Atomic Sets and Extending Machinery from Low-Rank Optimization

Thomas Zhang

Contents

1 Atomic Sets: Definitions and Basic Properties

An atomic set $\mathcal{A}$ is a (possibly infinite) collection of "atoms", which can be anything from basis vectors, low-rank matrices to dirac measures. The reason we care about atomic sets is that they admit ideas about sparsity quite nicely, e.g. we can generalize the problem of regularization.

For the atomic norm to exist, we require that atomic sets are symmetric, i.e. if $a \in \mathcal{A}$, then $-a \in \mathcal{A}$. Given a member x of the same space that the atoms live in, we might wonder how to represent x in terms of the atoms. Say that

$$x = \sum_{\alpha \in \mathcal{A}_t} c_\alpha \alpha, \quad \mathcal{A}_t \subset \mathcal{A}$$

We immediately observe that since $\mathcal{A}$ is symmetric, we can always represent x as a conic combination, i.e. $c_\alpha \geq 0$ for all $\alpha \in \mathcal{A}_t$. If for some choice $\tau \geq 0$, we have $\sum_{\alpha \in \mathcal{A}_5} c_\alpha \leq \tau$, we can also say

$$x \in \mathbf{conv}\left(\mathcal{A}_t, \tau\right).$$

Echoing some concepts from functional analysis, atomic sets induce a norm; namely, since atomic sets are symmetric, one can define a gauge functional with respect to the convex hull of a atomic set, and that turns out to be a norm:

$$\|x\|_{\mathcal{A}} := \inf\left\{t \mid x \in \mathbf{conv}\left(t\mathcal{A}\right)\right\}.$$

Sometimes we might be interested in the dual atomic norm, which is defined in the following way

$$\|x\|_{\mathcal{A}}^* := \sup_{\|u\|_{\mathcal{A}} \leq 1} \langle u, x \rangle,$$

if we are concerned with recovering the minimizing direction/atom in an optimization problem.

What is the point of atomic sets?

Atomic sets in some sense are a general model designed to capture the sparsifying, yet convex, structure of things like ℓ^1 regularization and matrix completion. In the case of ℓ^1 regularization, the ultimate goal is to retrieve a solution that has a small support (which can be seen as ℓ^0), but that is not a convex problem (since ℓ^0 is not a norm). In other words, we want to solve some problem like this: $f(x)$ is convex

$$\begin{aligned} &\min\ f(x) \\ &\text{s.t. } \operatorname{supp}(x) \leq k. \end{aligned}$$

If we treat the elementary unit vectors as the atoms of $\mathcal{A}$, then we observe that the unit atomic norm ball, in other words $\mathbf{conv}(\mathcal{A})$, is identical to the ℓ^1 norm ball. The properties of the ℓ^1 norm ball gives us an idea of why we would want to find a generalization that preserves them. One may recall coordinate descent for a strongly convex, differentiable objective, where the optimization subroutine at each iteration looks something like

$$\begin{aligned} \Delta x_{CD} &= \arg\min_{\|v\|_1=1} \left\langle \nabla f(x^{(t)}), v \right\rangle \\ &= \arg\min_{\|v\|_1=1} \sum_{i=1}^{n} \frac{\partial f}{\partial x_i^{(t)}} v_i \\ &= -\arg\max_i \left\{ \left| \frac{\partial f}{\partial x_i^{(t)}} e_i \right| \right\} \\ &\text{where } v_{\min} \text{ would be the corresponding } e_i. \end{aligned}$$

As one can see, imposing an ℓ^1 restriction on the subroutine induces the descent direction to be a coordinate direction, i.e. a sparse direction. More complex phenomena related to the ℓ^1 norm are described in the field of compressed sensing. However, from a geometric standpoint, many of these phenomena can be attributed to the "polytopal" structure of the ℓ^1 unit ball. In fact, many phenomena that could be described as "sparsifying" can be roughly ascribed to the underlying polytopal/ridge structures. In our above example, it is clear to see why if we had replaced the ℓ^1 unit ball with some other (I guess finite for now) atomic set, we would also get a similar result: since the atomic-norm ball is the convex hull of finitely many vectors, it is a polytope. Since the function $\langle \nabla f(x), \cdot \rangle$ is trivially a concave function, the minimum on a convex set is attained at a extreme point (a.k.a. vertex). Therefore, we could have easily replaced the constraint set $\|v\|_1 = 1$ with $\|v\|_{\mathcal{A}} = 1$ and instead of the minimizing direction being a coordinate direction, we would have gotten a minimizing direction that is an atom.

