Sparse Optimization on General Atomic Sets: Greedy and Forward-Backward Algorithms

Thomas Zhang[1]

[1]Yale University , thomas.zhang@yale.edu

Abstract

We consider the problem of sparse atomic optimization, where the notion of "sparsity" is generalized to meaning some linear combination of few atoms. The definition of atomic set is very broad; popular examples include the standard basis, low-rank matrices, overcomplete dictionaries, permutation matrices, orthogonal matrices, etc. The model of sparse atomic optimization therefore includes problems coming from many fields, including statistics, signal processing, machine learning, computer vision and so on. Specifically, we consider the problem of maximizing a restricted strongly convex (or concave), smooth function restricted to a sparse linear combination of atoms. We extend recent work that establish linear convergence rates of greedy algorithms on restricted strongly concave, smooth functions on sparse vectors to the realm of general atomic sets, where the convergence rate involves a novel quantity: the "sparse atomic condition number". This leads to the strongest known multiplicative approximation guarantees for various flavors of greedy algorithms for sparse atomic optimization; in particular, we show that in many settings of interest the greedy algorithm can attain strong approximation guarantees while maintaining sparsity. Furthermore, we introduce a scheme for forward-backward algorithms that achieves the same approximation guarantees. Secondly, we define an alternate notion of weak submodularity, which we show is tightly related to the more familiar version that has been used to prove earlier linear convergence rates. We prove analogous multiplicative approximation guarantees using this alternate weak submodularity, and establish its distinct identity and applications.

Keywords: Greedy algorithms, convex optimization, atomic sets, weak submodularity, approximation ratios, feature selection.

1 Background and Definitions

Sparsity in its many forms is central to a variety of problems across statistics and computer science. In general, these problems usually require the estimation of some model whose dimension is much higher than the number of measurements that can be feasibly made. However, if one has the belief or knowledge that the model is constrained in some way that makes it feasibly estimable by few measurements, then sparse optimization becomes a problem of interest. The notion of sparsity differs from problem to problem: in linear least squares, one seeks sparsity in the support of the coefficient vector; in matrix completion, one seeks sparsity in the spectrum of a matrix; in ranked elections, one seeks sparsity in the number of permutations. Atomic sets are an incorporating model that often elegantly capture these different notions of "sparsity".

Given an inner product space $\mathcal{H}$, e.g. $\mathbb{R}^n$, $M^{m\times n}(\mathbb{R})$, an atomic set is a (possibly uncountable) set of vectors $\mathcal{A} = \{v_i\} \subseteq \mathcal{H}$ that is symmetric: if $v \in \mathcal{A}$ then $-v \in \mathcal{A}$. For convenience of notation, many of the atomic sets we mention later are not immediately symmetric; in these cases it is sufficient to assume we are instead dealing with $\mathcal{A} \cup -\mathcal{A}$. We note that the convex hull $\mathbf{conv}(\mathcal{A})$ contains 0 and is a polytope when $\mathcal{A}$ is finite. The symmetricity of $\mathcal{A}$ is important for the following property: $\mathbf{conv}(\mathcal{A})$ induces a norm from its gauge function that we call the "atomic norm" induced by $\mathcal{A}$:

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

The dual atomic norm is then defined

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

Familiar examples of atomic sets include the aforementioned coordinate basis vectors $\{e_i\} \subset \mathbb{R}^n$ and rank-one matrices $\left\{uv^\top\right\} \subset \mathbb{R}^{n\times d}$. 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 [1, 2, 3, 4]. However, in this paper we are concerned with greedy algorithms that *construct* sparse atomic solutions. Such an approach is desirable for multiple reasons. Firstly, atomic norm regularization requires solving a convex program at each iteration; for many atomic sets, atomic norm regularization requires semi-definite programming, which is prohibitively costly for any somewhat high-dimensional problem. Secondly, atomic norm regularization only implicitly induces sparsity, and requires fine-tuning of parameters to get the right degree of sparsity. Thirdly, sharp sufficient conditions under which atomic norm regularization will recover the true sparse solution are in general unknown, and the conditions that are known (e.g. Restricted Isometry Property) are often computationally infeasible [5]. Therefore, one may be motivated to consider the greedy approach to sparse atomic optimization, where at each iteration the locally optimal (by some metric) atom is added to the active set, thus giving us explicit control over the sparsity of the resulting solution. However, whereas we are guaranteed to converge to the optimal solution in atomic norm regularization, we must face the possibility of a suboptimal solution. One of the main goals of this paper is therefore to establish that the possibly suboptimal greedy solution is comparable to the optimal sparse solution. The problem we consider in this paper is the following "sparse atomic" maximization:

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

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

Definition 1.1 (Restricted Strong Concavity, Restricted Smoothness [6, 7]) *A function $f : \mathcal{H} \to \mathbb{R}$ is restricted strongly concave with parameter μ_Ω and restricted smooth with parameter L_Ω if for all $x, y \in \Omega \subset \mathcal{H}$,*

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

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

$$L_{\Omega'} \leq L_{\Omega}, \quad \mu_{\Omega'} \geq \mu_{\Omega}. \tag{1.1}$$

In our paper, we often shorten notation and treat Ω as the whole ambient space, such that restricted strong concavity and restricted smoothness just become their unrestricted counterparts. However, strictly speaking, setting r as the sparsity constraint of (P), then it is sufficient to treat Ω as the set of all elements of the ambient space that can be written as a linear combination of no more than $2r$ atoms. As a shorthand, we may write the corresponding strong concavity and smoothness parameters as μ_{2r} and L_{2r}. We note that the additional flexibility of "restricted" strong concavity and smoothness turns out to be crucial in admitting important problems and objective functions into the model [6, 7, 8, 9].

Khanna et al. [10, 11, 9] have shown that greedy algorithms attain multiplicative 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 in some way to general atomic sets. However, at face value, the definition of "atomic set" is extremely broad, and therefore one can easily construct poorly behaving atomic sets where greedy algorithms will not achieve any sort of approximation to the optimal sparse solution in many, many iterations. Therefore, we must introduce a way to measure the structure of an atomic set, in particular its suitability for greedy algorithms. This will come in the form of the "atomic set condition number" that will be introduced in the next section. We will end up showing that the approximation guarantee of the greedy algorithm has a very intuitive dependency on three components:

- conditioning of the objective function;
- conditioning of the underlying atomic set;
- the number of greedy steps.

The precise meaning of the first two items will be formalized later. This echoes the form of earlier approximation guarantees [3]. However, the concrete notion of "atomic condition number" that we introduce has many immediate benefits. Firstly, it generally leads to tighter approximation guarantees than other structural measures of an atomic set. Secondly, the sub-problem of explicitly computing or bounding the atomic condition number is relatively straightforward, as it primarily involves elementary linear algebra, and does not require computing complicated geometric values involving the width or volume of convex bodies. Therefore, if one were to formulate a structured optimization problem in the language of sparse atomic optimization, one could obtain explicit *post hoc* numerical approximation guarantees by deriving the atomic condition numbers. We demonstrate this by deriving a number of atomic condition numbers for common atomic sets in the appendix.

After establishing the importance of atomic condition numbers, our second aim is to revisit and similarly extend to the general atomic setting a notion that has recently been the power tool in establishing greedy approximation guarantees: weak submodularity. One may be familiar with the $\left(1 + e^{-1}\right)$ approximation guarantee of the greedy algorithm on non-negative submodular functions [12]. It has been shown in recent work [9, 13] that certain families of restricted strongly convex, smooth functions can be transformed into "weakly" submodular functions, for which the greedy algorithm attains good approximation guarantees that decay gracefully depending on how "weakly"

submodular the function is. We show that greedy algorithms also attain nice approximation guarantees in the language of weak submodularity that are distinct from the ones derived using atomic condition numbers: weak submodularity is a notion that has an identity distinct from continuous optimization. We give a motivating example that shows weak submodularity has utility outside the realm of sparse atomic optimization. Roughly speaking, whereas greedy approximation guarantees for sparse atomic optimization have a separate dependence on the conditioning of the atomic set and the conditioning of the objective function, greedy approximation guarantees for weak submodular maximization technically depends only on the weakly submodular function. In previous bounds relating sparse optimization and weak submodularity, one could say the stars aligned and allowed the good conditioning of the objective function and atomic set to translate to a useful notion of weak submodularity.

2 Atomic Set Properties and Greedy Improvements

Atomic sets in full generality can be succinctly characterized as any set $\mathcal{A}$, possibly uncountable, in a Hilbert space that is symmetric: if $v \in \mathcal{A}$, then $-v \in \mathcal{A}$. This is the definition used in recent literature concerning the convergence properties of Frank-Wolfe-type algorithms [1, 3, 14]: in the past, the term "atom set" has been used to refer to certain particular examples of the above definition, in particular the finite-dimensional elementary basis vectors $\{e_i\} \subset \mathbb{R}^n$ and elements of a dictionary [15, 16, 17, 18, 19]. Examples of atomic sets include the aforementioned instances in addition to rank-one matrices, Dirac measures, orthogonal matrices, permutation matrices, as well as group-sparse atoms [2, 1].

