^2 \geq \beta^2 \max_{\substack{L\subset\mathcal{A}\\|L|=r}} \|P_L\left(\nabla f(0)\right)\|^2 =: \beta^2 P^2.$$

Let L^* be the optimal ε-incoherent atomic solution of size r: $L^* = \arg\max g(L)$. Applying Lemma 4.4 on L^*:

$$P^2 \geq \|P_{L^*}\left(\nabla f(0)\right)\|^2 \geq \frac{2m_r}{1+\xi} g(L^*).$$

We establish the approximation ratio through the following series of inequalities:

$$\begin{aligned}
g(L) &\geq \frac{1-\tau}{1+\delta}\frac{1}{2\tilde{M}_1}\left\|P_L\left(\nabla f\left(B^{(K)}\right)\right)\right\|^2 \\
&\geq \beta^2\frac{1-\tau}{1+\delta}\frac{1}{2\tilde{M}_1}P^2 \\
&\geq \beta^2\frac{1-\tau}{1+\delta}\frac{1}{2\tilde{M}_1}\|P_{L^*}\left(\nabla f(0)\right)\|^2 \\
&\geq \beta^2\frac{1-\tau}{(1+\delta)(1+\xi)}\frac{m_r}{\tilde{M}_1}g(L^*) \\
&\geq \beta^2\frac{1-5(r-1)\sqrt{\varepsilon}}{1+4r^{3/2}\sqrt{\varepsilon}}\frac{m_r}{\tilde{M}_1}g(L^*).
\end{aligned}$$

■

5 Applications

When $\mathcal{A}$ is the set of coordinate vectors $\{e_i\} \subset \mathbb{R}^n$ or the set of low-rank matrices $\left\{uv^\top\right\} \subset \mathbb{R}^{n\times d}$, variants of GECO and Oblivious have been shown to perform well in a multitude of important settings. These include: feature selection [3], collaborative filtering [13, 15], image recovery [15], and clustering [7].

The implementations of GECO and Oblivious in the aforementioned settings easily extend to problems like (disjoint) group lasso and tensor completion. For the former problem, GECO can easily be implemented by finding the best remaining group at each iteration rather than the best coordinate as in sparse recovery. Since the groups are disjoint, they satisfy the Orthogonality Condition 1.5. At a high-level, GECO and Oblivious for tensor completion and matrix completion have very implementations. An iteration of GECO for matrix completion reduces to finding the top singular vector pair of the gradient matrix, and Oblivious reduces to evaluating the r-singular value decomposition of the gradient at 0. This subroutine is not so simple in tensor completion. However,

there has been recent work [6] that demonstrates finding the best rank-one tensor approximation admits a semi-definite relaxation, whose solution can be effectively computed. This SDP relaxation heuristically almost always returns a rank-one solution that corresponds with the optimal rank-one tensor both for structured and random examples, and if not, the rank-one tensor can be retrieved by adding small perturbations to the original tensor. This implies that GECO and Oblivious can be feasibly implemented for tensor completion problems, which we remark satisfy the Orthogonality Condition 1.5.

An interesting implication of the Incoherence Condition 1.7 is that certain infinite atomic sets, specifically those living in locally compact metric spaces, can be systematically discretized into a finite set without significantly affecting the performance guarantees of the proposed greedy algorithms. For example, the set of unit rank-one matrices $\mathcal{A} = \left\{uv^\top\right\}$ is uncountably infinite. However, the matrix space $\mathbb{R}^{n\times d}$ is a finite-dimensional Euclidean space, and thus is locally compact. Given that the sparsity constraint on a restricted strongly concave objective is r, we choose a sufficiently small $\varepsilon < r^{-5}$ and form an ε-covering for the set of rank-one matrices. Let us collect the centers of the balls in the ε-covering to form $\mathcal{A}'$. Observe that by construction $\mathcal{A}'$ satisfies the Incoherence Condition 1.7, and thus GECO and Oblivious attain similar approximation guarantees.

6 Discussion

We have connected the problem of greedy sparse atomic optimization to submodular optimization and demonstrated that two distinct flavors of greedy algorithms that obtain sparse solutions within a multiplicative factor of the optimal. This unifies and generalizes existing results in sparse recovery, low-rank matrix optimization, and group lasso.

References

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

[2] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[3] 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.

[4] Rémi Gribonval and Pierre Vandergheynst. On the exponential convergence of matching pursuits in quasi-incoherent dictionaries. *IEEE Transactions on Information Theory*, 52(1):255–261, 2006.

[5] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 1990.

[6] Bo Jiang, Shiqian Ma, and Shuzhong Zhang. Tensor principal component analysis via convex optimization. *Mathematical Programming*, 150(2):423–457, 2015.

[7] Rajiv Khanna, Ethan Elenberg, Alexandros G Dimakis, and Sahand Negahban. On approximation guarantees for greedy low rank optimization. *arXiv preprint arXiv:1703.02721*, 2017.

[8] Rajiv Khanna, Ethan Elenberg, Alexandros G Dimakis, Sahand Negahban, and Joydeep Ghosh. Scalable greedy feature selection via weak submodularity. *arXiv preprint arXiv:1703.02723*, 2017.

[9] Francesco Locatello, Rajiv Khanna, Michael Tschannen, and Martin Jaggi. A unified optimization view on generalized matching pursuit and frank-wolfe. *arXiv preprint arXiv:1702.06457*, 2017.

[10] Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484, 2013.

[11] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[12] Nikhil Rao, Parikshit Shah, and Stephen Wright. Forward–backward greedy algorithms for atomic norm regularization. *IEEE Transactions on Signal Processing*, 63(21):5798–5811, 2015.

[13] Shai Shalev-Shwartz, Alon Gonen, and Ohad Shamir. Large-scale convex minimization with a low-rank constraint. *arXiv preprint arXiv:1106.1622*, 2011.

[14] Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004.

[15] Zheng Wang, Ming-Jun Lai, Zhaosong Lu, Wei Fan, Hasan Davulcu, and Jieping Ye. Rank-one matrix pursuit for matrix completion. In *International Conference on Machine Learning*, pages 91–99, 2014.