2 Inspiration from Low-Rank Matrix Optimization

In Khanna et al's paper (**cite**) (among other papers [**cite**]), the connection is drawn between weak submodularity and the atomic set of rank-one matrices, and a greedy algorithm (Section 4.2) is proposed. However, the convergence properties of the algorithm relied heavily on the special properties of rank-one matrices and SVD. Our goal in this section is to first revise the proof of the weak submodularity of a restricted-strong-concave utility function ℓ (Theorem 2 in [**cite**]) so that it fits better under a more general atomic set narrative. Then, we will see if we can extend the OMPSel algorithm to atomic sets (under some conditions mentioned later).

2.1 Generalized Proof of Weak Submodularity over Rank-One Matrices

Background

The problem we address in this section is the following:

$$\begin{aligned}(\arg)\max_{X}\ & \ell(X)\\ \text{s.t. }& \operatorname{rank}(X)\le r.\end{aligned}$$

A core assumption is that the function we want to optimize (maximize) satisfies what is called "restricted strong concavity" (convexity for minimization) over the set we care about, in this case low-rank matrices.

Definition 2.1 (Low Rank Restricted Strong Concavity, Restricted Smoothness) *A function $\ell : \mathbb{R}^{n\times d} \to \mathbb{R}$ is restricted strong concave with parameter m_Ω and restricted smooth with parameter M_Ω if for all $X, Y \in \Omega \subset \mathbb{R}^{n\times d}$,*

$$\begin{aligned}-\frac{m_\Omega}{2}\|Y-X\|_F^2 &\ge \ell(Y)-\ell(X)-\langle\nabla\ell(X), Y-X\rangle\\ &\ge -\frac{M_\Omega}{2}\|Y-X\|_F^2.\end{aligned}$$

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

$$M_{\Omega'} \le M_\Omega, \quad m_{\Omega'} \ge m_\Omega. \tag{1}$$

In related work (**cite**), it is shown that restricted strong concave/convex functions are closely related to the concept of weak submodularity, which generalizes the concept of submodularity of a set function.

Definition 2.2 (Weak Submodularity Ratio) *Let $S, L \subset [p]$ be two disjoint sets, and $f(\cdot) : [p] \to \mathbb{R}$. The submodularity ratio of L with respect to S is defined*

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

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 weak submodularity ratio of a set $U \subseteq [p]$ with respect to a number k:

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

In order to translate the RSC-RS function utility function $\ell(\cdot)$ to a weakly submodular function f, We introduce a centered set function based on $\ell(\cdot)$, which we will want to show satisfies weak submodularity if $\ell(\cdot)$ satisfies RSC and RS.

Definition 2.3 *Given atomic set L, we define the set function*

$$f(L) = \max_{H \in diag(\mathbb{R}^{|L|})} \ell\left(U_L^\top H V_L\right) - \ell(0),$$

such that the low-rank maximization problem can be reformulated

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

In its current state, given a restricted strong concave function $\ell(X)$, $X \in M_{m,n}$, the proof of its weak submodularity relies on the matrix structure of X and SVD. Namely, the proof defines and uses

$$\begin{aligned}\hat{H}^{(L)} &:= \operatorname*{arg\,max}_{H\in\mathbb{R}^{|L|\times|L|}} \ell\left(U_L H V_L^\top\right) \\ B^{(L)} &:= U_L \hat{H}^{(L)} V_L^\top,\end{aligned}$$

where L is a set of indices corresponding to vectors in $\mathbb{R}^m$ and and $\mathbb{R}^n$ such that U_L is $|L|$ columns of $\mathbb{R}^m$ vectors, and $V_L^\top$ is $|L|$ rows of $\mathbb{R}^n$ vectors. However, since H is an $R^{|L|\times|L|}$ matrix, if we expand $B^{(L)}$, we get

$$B^{(L)} = \sum_{i=1}^{|L|} \alpha_{ii} u_i v_i^\top + \sum_{i\neq j} \alpha_{ij} u_i v_j^\top,$$