A unifying goal in introducing the notion of atoms is sparsity. Consider the convex hull $\mathbf{conv}\,(\mathcal{A})$: the maximum of a convex function (in particular linear) over it is attained at an extreme point, i.e. an atom. This is part of the intuition behind Frank-Wolfe-type algorithms, and also underlies the motivation for atomic regularization [2, 1, 3]. In atomic regularization, an objective function of the form $f(x) + \lambda \|x\|_{\mathcal{A}}$, where $\|x\|_{\mathcal{A}}$ is the gauge norm induced by $\mathbf{conv}\,(\mathcal{A})$, is considered. One may be familiar with the analysis of particular examples of atomic regularization, such as LASSO, where $\mathcal{A} = \{e_i\}^n$, $\|\cdot\|_{\mathcal{A}} = \|\cdot\|_1$ [20, 15, 21], and nuclear-norm minimization, where $\mathcal{A} = \left\{uv^\top\right\}$, $\|\cdot\|_{\mathcal{A}} = \|\cdot\|_*$ [22, 23]. If $f(x)$ is convex, then gradient-based methods have provably good convergence properties and enjoy Nesterov-type acceleration [24, 25], since $f(x) + \lambda \|x\|_{\mathcal{A}}$ is a convex composite objective. Regardless, there are two immediate drawbacks to the regularization approach to sparse atomic optimization. Firstly, blackbox convex programming solvers are difficult to scale to many important modern problems. For example, matrix completion (low-rank optimization) turns into semidefinite programming, which struggles in practical efficiency as the dimension of the matrix inflates. Secondly, the sparsity induced by the regularization term is implicit: many of the sharper conditions to guarantee a sparse solution are provably expensive to verify [5, 26]. This serves as the motivation for a greedy algorithm approach to sparse atomic optimization. In particular, we consider the approach of adding one atom greedily to an active set per iteration, such that we have explicit control over the sparsity. Certainly, the greedy approach may not find the optimal sparse solution, but the ultimate aim is to bound the improvement per iteration to guarantee that the greedy algorithm can generate a sparse solution that is a good approximation of the optimal.

Atomic sets are not created equal. For such a broad definition, we cannot hope to find approximation guarantees that are both strong and universally applicable for sparse atomic optimization. To capture the effectiveness of greedy steps on a given atomic set, let us define the following quantities.

Definition 2.1 (Atomic Condition Number) *Let $\mathcal{A} \subseteq \mathcal{H}$ be a given atomic set in inner product space $\mathcal{H}$. We define the atomic condition number of $\mathcal{A}$ to be the largest value such that for any vector $v \in \mathcal{H}$, $v \neq 0$, there exists an atom $a \in \mathcal{A}$ such that*

$$\frac{|\langle v, a \rangle|}{\|v\| \, \|a\|} \geq \theta.$$

In other words,

$$\theta := \min_{\|v\|=1} \max_{a \in \mathcal{A}} \frac{|\langle v, a \rangle|}{\|a\|}.$$

Definition 2.2 (Sparse Atomic Condition Number) *Let $\mathcal{A} \subseteq \mathcal{H}$ be a given atomic set in inner product space $\mathcal{H}$. Given $L \subset \mathcal{A}$, let the atomic condition number with respect to L, $\theta(L) \geq 0$, to be*

$$\theta(L) := \min_{\substack{\|v\|=1 \\ v \in \boldsymbol{span}(L)}} \max_{a \in \mathcal{A}} \frac{|\langle v, a \rangle|}{\|a\|}.$$

The r-sparse, or simply sparse, atomic condition number is defined as the minimum atomic condition number over all r-subsets: $\theta_r := \min_{|L| \leq r} \theta(L)$. In other words,

$$\theta_r := \min_{\substack{L \subseteq \mathcal{A} \\ |L| \leq r}} \min_{\substack{\|v\|=1 \\ v \in \boldsymbol{span}(L)}} \max_{a \in \mathcal{A}} \frac{|\langle v, a \rangle|}{\|a\|},$$

where $|L|$ indicates the cardinality of L.

Essentially, the atomic condition number measures how dense an atomic set is. A large θ means every vector is reasonably close to an atom. We can identify some representative atomic sets with properties that lead to tight lower bounds on θ_r. We will see that larger values of θ_r directly correspond to stronger bounds on greedy improvement.

- (Topologically) dense atomic sets, e.g. Euclidean unit sphere. In that case, $\theta = 1$ trivially, since any vector is well-approximated by an atom. That said, sparsity with respect to these atomic sets is not often very meaningful.
- Orthogonal basis. $\theta_r \geq r^{-1/2}$. This class of atomic sets is perhaps the nicest "meaningful" atomic set, particularly due to the following property: given two sets of atoms S and T, where $|T| = k$, if we define T' as the minimal set of atoms such that

  $$\mathbf{span}\left(T'\right) \supseteq \mathrm{proj}_{S^\perp}(T),$$

  we observe $|T'| \leq k$, as T' is simply T minus the atoms it shares with S. This is a key property used in Elenberg et al. [9] to prove strong approximation guarantees for greedy algorithms where $\mathcal{A}$ is the set of elementary basis vectors (sparse optimization). This is a property lost even when considering union of orthogonal bases (dictionary learning).
- Other atomic sets with θ_r dependent only on r, for example rank-one matrices: $\theta_r \geq r^{-1/2}$ as the linear combination of r rank-one matrices is at most rank r. We observe that rank-one matrices do not satisfy the additional property mentioned above.

Atomic set	**Atomic cond. number θ**	**Sparse θ_r value**
Standard basis $\{e_i\}_{i=1}^n \subset \mathbb{R}^n$	$n^{-1/2}$	$r^{-1/2}$
$m \times n$ rank-one matrices $uv^\top$	$(\min\{m,n\})^{-1/2}$	$r^{-1/2}$
Disjoint group-sparse atoms $\mathcal{P}(\{e_i\}_{i=1}^n)$	$(\#\text{ groups})^{-1/2}$	$r^{-1/2}$
2-ortho basis $\{\phi_i\}^n \cup \{\psi_j\}^n \subset \mathbb{R}^n$	$\Omega\left(n^{-1/2}\right)$	$\Omega\left(r^{-1/2}\right)$ when $r \leq \mu(\mathcal{A})^{-1}$
Binary sign vectors $\{\pm 1\}^n$	$n^{-1/2}$	$n^{-1/2}$
$n \times n$ orthogonal matrices	$n^{-1/2}$	$n^{-1/2}$

Table 1: Sparse Atomic Condition Numbers

- Atomic sets with θ_r bounded away from 0, but possibly dependent on the ambient dimension. For example, the set of orthogonal matrices $U \in M_n$, where $\theta_r = n^{-1/2}$. The proof of this bound will be relegated to the appendix.

Table 1 contains a larger sample of atomic sets and their respective lower bounds on the atomic condition numbers. The proofs of the bounds in Table 1 can be found in the appendix.

We note that even though adding more vectors to an atomic set can only increase θ, and $\theta_r \geq \theta$, it is not necessarily the case that θ_r monotonically increases with the addition of vectors to the atomic set. In fact, the gap between θ_r and θ, if it exists, can be made arbitrarily small by adding just one vector to the atomic set. Consider the following simple example. Let $\mathcal{A} = \{e_i\}_{i=1}^n$. Consider adding to $\mathcal{A}$ the vector $v = e_1 + \varepsilon\mathbb{1}$, where $\mathbb{1}$ is the all-ones vector, and ε is an arbitrary small value. Then for any $r \geq 2$, we consider the subset containing e_1 and v. The span of that subset will contain $n^{-1}\mathbb{1}$, which satisfies

$$\max_{a \in \mathcal{A} \cup \{v\}} \frac{|\langle \mathbb{1}, a\rangle|}{\sqrt{n}\,\|a\|} \leq \frac{1}{\sqrt{n}} + \delta,$$

where δ can be made arbitrarily small by shrinking ε. Therefore, we have shown that we can make the sparse atomic condition number θ_r for the standard basis, which is normally $r^{-1/2}$, arbitrarily close to its atomic condition number $\theta = n^{-1/2}$ simply by corrupting the atomic set with one vector. We will later see that this phenomenon is closely tied to how well a greedy approach can find good sparse solutions. In short, the structure of an atomic set is important!

2.1 Atomic Sets and Greedy Optimization

The atomic condition numbers have a direct relationship with bounds on greedy algorithmic performance. First, we state the following combinatorial reformulation of problem (P) on which we apply greedy algorithms:

$$\begin{aligned} (P) \quad & \max\ g(U) \\ & \text{s.t. } U \subset \mathcal{A},\ |U| \leq r. \end{aligned}$$

In the cases of our concern, $g : 2^{\mathcal{A}} \to \mathbb{R}$ is a set function defined on subsets of the atomic set:

$$g(U) := \max_{x \in \mathbf{span}(U)} f(x) - f(0),$$

where $f(x)$ is some (restricted) strongly concave and smooth function that we want to maximize. Defining $g(\emptyset) = 0$, we observe that g is a non-negative function. Let $g(U^*) = f(x^*) - f(0)$ be the optimal value of (P). Similarly, given set U, we define

$$B^{(U)} := \underset{x \in \mathbf{span}(U)}{\arg\max} f(x) - f(0).$$

In other words, $B^{(U)}$ is the vector in $\mathcal{H}$ such that $f(B^{(U)}) = g(U) + f(0)$. We have the following standard lemma relating the norm of the gradient at a given point to the objective value.