where α_{ij} are entries in H. This is where the reliance on the matrix structure of the problem comes in. If we are given another index set S corresponding to vectors orthogonal to L, establishing the weak submodularity bounds requires projections onto the span of S, and the expansion discussed does not necessarily exist for more abstract vector spaces. However, if we instead define $\hat{H}^{(L)}$ and $B^{(L)}$:

$$\begin{aligned}\hat{H}^{(L)} &:= \operatorname*{arg\,max}_{H\in\mathrm{diag}(\mathbb{R}^{|L|})} \ell\left(U_L H V_L^\top\right) \\ B^{(L)} &:= U_L \hat{H}^{(L)} V_L^\top \\ &= \sum_{i=1}^{|L|} \alpha_i u_i v_i^\top,\end{aligned}$$

such that $\hat{H}^{(L)}$ is diagonal and hence no cross terms appear in $B^{(L)}$, we can treat $\{u_i v_i^\top\}_{i\in L}$ as a sparse atomic set. The only real barrier to generalizing the upcoming proof to general atomic sets is the fact that we are able to pick orthogonal $u_i v_i^\top$ for the atomic set. Note that this is also an underlying structure for ℓ^1 sparse optimization, where the atoms are the (orthogonal) elementary basis vectors. We will discuss a relaxation of this condition later (hopefully).

The Theorem and Proof

Theorem 2.4 (Weak Submodularity Ratio) *Let L be a set of k rank 1 atoms and S be a set of r rank 1 atoms that have been sequentially orthogonalized against L. If $\ell(\cdot)$ is m_i-strongly concave over matrices of rank i, and $\tilde{M}_1$-smooth over the set $\tilde{\Omega} := \{(X, Y) : rank(X - Y) = 1\}$, then we have the following submodularity ratio for $f(\cdot)$:*

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

Proof: essentially this boils down to lower bounding the numerator and upper bounding the denominator. An important assumption is that the atoms in S are all mutually orthogonal and are orthogonal to the space spanned by L. We now lower bound the numerator.

We observe that

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

and thus it suffices to deal with $\ell(\cdot)$. We denote $B_{ii},\ i \in S$ as the projection of B onto the rank-one matrix $u_i v_i^\top$:

$$B_{ii} = \left\langle B, u_i v_i^\top \right\rangle u_i v_i^\top,$$

where $\|u_i\|_2 = \|v_i\|_2 = 1$ such that $u_i v_i^\top$ is normalized

$$\left\|u_i v_i^\top\right\|_F^2 = \mathrm{tr}\left(v_i u_i^\top u_i v_i^\top\right) = 1.$$

We now have the following series of inequalities

$$\begin{aligned}
\ell\left(B^{(L\cup\{i\})}\right) - \ell\left(B^{(L)}\right) &\geq \ell\left(B^{(L)} + \alpha_i B_{ii}^{(L\cup S)}\right) - \ell\left(B^{(L)}\right) \\
&\geq \left\langle \nabla\ell\left(B^{(L)}\right), \left(B^{(L)} + \alpha_i B_{ii}^{(L\cup S)}\right) - B^{(L)} \right\rangle \\
&\quad - \frac{\tilde{M}_1}{2}\left\|\left(B^{(L)} + \alpha_i B_{ii}^{(L\cup S)}\right) - B^{(L)}\right\|_F^2 \\
&= \left\langle \nabla\ell\left(B^{(L)}\right), \alpha_i B_{ii}^{(L\cup S)} \right\rangle - \alpha_i^2 \frac{\tilde{M}_1}{2}\left\|B_{ii}^{(L\cup S)}\right\|_F^2,
\end{aligned}$$

where α_i is an arbitrary constant. Let us set

$$\alpha_i = \frac{\left\langle \nabla\ell\left(B^{(L)}\right), B_{ii}^{(L\cup S)} \right\rangle}{\tilde{M}_1\left\|B_{ii}^{(L\cup S)}\right\|_F^2}.$$

Plugging this back into the inequality, we get

$$
\begin{aligned}
&\left\langle \nabla \ell\left(B^{(L)}\right), \alpha_i B_{ii}^{(L\cup S)} \right\rangle - \alpha_i^2 \frac{\tilde{M}_1}{2} \left\| B_{ii}^{(L\cup S)} \right\|_F^2 \\
&= \left\langle \nabla \ell(B^{(L)}), \frac{\left\langle \nabla \ell\left(B^{(L)}\right), B_{ii}^{(L\cup S)} \right\rangle}{\tilde{M}_1 \left\| B_{ii}^{(L\cup S)} \right\|_F^2} B_{ii}^{(L\cup S)} \right\rangle - \left(\frac{\left\langle \nabla \ell\left(B^{(L)}\right), B_{ii}^{(L\cup S)} \right\rangle}{\tilde{M}_1 \left\| B_{ii}^{(L\cup S)} \right\|_F^2} \right)^2 \frac{\tilde{M}_1}{2} \left\| B_{ii}^{(L\cup S)} \right\|_F^2 \\
&= \frac{\left\langle \nabla \ell\left(B^{(L)}\right), B_{ii}^{(L\cup S)} \right\rangle^2}{\tilde{M}_1 \left\| B_{ii}^{(L\cup S)} \right\|_F^2} - \frac{1}{2} \frac{\left\langle \nabla \ell\left(B^{(L)}\right), B_{ii}^{(L\cup S)} \right\rangle}{\tilde{M}_1 \left\| B_{ii}^{(L\cup S)} \right\|_F^2} \\
&= \frac{1}{2} \frac{\left\langle \nabla \ell\left(B^{(L)}\right), B_{ii}^{(L\cup S)} \right\rangle^2}{\tilde{M}_1 \left\| B_{ii}^{(L\cup S)} \right\|_F^2}.
\end{aligned}
$$

Therefore, we have established that

$$
\ell\left(B^{(L\cup\{i\})}\right) - \ell\left(B^{(L)}\right) \geq \frac{1}{2} \frac{\left\langle \nabla \ell\left(B^{(L)}\right), B_{ii}^{(L\cup S)} \right\rangle^2}{\tilde{M}_1 \left\| B_{ii}^{(L\cup S)} \right\|_F^2}.
$$

The next step is to sum this over all $i \in S$:

$$
\begin{aligned}
\sum_{i\in S} \ell\left(B^{(L\cup\{i\})}\right) - \ell\left(B^{(L)}\right) &\geq \sum_{i\in S} \frac{1}{2} \frac{\left\langle \nabla \ell\left(B^{(L)}\right), B_{ii}^{(L\cup S)} \right\rangle^2}{\tilde{M}_1 \left\| B_{ii}^{(L\cup S)} \right\|_F^2} \\
&= \frac{1}{2\tilde{M}_1} \sum_{i\in S} \frac{\left\langle \nabla \ell\left(B^{(L)}\right), B_{ii}^{(L\cup S)} \right\rangle^2}{\left\| B_{ii}^{(L\cup S)} \right\|_F^2}.
\end{aligned}
$$

The last expression looks awfully like something that results from an orthogonal projection. Let us denote $P_S(X)$ the projection of X onto the span of S, i.e. $\operatorname{span}\left(\left\{u_i v_i^\top\right\}\right)$. Let $\left\{c_i u_i v_i^\top\right\}$ be a basis for S (where we will draw a parallel between c_i and $B_{ii}^{(L\cup S)}$).

$$
\begin{aligned}
P_S\left(\nabla \ell\left(B^{(L)}\right)\right) &= \sum_{i\in S} \frac{\left\langle \nabla \ell\left(B^{L}\right), c_i u_i v_i^\top \right\rangle}{\left\langle c_i u_i v_i^\top, c_i u_i v_i^\top \right\rangle} c_i u_i v_i^\top \\
&= \sum_{i\in S} \frac{\left\langle \nabla \ell\left(B^{L}\right), c_i u_i v_i^\top \right\rangle}{c_i} u_i v_i^\top.
\end{aligned}
$$