Lemma 2.3 *Let x^* be the optimal solution to (P). Then for any x, we have*

$$\frac{\|\nabla f(x)\|^2}{2L} \leq f(x^*) - f(x) \leq \frac{\|\nabla f(x)\|^2}{2\mu}.$$

Proof: by the concavity of f, we have

$$\begin{aligned} f(x^*) &\leq f(x) + \nabla f(x)^\top (x^* - x) - \frac{\mu}{2} \|x^* - x\|^2 \\ &\leq f(x) + \|\nabla f(x)\| \|x^* - x\| - \frac{\mu}{2} \|x^* - x\|^2 \\ &\leq f(x) + \frac{\|\nabla f(x)\|^2}{2\mu}. \end{aligned}$$

On the other hand, from the smoothness of f, we have

$$\begin{aligned} f(x^*) &\geq f\left(x + \frac{1}{L}\nabla f(x)\right) \\ &\geq f(x) + \frac{1}{L}\nabla f(x)^\top \nabla f(x) - \frac{1}{2L} \|\nabla f(x)\|^2 \\ &= f(x) + \frac{\|\nabla f(x)\|^2}{2L}. \end{aligned}$$

■

Therefore, given any subset U, if we pick an atom v satisfying

$$\frac{\langle \nabla f(x_U), v \rangle}{\|\nabla f(x_U)\| \|v\|} \geq \theta,$$

where such a v is guaranteed to exist by the definition of θ, we can lower bound the gain from $g(U)$ to $g(U \cup \{v\})$. From now on, assume that the elements of the atomic set $\mathcal{A}$ are normalized: $\|v\| = 1$ for all $v \in \mathcal{A}$.

$$\begin{aligned} g(U \cup \{v\}) &\geq \max_t f(x_U^* + tv) - f(0) \\ &\geq f(x_U^*) + \max_t \left[t \nabla f(x_U^*)^\top v - t^2 \frac{L}{2} \|v\|^2 \right] - f(0) \\ &\geq g(U) + \theta^2 \frac{\|\nabla f(x_U^*)\|^2}{2L}, \\ g(U \cup \{v\}) - g(U) &\geq \theta^2 \frac{\|\nabla f(x_U^*)\|^2}{2L} \\ &\geq \theta^2 \frac{\mu}{L} (g(U^*) - g(U)) \end{aligned}$$

by Lemma 2.3. In other words, greedily adding an atom yields an objective gain toward the optimal that can be lower bounded multiplicatively. We now introduce the simple greedy algorithm.

Algorithm 1 Greedy$(\mathcal{A}, f, r, \beta)$

$U_0 \leftarrow \emptyset$
for $t = 1, \ldots, r$ **do**
    Compute $B^{(U_{t-1})}$, $\nabla f\left(B^{(U_{t-1})}\right)$
    $v_t \leftarrow \texttt{PureGreedy}(\mathcal{A}, f, U_{t-1}, \beta)$
    $(v_t \leftarrow \texttt{OMPSel}(\mathcal{A}, f, U_{t-1}, \beta))$
    $U_t \leftarrow U_{t-1} \cup \{v_t\}$
end for
return $U_r, B^{(U_r)}, f(U_r)$

$\texttt{PureGreedy}(\mathcal{A}, f, U_{t-1}, \beta)$ is an oracle that returns an atom v_t such that

$$g\left(U_{t-1} \cup \{v_t\}\right) - g\left(U_{t-1}\right) \geq \beta \max_{v \in \mathcal{A}} \left(g\left(U_{t-1} \cup \{v\}\right) - g\left(U_{t-1}\right)\right).$$

On the other hand, $\texttt{OMPSel}(\mathcal{A}, f, U_{t-1}, \beta)$ is a linear oracle that returns an atom v_t such that

$$\left|\left\langle \nabla f\left(B^{(U_{t-1})}\right), v_t \right\rangle\right| \geq \beta \max_{v \in \mathcal{A}} \left|\left\langle \nabla f\left(B^{(U_{t-1})}\right), v \right\rangle\right|.$$

We note that either oracle can be used without affecting any our approximation guarantees. However, $\texttt{OMPSel}$ is usually the more computationally feasible option.

Applying the greedy algorithm to (P), from our lower bound on the gain of greedy addition, the greedy algorithm attains the following multiplicative approximation of the optimal solution.

$$\begin{aligned}
g(U^*) - g(U_t) &\leq \left(1 - \beta\theta^2 \frac{\mu}{L}\right)\left(g(U^*) - g(U_{t-1})\right) \\
&\leq \left(1 - \beta\theta^2 \frac{\mu}{L}\right)^t \left(g(U^*) - g(\emptyset)\right) \\
&\leq \left(1 - \beta\theta^2 \frac{\mu}{L}\right)^t g(U^*) \\
&\leq \exp\left(-\beta\theta^2 \frac{\mu}{L}\right) g(U^*) \\
\implies g(U_t) &\geq \left(1 - \exp\left(-\beta\theta^2 t \frac{\mu}{L}\right)\right) g(U^*).
\end{aligned}$$

Up to this point, analogous convergence rates have been shown for other greedy-type methods [3, 4]: substituting the $\theta = \frac{1}{\sqrt{n}}$ value for sparse optimization, we get an approximation guarantee of the form

$$g(L_t) \geq \left(1 - \exp\left(-\beta \frac{t}{n} \frac{\mu}{L}\right)\right) g(U^*).$$

According to this approximation guarantee, if we have run the greedy algorithm for $O(n)$ iterations, we get an approximation ratio solely dependent on the condition number μ/L and the precision constant β. However, this is not the end-goal of sparse optimization, as the solution will have $O(n)$ non-zero entries.

2.2 Tightening Greedy Bound

As previewed earlier, our goal is to create a bound on the greedy performance that attains a "constant-factor" approximation ratio (that is, solely dependent on the condition number and the precision constant) *while* maintaining sparsity of the solution. Here we will show that we can replace θ in our earlier bounds with θ_{2r}. Assume we have applied the greedy algorithm on (P) for r iterations and attained atom set U_r, and that the optimal solution to (P) is U^*. Define $V = U_r \cup U^*$. We consider a restricted version of (P):

$$\begin{aligned}(P_R) \quad & \max \tilde{f}(x) \\ & \text{s.t. } x \in \mathbf{span}\,(U) \\ & \quad U \subseteq V,\ |U| \leq r,\end{aligned}$$

where $\tilde{f}(x) := f\left(\text{proj}_V(x)\right)$, where f is the restricted strongly concave, smooth objective function in (P). $\tilde{f}(x)$ is concave, as the projection operator is linear. Additionally, $\tilde{f}(x)$ also inherits the strong concavity and smoothness of f, as long as x is restricted to $\mathbf{span}\,(V)$. In other words, if f is m_Ω-restricted strongly concave and M_Ω-smooth on Ω, then $\tilde{f}(x)$ is m_Ω-restricted strongly concave and M_Ω-smooth on $\Omega \cap \mathbf{span}\,(V)$. Let us introduce the sparse condition number of the objective function f:

$$\sigma_r := \min_{\substack{L \subseteq \mathcal{A} \\ |L| \leq r}} \min_{u \in \mathbf{span}(L)} \frac{\mu(u)}{L(u)}.$$

In other words, σ_r is the condition number of the function f over all subspaces of dimension at most r. Observe that $\sigma_r \geq \sigma := \mu/L$. Recalling our notation μ_r and L_r, indicating the restricted strong convexity and smoothness constants over all subspaces of dimension at most r, we also have $\sigma_r \geq \mu_r/L_r$. We note that the latter expression may be more practical to estimate.

Observe that the optimal values of (P) and (P_R) are the same. However, since the dimension of the search space of (P_R) is at most $2r$, we can replace the θ in previous derivations with θ_{2r}, and μ/L with σ_r. We note that θ_{2r} may not always have a polynomial dependence on r; in some cases, θ_{2r} might be no better than θ, as one can see from the table of atomic condition numbers. The greedy algorithm applied on (P_R), Greedy$(V, \tilde{f}, r, \beta)$, therefore has a convergence rate of

$$\tilde{g}(U_t) \geq \left(1 - \exp\left(-\beta\theta_{2r}^2 t \sigma_{2r}\right)\right) \tilde{g}(U^*).$$

Since $\tilde{g}(U^*) = g(U^*)$, if we show that the iterates of Greedy$(V, \tilde{f}, r, \beta)$ are identical with the iterates of the greedy algorithm applied to (P), Greedy$(\mathcal{A}, f, r, \beta)$, then the above convergence rate is actually the convergence of the greedy algorithm on (P). If `PureGreedy` is used, this is trivial, since $\tilde{f}$ is a restriction of f, and all the locally optimal choices for the greedy algorithm on (P) are available in the search space of (P_R). If `OMPSel` is used instead, the iterates are still identical by applying the chain rule: let P_V denote the projection matrix projecting onto $\mathbf{span}\,(V) = \mathbf{span}\,(U_r \cup U^*)$ such that $\tilde{f}(x) = f\left(P_V x\right)$. By the chain rule we have

$$\nabla \tilde{f}(x) = P_V \nabla f\left(P_V x\right)$$

Each v_t chosen by the greedy algorithm on (P) satisfies

$$\left\langle \nabla f\left(B^{(U_{t-1})}\right), v_t \right\rangle = \max_{v \in \mathcal{A}} \left\langle \nabla f\left(B^{(U_{t-1})}\right), v \right\rangle.$$

By definition of U_r, we have $v_t \in U_r$ for all $t = 1, \ldots, r$. Since $B^{(U_{t-1})} \in \mathbf{span}(V)$, v_t also satisfies

$$\begin{aligned}\left\langle \nabla f\left(B^{(U_{t-1})}\right), v_t \right\rangle &= \max_{v \in \mathcal{A}} \left\langle \nabla f\left(B^{(U_{t-1})}\right), P_V v \right\rangle \\ &= \max_{v \in \mathcal{A}} \left\langle P_V \nabla f\left(P_V B^{(U_{t-1})}\right), v \right\rangle \\ &= \max_{v \in \mathcal{A}} \left\langle \nabla \tilde{f}\left(B^{(U_{t-1})}\right), v \right\rangle.\end{aligned}$$

Therefore, the locally optimal atom at each iteration on (P) as decided by `OMPSel` agrees with the locally optimal atom on (P_R). By induction, this implies that the iterates of the greedy algorithm on (P) agree with the iterates of the greedy algorithm on (P_R). Therefore, we have established the improved approximation guarantee.

Theorem 2.4 *Greedy$(\mathcal{A}, f, k, \beta)$ has the following multiplicative improvement ratio and approximation guarantee:*

$$\begin{aligned}g(U^*) - g(L_t) &\leq \left(1 - \beta\theta_{2r}^2 \sigma_{2r}\right)\left(g(U^*) - g(L_{t-1})\right) \\ g(L_t) &\geq \left(1 - \exp\left(-\beta\theta_{2r}^2 t\sigma_{2r}\right)\right) g(U^*).\end{aligned}$$

Referring to our earlier discussion of θ_r values for particular atomic sets, we have the following examples of improved greedy approximation guarantees.

Corollary 2.5 (Greedy Feature Selection Convergence Rate) *Consider problem (P), where $\mathcal{A} = \{e_i\}_{i=1}^n \subset \mathbb{R}^n$. Given a function f (and corresponding function g) that satisfies RSC-RS, we have the following lower bound for the performance of the greedy algorithm*

$$g(U_t) \geq \left(1 - \exp\left(-\beta\theta_{2r}^2 t\sigma_{2r}\right)\right) g(U^*) \geq \left(1 - \exp\left(-\beta\frac{t}{2r}\frac{\mu_{2r}}{L_{2r}}\right)\right) g(U^*)$$

Corollary 2.6 (Greedy Low-Rank Optimization Convergence Rate) *Consider problem (P), where $\mathcal{A} = \left\{uv^\top\right\} \subset \mathbb{R}^{m \times n}$. Given a function f (and corresponding function g) that satisfies RSC-RS, we have the following lower bound for the performance of the greedy algorithm*

$$g(U_t) \geq \left(1 - \exp\left(-\beta\theta_{2r}^2 t\sigma_{2r}\right)\right) g(U^*) \geq \left(1 - \exp\left(-\beta\frac{t}{2r}\frac{\mu_{2r}}{L_{2r}}\right)\right) g(U^*)$$

Note that the above approximation guarantees imply that given the sparsity constraint r, the greedy algorithm will find a constant-factor approximation of the optimal r-sparse solution within $O(r)$ iterations, instead of $O(n)$ iterations. These agree with the current best approximation guarantees (up to small constant factors) for greedy-type algorithms for the above settings [9, 10, 11, 14, 27].

2.3 Forward-Backward Schemes

Our goal is to extend an analogous approximation guarantee to a flexible family of forward-backward algorithms. The motivation of forward-backward algorithms is that "bad atoms" contributing little to the objective chosen earlier by the myopic forward steps may be removed later by backward

steps to improve the quality of the sparse solution. At its core, the forward-backward paradigm is heuristic, and thus bounds on its performance even in familiar settings and for popular objectives are few and far between, despite bounds existing on the forward-only procedure. We will show that a large class of forward-backward schemes have approximation guarantees no worse than the corresponding forward-only scheme. We propose the following framework for the Forward-Backward scheme.

Algorithm 2 FoBa($\mathcal{A}, f, k, \beta, \nu$)

// ν is a thresholding constant $0 \leq \nu < 1$
$S_0 \leftarrow \emptyset$
$t \leftarrow 0$
while $|S_t| < k$ **do**
    $v_{t+1} \leftarrow \texttt{PureGreedy}(\mathcal{A}, f, S_t, \beta)$
    $(v_{t+1} \leftarrow \texttt{OMPSel}(\mathcal{A}, f, S_t, \beta))$
    $S_{t+1} \leftarrow S_t \cup \{v_{t+1}\}$
    $t \leftarrow t + 1$
    // Optional: additional conditions to enter backward step
    $\texttt{Dmg} \leftarrow 0$ // variable tracking the damage done by backward steps
    $d^+ \leftarrow g(S_t) - g(S_{t-1})$
    while true do
        $p \leftarrow \arg\max_{v \in \mathcal{A}} g(S_t \setminus \{p\})$ // picking the element contributing the least
        $d^- \leftarrow g(S_t) - g(S_t \setminus \{p\})$
        if $(\texttt{Dmg} + d^-) < \nu d^+$ **then**
            $S_{t+1} \leftarrow S_t \setminus \{p\}$
            $t \leftarrow t + 1$
            $\texttt{Dmg} \leftarrow \texttt{Dmg} + d^-$
        else
            break
        end if
    end while
end while
return $S_t, g(S_t)$

Theorem 2.7 (FoBa Convergence Rate) *Let x^*_S denote the optimal solution to (P), with sparsity constraint k. If for some $0 < c \leq 1$ the forward-only procedure Greedy$(\mathcal{A}, f, k, \beta)$ satisfies a convergence rate of the form*

$$g(S^*) - g(S_{t+1}) \geq (1 - c)\,(g(S^*) - g(S_t)),$$

then FoBa$(\mathcal{A}, f, k, \beta, \nu)$ has a convergence rate

$$g(S_t) \geq \left(1 - \left(1 - \frac{c}{k}\right)^{|S_t|}\right) g(S^*) \geq (1 - \exp(-c\,|S_t|))\, g(S^*)$$

Note that this bound is independent of the thresholding constant $0 \leq \nu < 1$.

Proof of Theorem 2.7: We use induction. First we verify the base cases: $t = 0 \implies S_0 = \emptyset$. By our algorithm, $t = 1$ must be a forward step, and therefore the bound is true due to our assumption of

the forward-only convergence rate. Assume the induction hypothesis: at steps $j < t+1$ we have

$$g(S_j) \geq \left(1 - \left(1 - \frac{c}{k}\right)^{|S_j|}\right) g(S^*).$$

After step t, we are at one of the following two scenarios.

- Case 1: step $t+1$ will be a forward step. Therefore, $|S_{t+1}| = |S_t| + 1$. By the same argument made in the proof of Theorem 3.4, we have

$$g(S_{t+1}) \geq \left(1 - \left(1 - \frac{c}{k}\right)^{|S_{t+1}|}\right) g(S^*)$$

- Case 2: step $t+1$ will be a backward step. Since we can only take a backward step after making at least one forward step, say our last forward step was at step $t-i$, $i \geq 0$. By the thresholding in the backward step, we have that

$$g\left(S_{t+1}\right) \geq g\left(S_{t-i-1}\right),$$

and since all steps since $t-i$ are backward steps, we have

$$|S_{t+1}| = |S_{t-i-1}| - i \leq |S_{t-i-1}|.$$

By the induction hypothesis we have

$$\begin{aligned} g\left(S_{t+1}\right) &\geq g\left(S_{t-i-1}\right) \\ &\geq \left(1 - \left(1 - \frac{c}{k}\right)^{|S_{t-i-1}|}\right) g(S^*) \\ &\geq \left(1 - \left(1 - \frac{c}{k}\right)^{|S_{t+1}|}\right) g(S^*). \end{aligned}$$

Therefore, when the algorithm terminates at step N, we have $|S_N| = k$ and thus

$$g\left(S_N\right) \geq \left(1 - \left(1 - \frac{c}{k}\right)^{k}\right) g(S^*) \geq \left(1 - \exp(-c)\right) g(S^*).$$

■

Substituting $c = \beta \frac{\mu}{L} \theta_{2k}^2$, we recover the exact same approximation guarantee for the forward-backward scheme as the purely greedy scheme. We cannot hope for a better guarantee in general, as it is possible for no backward steps to have been taken. To some degree, it is also not surprising that the forward-backward scheme is "as good" as the purely greedy scheme, but we note that the qualifications and conditions to enter the backward phase and/or to take a backward step can be modified in numerous ways to get better empirical results and will likely still result in similar approximation guarantees; we are proposing but one popular class of forward-backward schemes [28].

3 Weak Submodularity

Recently, strong multiplicative bounds for greedy performance on particular atomic sets (i.e. $\mathcal{A} = \{e_i\}^n$, $\mathcal{A} = \{uv^\top\}$) were established using the notion of weak submodularity [9, 10, 11]. In the previous section, we have recovered and extended these bounds to the general atomic setting independent of weak submodularity. However, we note that weak submodularity in the aforementioned papers served predominantly as a convenient combinatorial interpretation of a continuous convex problem. Namely, a notion known as "submodularity ratio" [13] is developed, and it is essentially shown that a μ-strongly convex, L-smooth function can be turned into a μ/L-weakly submodular set function, for which greedy maximization attains a $(1 - \exp(-\mu/L))$-approximation guarantee, reminiscent of the $(1 - 1/e)$ guarantee for greedy maximization of submodular functions in the seminal paper by Nemhauser et al. [12]. We note that the ground sets of the weakly-submodular functions in the aforementioned literature are limited to the standard basis [10, 11, 13] or the rank-one matrices [9]. We can generalize bounds on weak submodularity to general atomic sets that satisfy a certain conditions.

We note that there are rich classes of set functions that cannot be fruitfully converted into the language of convexity, but weak submodularity may be ascribed. In this section, we develop a new notion of weak submodularity which we will prove is interchangeable with versions introduced in earlier papers. Weak submodularity allows us to establish strong multiplicative bounds for the greedy algorithm, and hence the forward-backward algorithm. We define the *subset* submodularity ratio, which is related to what will be called the *disjoint* submodularity ratio seen in earlier literature [13, 9, 10].