Note that $B_{ii}^{(L\cup S)} = \left\langle B^{(L\cup S)}, u_i v_i^\top \right\rangle u_i v_i^\top$, such that $c_i = \left\langle B^{(L\cup S)}, u_i v_i^\top \right\rangle = \left\| B_{ii}^{(L\cup S)} \right\|_F$. Plug-

ging this in, we get

$$P_S\left(\nabla\ell\left(B^{(L)}\right)\right) = \sum_{i\in S} \frac{\left\langle \nabla\ell\left(B^{(L)}\right), B_{ii}^{(L\cup S)}\right\rangle}{\left\|B_{ii}^{(L\cup S)}\right\|_F} u_i v_i^\top$$

$$\begin{aligned}\left\|P_S\left(\nabla\ell\left(B^{(L)}\right)\right)\right\|_F^2 &= \left\langle \sum_{i\in S} \frac{\left\langle \nabla\ell\left(B^{(L)}\right), B_{ii}^{(L\cup S)}\right\rangle}{\left\|B_{ii}^{(L\cup S)}\right\|_F} u_i v_i^\top, \sum_{i\in S} \frac{\left\langle \nabla\ell\left(B^{(L)}\right), B_{ii}^{(L\cup S)}\right\rangle}{\left\|B_{ii}^{(L\cup S)}\right\|_F} u_i v_i^\top \right\rangle \\ &= \sum_{i\in S} \frac{\left\langle \nabla\ell\left(B^{(L)}\right), B_{ii}^{(L\cup S)}\right\rangle^2}{\left\|B_{ii}^{(L\cup S)}\right\|_F^2}.\end{aligned}$$

Therefore, we get

$$\sum_{i\in S} \ell\left(B^{(L\cup\{i\})}\right) - \ell\left(B^{(L)}\right) \geq \sum_{i\in S} \frac{\left\langle \nabla\ell\left(B^{(L)}\right), B_{ii}^{(L\cup S)}\right\rangle^2}{\left\|B_{ii}^{(L\cup S)}\right\|_F^2} \tag{2}$$

$$= \frac{1}{2\tilde{M}_1}\left\|P_S\left(\nabla\ell\left(B^{(L)}\right)\right)\right\|_F^2. \tag{3}$$

We are done with the numerator. Now we want to bound the denominator. By the restricted strong concavity by $\ell(\cdot)$:

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

Therefore, we can use this to bound the denominator

$$\begin{aligned}\ell\left(B^{(L\cup S)}\right) - \ell\left(B^{(L)}\right) &\leq \left\langle \nabla\ell\left(B^{(L)}\right), B^{(L\cup S)} - B^{(L)}\right\rangle - \frac{m_{k+r}}{2}\left\|B^{(L\cup S)} - B^{(L)}\right\|_F^2 \\ &\leq \max_{\substack{X = U_{(L\cup S)} H V_{(L\cup S)}^\top \\ H\in \mathrm{diag}(\mathbb{R}^{|L\cup S|})}} \left\langle \nabla\ell\left(B^{(L)}\right), X - B^{(L)}\right\rangle - \frac{m_{k+r}}{2}\left\|X - B^{(L)}\right\|_F^2.\end{aligned}$$

Now we use the fact that $B^{(L)}$ is by construction the argmax of ℓ over $\mathrm{span}\,(L)$. Therefore, by the restricted strong concavity of $\ell(\cdot)$, we have

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

since $P_L(X) - B^{(L)} \in \mathrm{span}\,(L)$. We want to know what the optimal X looks like. First, we observe that X must be of the form

$$X = B^{(L)} + X_S,$$

where X_S is in the span of 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 $\frac{m_{k+r}}{2}\left\|X - B^{(L)}\right\|_F^2$. We

want X_L to cancel out $B^{(L)}$, since

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

This leaves us to determine X_S. Notice that our problem has been reduced to

$$\begin{aligned}&\max_{X_S} \left\langle \nabla \ell\left(B^{(L)}\right), X_S\right\rangle - \frac{m_{k+r}}{2}\|X_S\|_F^2 \\ =\ &\max_{X_S} \left\langle P_S\left(\nabla \ell\left(B^{(L)}\right)\right), X_S\right\rangle - \frac{m_{k+r}}{2}\|X_S\|_F^2,\end{aligned}$$

since $X_S \in \operatorname{span}(S)$. We observe that the optimal X_S will be collinear with $P_S\left(\nabla \ell\left(B^{(L)}\right)\right)$, in which case this reduces to a calculus problem:

$$\max_c\ c\left\|P_S\left(\nabla \ell\left(B^{(L)}\right)\right)\right\|_F^2 - c^2\frac{m_{k+r}}{2}\left\|P_S\left(\nabla \ell\left(B^{(L)}\right)\right)\right\|_F^2.$$

It is simple to verify that the maximizer is

$$X^* = c^* P_S\left(\nabla \ell\left(B^{(L)}\right)\right) = \frac{1}{m_{k+r}} P_S\left(\nabla \ell\left(B^{(L)}\right)\right).$$

Plugging this back into the inequality for the denominator, we get

$$\begin{aligned}\ell\left(B^{(L\cup S)}\right) - \ell\left(B^{(L)}\right) &\le \left\langle \nabla \ell\left(B^{(L)}\right), \frac{1}{m_{k+r}} P_S\left(\nabla \ell\left(B^{(L)}\right)\right)\right\rangle - \frac{m_{k+r}}{2}\left\|\frac{1}{m_{k+r}} P_S\left(\nabla \ell\left(B^{(L)}\right)\right)\right\|_F^2 \\ &\le \frac{1}{2m_{k+r}}\left\|P_S\left(\nabla \ell\left(B^{(L)}\right)\right)\right\|_F^2.\end{aligned}$$

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

$$\begin{aligned}\gamma_{L,S} &:= \frac{\sum_{i\in S}\left(f(L\cup\{i\}) - f(L)\right)}{f\left(L\cup S\right) - f(L)} \\ &= \frac{\sum_{i\in S}\left(\ell(B^{(L\cup\{i\})} - \ell(B^{(L)})\right)}{\ell\left(B^{(L\cup S)}\right) - \ell(B^{(L)})} \\ &\ge \frac{m_{k+r}}{\tilde{M}_1} \quad \blacksquare\end{aligned}$$

2.2 Proving Greedy Algorithm Gets Good Bounds

We now want to show that we can use weak submodularity to obtain approximation guarantees on a simple greedy algorithm that depend only on the underlying rank of the problem. Since we rely very little on the matrix structure, the hope is that these approximation guarantees will transfer easily to simple greedy algorithms for more general sparse atomic

optimization. Given the definitions from above, we recall that the problem we are trying to solve is

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

where L are sets of rank-one matrices $\left\{u_i v_i^\top\right\}$. We propose the following algorithm known as GECO (Greedy Efficient Component Optimization) (**cite**), where the parameters are GECO(n, d, k, β): n is the dimension of the left singular vectors, d the dimension of the right vectors ($\mathbb{R}^{n\times d}$ is the ambient space), k is the sparsity parameter, and β is the precision parameter for the linear oracle in the algorithm.

Algorithm 1 GECO(n, d, k, β)

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

$\texttt{OMPSel}(\beta, L)$ is a linear oracle that given precision parameter β and atomic set L returns atom $\hat{u}\hat{v}^\top$ such that

$$\left\langle \nabla\ell(B^{(L)}), \hat{u}\hat{v}^\top \right\rangle \geq \beta \max_{uv^\top \perp L} \left\langle \nabla\ell(B^{(L)}), uv^\top \right\rangle. \tag{4}$$

We choose this particular oracle because it can be easily computed at least in the matrix case by computing the left and right singular vectors corresponding to the largest singular value of $\nabla\ell\left(B^{(L)}\right) \leftrightarrow \left\|\nabla\ell\left(B^{(L)}\right)\right\|_2$. We now present the main result of this subsection

Theorem 2.5 (Approximation Guarantee for GECO) *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 $\ell(\cdot)$ is m_{k+r} restricted strong concave on the set of matrices with rank less than or equal to $k + r$, and $\tilde{M}_1$ restricted smooth on the set of matrices $\tilde{\Omega}$. Then*

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

where $w = \beta^2 \frac{m_{k+r}}{\tilde{M}_1} \frac{k}{r}$.

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. This is an improvement on certain results (**cite**) that show a global linear convergence, except the constant factor approximation depends on n or d times, i.e. the dimension of the problem.

Toward proving the above theorem, we introduce some notation and prove a key lemma. Define $D(t) = f(L_t) - f(L_{t-1}),\ D(0) := 0$ as the improvement after the t-th iteration of the

algorithm. Let us also define $R(t) = f(L^*) - f(L_t)$ 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 2.6 *At iteration t, we have the following lower bound on the incremental gain*

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

where the second inequality comes for free from the properties of weak submodularity (1).