Definition 3.1 (Disjoint Submodularity Ratio [13]) *Let $P, Q \subset [p]$ be two disjoint index sets, and $g : [p] \to \mathbb{R}$. The* disjoint *submodularity ratio of P with respect to Q is defined as*

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

Intuitively, the disjoint submodularity ratio measures the diminishing marginal returns property. The additional adjective "disjoint", which is not seen in earlier literature [13, 9], is introduced here to distinguish it from a separate notion of submodularity ratio we define next. Note that $\gamma_{P,Q} \geq 1$ if f is submodular. We can also define the disjoint submodularity ratio of a set $U \subseteq [p]$ with respect to an integer $k > 0$:

$$\gamma_{U,k} := \min_{\substack{P,Q : P \cap Q = \emptyset; \\ P \subseteq U \\ |Q| \leq k}} \gamma_{P,Q}.$$

Elenberg et al. [9] recently demonstrated in the standard basis setting that functions for satisfying RSC-RS are weakly submodular, where the disjoint submodularity ratio is essentially lower bounded by the restricted condition number: $\gamma_{U,k} \geq \mu_{|U|+k}/L_{|U|+k}$. This leads to a greedy convergence rate of the form

$$g(P_t) \geq \left(1 - \exp\left(-\frac{\mu_{|U|+k}}{L_{|U|+k}}\right)\right) g(P^*),$$

which is precisely analogous to the bound we established for the general atomic setting in the previous section. The reader is directed to the paper by Elenberg et al. [9] or the paper by Das and Kempe [13] for more details. A persistent barrier in simply extending their results to connect

general sparse atomic optimization and weak submodularity is summarized by the following: not every atomic set contains an orthogonal basis, let alone an orthogonal basis that can be formed starting with any arbitrary element in the atomic set.

Let us introduce a new notion of weak submodularity below, which directly relies on the sense of approximately diminishing marginal returns for a set function.

Definition 3.2 (Subset Submodularity Ratio) *Let $U, V \subset [p]$ be index sets such that $U \subseteq V$, and $g : [p] \to \mathbb{R}$. The* subset *submodularity ratio of U with respect to V is defined as*

$$\kappa_{U,V} := \min_{i \in [p] \setminus V} \frac{g\left(U \cup \{i\}\right) - g\left(U\right)}{g\left(V \cup \{i\}\right) - g\left(V\right)}$$

Similarly, we may define the subset submodularity ratio of a set $U \subseteq [p]$ with respect to an integer $k > 0$:

$$\kappa_{U,k} := \min_{\substack{T,V : T \subseteq U, U \subseteq V \\ |V \setminus U| \leq k}} \kappa_{T,V}.$$

Now that we have stated results for the disjoint submodularity ratio, it is interesting to note that the disjoint and subset submodularity ratios are tightly related.

Theorem 3.3 *Given $\gamma = \gamma_{U,k}$ and $\kappa = \kappa_{U,k}$, we have the following relationship:*

$$\begin{aligned} &\frac{\gamma}{2-\gamma} \leq \kappa \leq \gamma, \\ \implies &0.5\gamma \leq \kappa \leq \gamma. \end{aligned}$$

Proof of Theorem 3.3: we first prove that $\kappa \leq \gamma$. Consider any $L \subseteq U, |S| \leq k$, $L \cap S = \emptyset$. Let $S = \left\{x_1, \ldots, x_{|S|}\right\}$. We have

$$\begin{aligned} g(L \cup S) - g(L) &= \sum_{j=1}^{|S|} \left[g\left(L \cup \{x_1, \ldots, x_j\}\right) - g\left(L \cup \{x_1, \ldots, x_{j-1}\}\right)\right] \\ &\leq \sum_{j=1}^{|S|} \frac{1}{\kappa} \left[g\left(L \cup \{x_j\}\right) - g(L)\right], \\ \implies \kappa &\leq \frac{\sum_{j=1}^{|S|} \left[g\left(L \cup \{x_j\}\right) - g(L)\right]}{g(L \cup S) - g(L)}. \end{aligned}$$

Taking the minimum of the right-hand side of the last equation, we get $\kappa \leq \gamma$ as desired.

We now prove that $\frac{1}{2}\gamma \leq \frac{\gamma}{2-\gamma} \leq \kappa$. Consider any set $\{j, k\}$. We have

$$\begin{aligned} g(L \cup \{j\}) - g(L) &\geq \kappa \left(g(L \cup \{j,k\}) - g(L \cup \{k\})\right) \\ &= \kappa \left(g(L \cup \{j,k\}) - g(L)\right) - \kappa \left(g(L \cup \{k\}) - g(L)\right) \qquad (3.1) \\ g(L \cup \{k\}) - g(L) &\geq \kappa \left(g(L \cup \{j,k\}) - g(L)\right) - \kappa \left(g(L \cup \{j\}) - g(L)\right) \qquad (3.2) \end{aligned}$$

Define

$$\gamma_1 = \frac{g(L \cup \{j\}) - g(L)}{g(L \cup \{j,k\}) - g(L)}$$
$$\gamma_2 = \frac{g(L \cup \{k\}) - g(L)}{g(L \cup \{j,k\}) - g(L)}.$$

Rearranging inequalities 3.1 and 3.2, we have

$$\gamma_1 + \kappa\gamma_2 \geq \kappa$$
$$\gamma_2 + \kappa\gamma_1 \geq \kappa.$$

Furthermore, we have

$$\gamma_1 + \gamma_2 = \frac{(g(L \cup \{j\}) - g(L)) + (g(L \cup \{k\}) - g(L))}{g(L \cup \{j,k\}) - g(L)} \geq \gamma.$$

We consider the following simple minimization problem

$$\begin{aligned} \min\ & \gamma_1 + \gamma_2 \\ \text{s.t. } & \gamma_1 + \kappa\gamma_2 \geq \kappa \\ & \gamma_2 + \kappa\gamma_1 \geq \kappa \\ & \gamma_1, \gamma_2 \geq 0. \end{aligned}$$

The optimal value of the above problem is $\frac{2\kappa}{1+\kappa}$, which implies

$$\gamma \leq \frac{2\kappa}{1+\kappa} \iff \kappa \geq \frac{\gamma}{2-\gamma}.$$

One can verify that $0.5\gamma \leq \frac{\gamma}{2-\gamma}$, which completes the proof of the theorem. ■

As we have previously noted, even though there is some intersection between the world of convex sparse atomic optimization and weak submodular maximization, weak submodularity is independently an important notion to ensure the approximation ratios for greedy algorithms. These two alternative conditions lead to similar worst case performance guarantee, however the settings under which they operate may be completely different. To illustrate this point, let us examine below a weakly submodular function which is by no means related to convexity.

Let $h : \mathbb{R}^n \to \mathbb{R}$ be a componentwise increasing function, and moreover its partial derivatives satisfy $h_i(z) \geq h_i(z')$ for any $z \leq z'$ in the domain, where $i = 1, 2, ..., n$.

Such function h clearly exists. For instance, consider a symmetric doubly stochastic matrix Q, and define $q(y) = -\frac{1}{2}y^\top Qy + \mathbb{1}^\top y$ with $\nabla q(y) = -Qy + \mathbb{1}$ where $\mathbb{1}$ is the all-ones vector. Let $s(x) = 1/(1 + \exp(-x))$ be the sigmoid function, with $s'(x) = s(x)(1 - s(x))$. Finally, let

$$h(z) = q(s(z_1), s(z_2), ..., s(z_n))$$

where $z_i \in [0, \infty)$ for $i = 1, 2, ..., n$. Then, for any $0 \leq z_i \leq z_i'$ $(i = 1, 2, ..., n)$ we have

$$\begin{aligned} \frac{\partial h}{\partial z_i}(z) &= \frac{\partial q}{\partial y_i}(s(z))s(z_i)(1 - s(z_i)) \\ &\geq \frac{\partial q}{\partial y_i}(s(z'))s(z_i')(1 - s(z_i')) = \frac{\partial h}{\partial z_i}(z'), \end{aligned}$$

where $i = 1, 2, ..., n$. The above inequality holds because if $0 \leq z \leq z'$ then $0 < s(z) \leq s(z')$, and so $\nabla q(s(z)) - \nabla q(s(z')) \geq 0$; moreover, $s(x)(1 - s(x))$ is monotonically decreasing when $x \geq 0$.

We now continue our construction after finding such a function h. Let $u_i(t)$ be a unimodal function which attains its maximum at p_i (that is, $u_i(t)$ is increasing for $t < p_i$ and increasing for $t > p_i$). Let us also assume $u_i(0) = 0$.

Let $f(x) := h(u_1(x_1), u_1(x_2)..., u_n(x_n))$. Now we shall show that $g(U) = \max_{x_i, i \in U} f(x) - f(0)$ satisfies $(1/c)$-submodularity in the subset sense.

Consider any $U \subset V \subseteq \{1, 2, ..., n\}$ and $i \notin V$. Without losing generality, let us denote

$$U = \{1, 2, ..., m\},\ V = \{1, 2, ..., m, m+1, ..., m+\ell\},\ \text{and}\ i = m + \ell + 1.$$

Clearly,

$$\begin{aligned}
& g(U \cup \{i\}) - g(U) \\
= \ & h(u_1(p_1), ..., u_m(p_m), 0, ..., 0, u_{m+\ell+1}(p_{m+\ell+1}), 0, ..., 0) - h(u_1(p_1), ..., u_m(p_m), 0, ..., 0) \\
= \ & \int_0^{u_{m+\ell+1}(p_{m+\ell+1})} \frac{\partial h}{\partial x_{m+\ell+1}}(u_1(p_1), ..., u_m(p_m), 0, ..., 0, t, 0, ..., 0)dt.
\end{aligned}$$

Similarly,

$$\begin{aligned}
& g(V \cup \{i\}) - g(V) \\
= \ & h(u_1(p_1), ..., u_m(p_m), u_{m+1}(p_{m+1}), ..., u_{m+\ell}(p_{m+\ell}), u_{m+\ell+1}(p_{m+\ell+1}), 0, ..., 0) \\
& -h(u_1(p_1), ..., u_m(p_m), u_{m+1}(p_{m+1}), ..., u_{m+\ell}(p_{m+\ell}), 0, ..., 0) \\
= \ & \int_0^{u_{m+\ell+1}(p_{m+\ell+1})} \frac{\partial h}{\partial x_{m+\ell+1}}(u_1(p_1), ..., u_m(p_m), u_{m+1}(p_{m+1}), ..., u_{m+\ell}(p_{m+\ell}), t, 0, ..., 0)dt \\
\leq \ & \int_0^{u_{m+\ell+1}(p_{m+\ell+1})} \frac{\partial h}{\partial x_{m+\ell+1}}(u_1(p_1), ..., u_m(p_m), 0, ..., 0, t, 0, ..., 0)dt \\
= \ & g(U \cup \{i\}) - g(U).
\end{aligned}$$

Therefore, the subset submodularity holds with $\kappa = 1$. One notices that there is no concavity, restricted or not, required at all in this example.

We now derive the performance of the greedy algorithm with respect to the subset submodularity ratio. Consider the following maximization problem:

$$\begin{aligned}
\max \quad & g(S) \\
\text{s.t.} \quad & |S| \leq k, \quad S \subseteq [n],
\end{aligned}$$

where we make some assumptions on g.

1. Monotonicity: if $U \subseteq V \subseteq [n]$, then $g(U) \leq g(V)$.
2. Scaled: $g(\emptyset) = 0$.

Algorithm 3 Greedy(g, r)

$S_0 \leftarrow \emptyset$
for $t = 0, \ldots, r-1$ **do**
  $s_{t+1} \leftarrow \arg\max_{i \notin S_t} g\left(S_t \cup \{i\}\right)$
  $S_{t+1} \leftarrow S_t \cup \{s_{t+1}\}$
  $t \leftarrow t + 1$
end for
return $S_r, g(S_r)$

Theorem 3.4 *Let $S^* = \arg\max_{|S|=r} g(S)$. At iteration i of Greedy(g, r), we have*

$$g\left(S_i\right) \geq \left(1 - \left(1 - \frac{\kappa}{r}\right)^i\right) g(S^*),$$

where κ is the subset submodularity ratio. In particular, after the greedy algorithm terminates at step r, we have

$$\begin{aligned} g\left(S_r\right) &\geq \left(1 - \left(1 - \frac{\kappa}{r}\right)^r\right) g(S^*) \\ &\geq \left(1 - \exp(-\kappa)\right) g(S^*). \end{aligned}$$

Proof of Theorem 3.4: Let $S^* = \{x_1, \ldots, x_r\}$. By monotonicity, we have

$$\begin{aligned} g(S^*) &\leq g\left(S^* \cup S_i\right) \\ &= g(S_i) + \left(g\left(S_i \cup \{x_1\}\right) - g(S_i)\right) + \sum_{j=2}^{r} \left(g\left(S_i \cup \{x_1, \ldots, x_{j-1}, x_j\}\right) - g\left(S_i \cup \{x_1, \ldots, x_{j-1}\}\right)\right) \\ &\leq g(S_i) + \left(g(S_{i+1}) - g(S_i)\right) + \sum_{j=2}^{r} \frac{1}{\kappa} \left(g(S_i \cup \{x_j\}) - g(S_i)\right) \\ &\leq g(S_i) + \left(g(S_{i+1}) - g(S_i)\right) + \frac{r-1}{\kappa} \left(g(S_{i+1}) - g(S_i)\right) \\ &\leq g(S_i) + \frac{r}{\kappa} \left(g(S_{i+1}) - g(S_i)\right). \end{aligned}$$

To establish the approximation guarantee, we use induction. The base case $i = 0$ is trivial:

$$g(S_0) = 0 \geq \left(1 - \left(1 - \frac{\kappa}{r}\right)^0\right) g(S^*) = 0.$$

Assume the induction hypothesis

$$g(S_i) \geq \left(1 - \left(1 - \frac{\kappa}{r}\right)^i\right) g(S^*).$$

From the inequality we established earlier, we have

$$\begin{aligned}
& g(S^*) \leq g(S_i) + \frac{r}{\kappa}\left(g(S_{i+1}) - g(S_i)\right) \\
\iff\ & g(S_{i+1}) \geq g(S_i) + \frac{\kappa}{r}\left(g(S^*) - g(S_i)\right) \\
& = \frac{\kappa}{r} g(S^*) + \left(1 - \frac{\kappa}{r}\right) g(S_i) \\
& \geq \frac{\kappa}{r} g(S^*) + \left(1 - \frac{\kappa}{r}\right)\left(1 - \left(1 - \frac{\kappa}{r}\right)^i\right) g(S^*) \\
& \geq \left(1 - \left(1 - \frac{\kappa}{r}\right)^{i+1}\right) g(S^*).
\end{aligned}$$

This completes the induction. Setting $i = r$, we get the desired approximation guarantee. ■

4 Discussion and Remarks

In this paper, we borrowed the notion of atomic sets to capture a general notion of sparsity. We proved that for (restricted) strongly convex, smooth objectives, the performance of the greedy algorithm for unconstrained optimization is connected to a geometric quantity of the objective, the condition number μ/L, and a geometric quantity of the atomic set, θ_r. By deriving the θ_r values for various atomic sets, we established explicit approximation guarantees for greedy-type algorithms in various settings. We recovered the "strong" approximation guarantees appearing in recent literature for the feature selection and low-rank matrix optimization settings, where "strong" is quantified by the fact that the greedy algorithm will find a constant-factor approximation of the optimal r-sparse solution within $O(r)$ iterations. Namely, we can guarantee the greedy algorithm will produce a *sparse* solution that attains objective value on the same order as the optimal solution. Through the θ_r value, we have provided a simple method of deriving greedy approximation guarantees for theoretically any atomic set. We believe that the θ_r value of an atomic set furthermore holds some degree of truth regarding the ability of greedy search to discover good sparse solutions, since it measures the ability of the atomic set to approximate any element of an arbitrary r-subspace. Finding the atom best aligned to the gradient of the objective function is precisely the "greedy" part of one variant of the greedy algorithm, which have previously mentioned to be the computationally feasible alternative of pure greedy selection.

A secondary contribution of this paper was disentangling the notion of weak submodularity from the performance of greedy algorithms for continuous, convex optimization. Certainly, strongly convex, smooth objective functions on certain atomic sets can be shown to satisfy explicit degrees of weak submodularity. However, two key properties required of an atomic set to fruitfully connect weak submodularity and convex optimization is first whether the atomic set contains (approximately) orthogonal bases to the ambient space, and second whether the union of two sparse sets of atoms can be sparsely reparameterized into a set of (approximately) orthogonal atoms such that the span of the former is equal to (or contained in) the span of the latter. If the atomic set cannot guarantee the above two properties, one is hard-pressed to leverage the properties of strong convexity and smoothness to bound weak submodularity. However, by separately proving approximation guarantees of greedy algorithms for strongly convex, smooth functions versus weakly submodular functions, we now have access to guarantees of greedy algorithms on a much richer variety of functions.

Since our model is unconstrained optimization, we note that our model does not account for certain atomic sets where the norm of the atoms comes into play. For example, take the problem of recovering a low-rank and sparse matrix decomposition, such as in robust PCA [29]. Earlier literature has shown that this problem admits a convex relaxation. Namely, one can frame it as an atomic norm regularization problem, where the corresponding atomic norm ball is the convex hull of rank-one matrices $\left\{uv^\top\right\}$ and some multiple of the singleton matrices $\gamma\left\{e_i e_j^\top\right\}$ [30]. However, our greedy algorithm is agnostic to the norm of atoms, and since singleton matrices are themselves rank-one, optimizing over the above atomic set is equivalent to optimizing over the normalized rank-one matrices. In general, this issue occurs when one wants to promote selecting from a subfamily of a larger atomic set. This is fundamental to the unconstrained nature of the optimization, and thus there is no obvious adjustment that can be made.

An possibly interesting direction of future research would be to address the sparsity parameter in the greedy algorithm. As stated in this paper, the sparsity parameter is entirely user-determined, and thus it is up to the user to somehow determine the "true" sparsity of the model. It may be interesting to consider principled, yet computationally cheap approaches to selecting an "optimal" sparsity. As an example, we previously mentioned that PCA is a special case of applying the greedy algorithm to the linear recovery objective. In that case, parallel analysis [31, 32] has long been known as a very good heuristic for estimating the rank, and much more recently been shown theoretically to select the optimal rank under some model and noise assumptions [33, 34]. Heuristics and theory either for more general objective functions or different atomic sets will likely involve vastly different tools, but it is nevertheless an interesting problem.

Acknowledgements: The author is currently a postgraduate researcher in the lab of Prof. Yuval Kluger, Yale Applied Mathematics Program, and would like to thank Prof. Kluger for his support. The author would also like to thank Prof. Sahand Negahban for introducing him to the topic as an undergraduate thesis advisor, as well as many helpful comments on a nascent version of this paper. Any errors and aberrant conclusions are the author's own.

References

[1] N. Rao, P. Shah, and S. Wright, "Forward–backward greedy algorithms for atomic norm regularization," *IEEE Transactions on Signal Processing*, vol. 63, no. 21, pp. 5798–5811, 2015.

[2] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[3] S. Lacoste-Julien and M. Jaggi, "On the global linear convergence of frank-wolfe optimization variants," in *Advances in Neural Information Processing Systems*, pp. 496–504, 2015.

[4] F. Locatello, R. Khanna, M. Tschannen, and M. Jaggi, "A unified optimization view on generalized matching pursuit and frank-wolfe," *arXiv preprint arXiv:1702.06457*, 2017.