Proof of Lemma 2.6: given L^*, which is an atomic set of size at most r, and $L_t =: L$, which has size t, we sequentially orthogonalize L^* with respect to L_t, and call the resulting set S. Observe that S can have at most r members. By Theorem 2.4, we have

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

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

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

We now choose $\alpha = \frac{\beta}{\tilde{M}_1} \left\| \nabla \ell\left(B^{(L)}\right) \right\|_2$, where $\|\cdot\|_2$ is the operator induced-2-norm, i.e. the largest singular value σ_1. Expressing $\nabla \ell\left(B^{(L)}\right)$ by its singular value decomposition

$$\nabla \ell\left(B^{(L)}\right) = \sum_{i=1}^{\min(n,d)} \sigma_i u_i {v_i}^\top.$$

It is therefore simple to verify that

$$\left\langle \nabla \ell\left(B^{(L)}\right), u v^\top \right\rangle = u^\top \nabla \ell\left(B^{(L)}\right) v \leq \sigma_1 = \left\| \nabla \ell\left(B^{(L)}\right) \right\|_2.$$

This implies that the projection of $\nabla \ell\left(B^{(L)}\right)$ onto the span of S satisfies

$$\begin{aligned} P_S\left(\nabla \ell\left(B^{(L)}\right)\right) &= \sum_{j=1}^{r} \left\langle \nabla \ell\left(B^{(L)}\right), u_j v_j^\top \right\rangle u_j v_j^\top \\ \implies \left\| P_S\left(\nabla \ell\left(B^{(L)}\right)\right) \right\|_F^2 &\leq r \left\| \nabla \ell\left(B^{(L)}\right) \right\|_2. \end{aligned}$$

We also recall that by `OMPSel`, $u_t v_t^\top$ satisfies

$$\left\langle \nabla \ell\left(B^{(L)}\right), u_t v_t^\top \right\rangle \geq \beta \left\| \nabla \ell\left(B^{(L)}\right) \right\|_2.$$

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

$$\begin{aligned}
&\left\langle \nabla \ell\left(B^{(L)}\right), \alpha u_t v_t^\top \right\rangle - \alpha^2 \frac{\tilde{M}_1}{2} \\
\geq\ & \frac{\beta^2}{2\tilde{M}_1} \left\| \nabla \ell\left(B^{(L)}\right) \right\|_2 \\
\geq\ & \frac{\beta^2}{\tilde{M}_1} \frac{1}{r} \left\| P_S\left(\nabla \ell\left(B^{(L)}\right)\right) \right\|_F^2 \\
\geq\ & \frac{\beta^2 m_{t+r}}{\tilde{M}_1} \frac{1}{r} \left(\ell\left(B^{(L\cup S)}\right) - \ell\left(B^{(L)}\right)\right) \quad \text{from proof of Theorem 2.4} \\
\geq\ & \beta^2 \frac{m_{t+r}}{\tilde{M}_1} \frac{1}{r} \left(\ell\left(B^{(L^*)}\right) - \ell\left(B^{(L)}\right)\right) \quad \text{since } \operatorname{span}(L^*) \subset \operatorname{span}(L \cup S) \\
=\ & \beta^2 \frac{m_{t+r}}{\tilde{M}_1} \frac{1}{r} R(t) \geq \beta^2 \frac{m_{k+r}}{\tilde{M}_1} \frac{1}{r} R(t). \quad \blacksquare
\end{aligned}$$

Proof of Theorem 2.5: we observe that the lemma gives us

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

which gives us

$$\begin{aligned}
R(t+1) \leq (1 - w/k) R(t) &\leq (1 - w/k)^{t+1} R(0) = (1 - w/k)^{t+1} f(L^*) \\
R(t) = f(L^*) - f(L_t) &\leq (1 - w/k)^t R(0) = (1 - w/k)^t f(L^*) \\
f(L_t) &\geq (1 - (1 - w/k)^t) f(L^*) \\
f(L_k) &\geq \left(1 - \frac{1}{e^w}\right) f(L^*) \\
&= \left(1 - \frac{1}{e^{\beta^2 \frac{m_{k+r}}{\tilde{M}_1} \frac{k}{r}}}\right) f(L^*). \quad \blacksquare
\end{aligned}$$