[5] A. M. Tillmann and M. E. Pfetsch, "The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1248–1259, 2013.

[6] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538–557, 2012.

[7] P.-L. Loh and M. J. Wainwright, "Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima," in *Advances in Neural Information Processing Systems*, pp. 476–484, 2013.

[8] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *Journal of Machine Learning Research*, vol. 13, no. May, pp. 1665–1697, 2012.

[9] E. R. Elenberg, R. Khanna, A. G. Dimakis, and S. Negahban, "Restricted strong convexity implies weak submodularity," *The Annals of Statistics*, vol. 46, no. 6B, pp. 3539–3568, 2018.

[10] R. Khanna, E. Elenberg, A. G. Dimakis, and S. Negahban, "On approximation guarantees for greedy low rank optimization," *arXiv preprint arXiv:1703.02721*, 2017.

[11] R. Khanna, E. Elenberg, A. G. Dimakis, S. Negahban, and J. Ghosh, "Scalable greedy feature selection via weak submodularity," *arXiv preprint arXiv:1703.02723*, 2017.

[12] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions," *Mathematical programming*, vol. 14, no. 1, pp. 265–294, 1978.

[13] A. Das and D. Kempe, "Approximate submodularity and its applications: subset selection, sparse approximation and dictionary selection," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 74–107, 2018.

[14] D. Goldfarb, G. Iyengar, and C. Zhou, "Linear convergence of stochastic frank wolfe variants," in *Artificial Intelligence and Statistics*, pp. 1066–1074, 2017.

[15] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[16] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[17] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[18] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[19] R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 255–261, 2006.

[20] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[21] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

[22] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[23] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the American control conference*, vol. 6, pp. 4734–4739, Citeseer, 2001.

[24] Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, vol. 140, no. 1, pp. 125–161, 2013.

[25] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[26] A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin, "Certifying the restricted isometry property is hard," *IEEE transactions on information theory*, vol. 59, no. 6, pp. 3448–3450, 2013.

[27] Z. Wang, M.-J. Lai, Z. Lu, W. Fan, H. Davulcu, and J. Ye, "Rank-one matrix pursuit for matrix completion," in *International Conference on Machine Learning*, pp. 91–99, 2014.

[28] T. Zhang, "Adaptive forward-backward greedy algorithm for learning sparse representations," *IEEE transactions on information theory*, vol. 57, no. 7, pp. 4689–4708, 2011.

[29] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[30] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[31] J. L. Horn, "A rationale and test for the number of factors in factor analysis," *Psychometrika*, vol. 30, no. 2, pp. 179–185, 1965.

[32] A. Buja and N. Eyuboglu, "Remarks on parallel analysis," *Multivariate behavioral research*, vol. 27, no. 4, pp. 509–540, 1992.

[33] E. Dobriban, "Permutation methods for factor analysis and pca," *arXiv preprint arXiv:1710.00479*, 2017.

[34] E. Dobriban and A. B. Owen, "Deterministic parallel analysis: An improved method for selecting the number of factors and principal components," *arXiv preprint arXiv:1711.04155*, 2017.

[35] M. Charikar and A. Wirth, "Maximizing quadratic programs: Extending grothendieck's inequality," in *45th Annual IEEE Symposium on Foundations of Computer Science*, pp. 54–60, IEEE, 2004.

[36] S. Shalev-Shwartz, A. Gonen, and O. Shamir, "Large-scale convex minimization with a low-rank constraint," *arXiv preprint arXiv:1106.1622*, 2011.

[37] M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing.* Springer Science & Business Media, 2010.

[38] R. A. Horn and C. R. Johnson, *Matrix analysis.* Cambridge university press, 2012.

A Computational Complexity of the Atomic Condition Number

An immediate question one might ask is: given an arbitrary atomic set $\mathcal{A}$, can we numerically compute $\theta(\mathcal{A})$ efficiently? As one may suspect, the atomic condition number cannot be computed efficiently in general. It suffices to restrict our attention to $\mathcal{A} \subseteq \mathbb{R}^n$ and $|\mathcal{A}| = m < \infty$. For convenience, let us assume $\|a_i\|_2 = 1$ for all $a_i \in \mathcal{A}$. Let $A \in \mathbb{R}^{n \times m}$ denote the matrix whose columns are $a_i \in \mathcal{A}$. Consider the following definition of θ:

$$\begin{aligned}\theta &= \min_{\|x\|_2=1} \max_{a \in \mathcal{A}} |\langle x, a \rangle| \\ &= \min_{x \in \mathbb{R}^n} \frac{\left\|A^\top x\right\|_\infty}{\|x\|_2}.\end{aligned}$$

Observe that when A is degenerate, i.e. $\mathcal{A}$ does not span $\mathbb{R}^n$, then $\theta = 0$. Therefore, we need only to consider A such that $m \geq n$ and $\operatorname{rank}(A) \geq n$. Observe that if we can solve the following problem efficiently, we can also compute θ efficiently:

$$\begin{aligned}(P) \quad \max \ & \|x\|_2^2 \\ \text{s.t. } & A^\top x \leq t\mathbb{1} \\ & A^\top x \geq -t\mathbb{1}.\end{aligned}$$

Setting $y = A^\top x$ and $Q := (AA^\top)^{-1} \succ 0$, we may re-write (P) as

$$\begin{aligned}(P') \quad \max \ & y^\top Q y \\ \text{s.t. } & -1 \leq y_i \leq 1 \quad \forall i \in [m].\end{aligned}$$

Since Q is positive-definite, (P') precisely coincides with a sub-case of the MaxQP problem considered in [35], where it is shown that quadratic programming over the ℓ^∞ hypercube is NP-hard, as the MaxCut problem is a special case MaxQP. Therefore, precisely computing θ is difficult in general. However, this does not preclude the possibility of estimating θ efficiently either through approximation schemes or local-search heuristics: estimation is certainly possible when $\mathcal{A} \subset \mathbb{R}^n$ and $|\mathcal{A}| < \infty$, but may not be as straightforward when $\mathcal{A}$ is uncountably infinite or a large combinatorial set, such as the cases of the rank-one matrices or the permutation matrices.

B Illustrative Experiments

We supplement our theoretical analyses with some numerical experiments. We note that there are plentiful experiments involving greedy algorithms applied on various large-scale real data for the more popular atomic sets, for example unit basis vectors [11, 9, 13] and low-rank matrices [10, 36]. In the first experiment, we measure the greedy algorithm's performance on linear recovery tasks.

That is, we generate a ground truth vector that is a linear combination of a predetermined (sparse) number of atoms, where the weights are ± 1 Bernoulli random variables. We feed to the greedy algorithm the least-squares objective function, where the target vector has been corrupted by 10% i.i.d. additive gaussian noise, as well as an overestimate of the sparsity. Since the realization of each iterate of the greedy algorithm does not depend on the sparsity parameter, one can easily determine the performance of the greedy algorithm had the sparsity been under, perfectly, or over-estimated. We validate the performance of the greedy algorithm on a different noisy measurement of the target vector. The lines indicate the true sparsity of the solution, as well as the objective/validation functions evaluated on the true target vector.

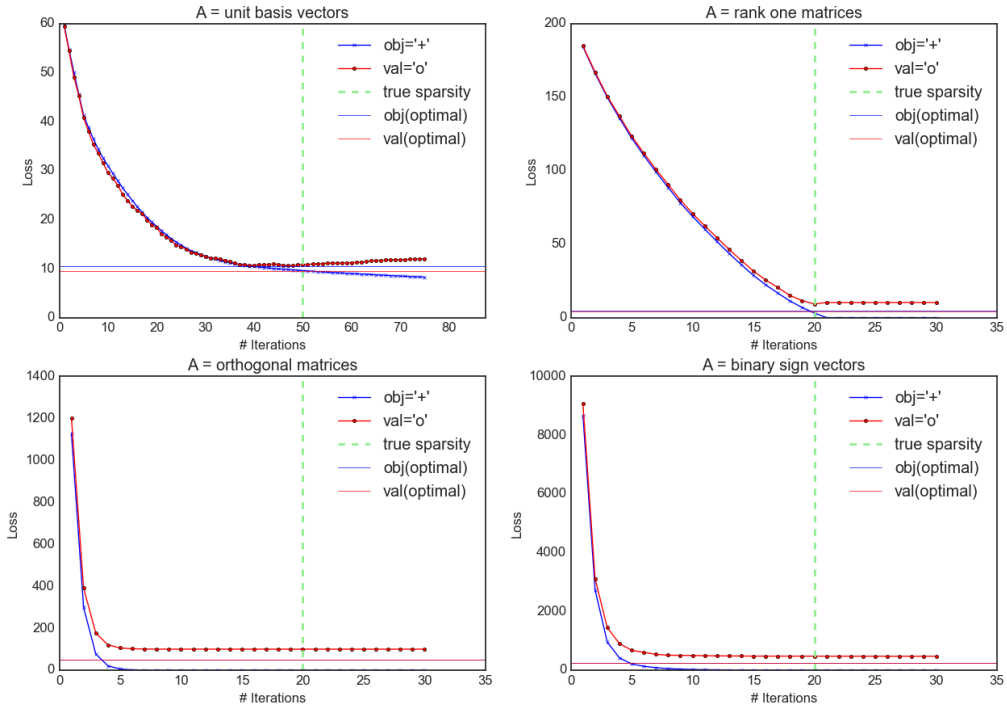

Figure 1: Performance of the greedy algorithm on linear recovery tasks for various atomic sets

We observe that the greedy algorithm performs extremely well for all the above recovery tasks, and finds optimal sparse solutions for all the above atomic sets. This is unsurprising for the unit basis vectors and rank-one matrices: for the unit basis vectors, each iteration of the greedy algorithm finds the next largest entry in absolute value; for the rank-one matrices, the greedy algorithm is equivalent to rank-k PCA or SVD of the measurement matrix, where k is the inputted sparsity parameter. Both these schemes will essentially find the optimal solution under our noise assumptions. What is slightly surprising is that the greedy algorithm converges to optimality in the orthogonal matrix and sign vectors case, especially how quickly it converges to 0. One possible explanation of this convergence rate might be related to the identifiability of linear combinations of those atoms. While in the unit basis and rank-one matrices case, the linear combination of k distinct atoms will (almost always) at best be k-sparse, in the orthogonal matrix and sign vectors case identifiability issues may arise, where certain linear combinations of k atoms are actually better represented by fewer atoms.

We demonstrate the relationship between certain quantities of an atomic set and the atomic condition number. Given an arbitrary set of vectors as an atomic set, it is a difficult problem to estimate the atomic condition number, as demonstrated in section A of the appendix. We restrict our attention to collection of unit vectors in $\mathbb{R}^n$, where computing the atomic condition number reduces to solving

$$\begin{aligned}\theta &= \min_{\|x\|_2=1} \max_{a\in\mathcal{A}} |\langle x, a\rangle| \\ &= \min_{x\in\mathbb{R}^n} \frac{\left\|A^\top x\right\|_\infty}{\|x\|_2},\end{aligned}$$

where A is the matrix that contains the atoms as column vectors. Given A, the latter value can be efficiently estimated in practice using local-search methods. As a preliminary, we observe that $\theta = n^{-1/2}$ is maximal for $A \in M_n$, and is attained if and only if A is orthogonal. This can be easily established by observing that, given an arbitrary non-singular matrix A, θ is monotonically increased by taking a vector from A and orthogonalizing it against the rest of the atoms. In the first experiment, we start with $\mathcal{A} = \{e_i\}_{i=1}^5$, $A = I_5$. We then corrupt the first column of A by setting $e_1 = v/\|v\|$, where $v = \lambda e_1 + (1-\lambda)s$ for different values of $0 \le \lambda \le 1$, where $s \in \{\pm 1\}^5$ is a fixed binary sign vector and $s(1) = -1$. For the latter two experiments, we generate many A randomly as i.i.d. gaussian random matrices and normalize the columns, then for each A we measure two values that are somewhat related to how close to "orthogonal" the matrix is, namely the mean coherence (mean of $|\langle a_i, a_j\rangle|$ for each $i \neq j$) and smallest singular value of A and plot the relationship with the atomic condition number.

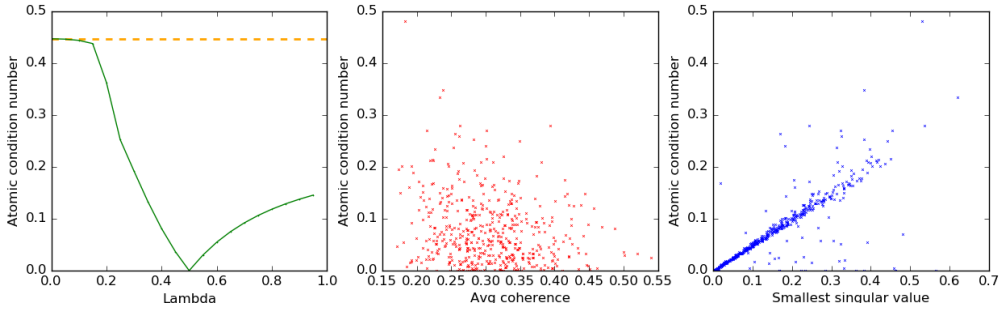

Figure 2: Relationship of the atomic condition number θ and various values

In the leftmost graph corresponding to the first experiment, the dotted orange line indicates the optimal $\theta = 1/\sqrt{5}$. The trend observed is not particularly surprising. At $\lambda = 0.5$, observe that since $s(1) = -1$, the first element of $v = e_1 + s$ is 0, which means that A is singular, and therefore the atomic condition number is 0. In the plot of atomic condition number versus the mean coherence, we observe a somewhat negative correlation, which makes sense, as a high average coherence indicates that many of the vectors in A are highly correlated, which implies θ is not optimal. Perhaps most surprising is the plot of atomic condition number versus the smallest singular value of A, where one observes an extremely tight linear relationship between the smallest singular value and atomic condition number (where many of the outliers can be attributed to the numerical inconsistency of the local search heuristic used to compute θ). At first glance, this might seem feasible, as the smallest singular value and θ can be written in similar ways:

$$\sigma_{\min} = \min_{x\in\mathbb{R}^n} \frac{\|Ax\|_2}{\|x\|_2}, \quad \theta = \min_{x\in\mathbb{R}^n} \frac{\|Ax\|_\infty}{\|x\|_2}.$$

However, there cannot be a linear relationship between the smallest singular value of A and the atomic condition number, as we have previously established in Appendix A that computing the atomic condition number of $\mathcal{A} \subset \mathbb{R}^n$ is NP-hard, while computing the smallest singular value is not.

C Computing Sparse Atomic Condition Numbers in Table 1

We recall the mathematical definitions of the atomic and sparse atomic condition numbers θ and θ_r: given the ambient vector space V of atomic set $\mathcal{A}$

$$\theta := \min_{\|v\|=1} \max_{a \in \mathcal{A}} \frac{|\langle v, a \rangle|}{\|a\|}$$
$$\theta_r := \min_{\substack{U \subset \mathcal{A} \\ |U| \leq r}} \min_{v \in \mathbf{span}(U)} \max_{a \in \mathcal{A}} \frac{|\langle v, a \rangle|}{\|a\|}.$$

θ measures the ability of the atomic set to approximate any vector in the ambient space, where θ_r measures the ability to approximate any vector given it comes from a subspace of dimension no more than r. We observe that $\theta_r \geq \theta$, and $\theta_r \geq \theta_k$ if $r \geq k$.

Standard Basis

The derivations of θ and θ_r are essentially identical. For θ, given any $v \in \mathbb{R}^n$, $\max_{i \in [n]} |\langle v, e_i \rangle| / \|v\|$ is attained by the largest entry in absolute value. It is then easy to see that

$$\theta = \min_{\|v\|=1} \max_{i \in [n]} |\langle v, e_i \rangle|$$

is attained by the scaled all-ones vector: $v = n^{-1/2}\mathbb{1}$. Similarly, if $U \subset \{e_i\}^n$, $|U| = r$, then θ_r is attained by the scaled indicator vector of U: $v = r^{-1/2}\mathbb{1}_U$, which gives us $\theta_r = r^{-1/2}$. We note that by finding the vectors that attain θ and θ_r, our lower bounds are tight.

The θ and θ_r values for the standard basis hold for any orthogonal basis, a fact that we will use in the analysis of the 2-ortho basis case.

Rank-One Matrices

The derivation of θ and θ_r for $\mathcal{A} = \left\{uv^\top\right\}$ is very similar to that for the standard basis, where we instead look at the spectrum of a given matrix. For θ, if $M \in \mathbb{R}^{m,n}$, we have $\text{rank}\,(M) \leq \min\{m, n\}$.

$$\begin{aligned}
\theta &= \min_{M} \max_{uv^\top, \|u\|\|v\|=1} \frac{\left\langle M, uv^\top \right\rangle}{\|M\|_F} \\
&= \min_{M} \max_{uv^\top, \|u\|\|v\|=1} \frac{1}{\|M\|_F} \text{tr}\left(u^\top M v\right) \\
&= \min_{M} \frac{1}{\|M\|_F} \sigma_1(M)
\end{aligned}$$

where $\sigma_1(M)$ is the largest singular value of M. Defining $\sigma(M)$ as the vector containing the singular values of M, we recall that $\|M\|_F = \|\sigma(M)\|_2$. Therefore, as in the case of the standard basis, the spectrum vector that attains θ is the scaled all-ones: $\sigma(M) = n^{-1/2}\mathbb{1}$, which of course has leading singular value $\sigma_1(M) = n^{-1/2}$.

For θ_r, given $U \subset \left\{uv^\top\right\}$, $|U| = r$, any matrix $M \in \mathbf{span}\,(V)$ is the linear combination of at most r rank-one matrices, and therefore is at most rank r. In other words, we have $|\mathrm{supp}\,(\sigma(M))| \leq r$. Following the argument for the standard basis, we have $\theta = r^{-1/2}$. The values derived for θ and θ_r are tight.

Disjoint Group Sparse Atoms

Given the standard basis $\{e_i\}^n \subset \mathbb{R}^n$, disjoint group-sparse atoms are defined as the elements of a partition of the basis: $\mathcal{A} = \mathcal{P}(\{e_i\}^n)$. Let $|\mathcal{A}| = L$. Note that these atoms are not vectors, and thus we must make a few adjustments to some definitions. Given objective function f, the accompanying set function g is now defined

$$g(\mathcal{S}) = \max_{x \in \mathbf{span}\left(\bigcup_{P\in\mathcal{S}} P\right)} f(x) - f(0),$$

where $\mathcal{S} \subset \mathcal{P}$ and P are the groups contained in $\mathcal{S}$. In other words, $g(\mathcal{S})$ returns the shifted optimal value of f searching over the vectors whose support lies in the union of the groups. Applying this new definition to our algorithm, $\texttt{PureGreedy}(\mathcal{A}, f, L_{t-1}, \beta)$ makes sense as is. For $\texttt{OMPSel}(\mathcal{A}, f, L_{t-1}, \beta)$, we modify the definition such that it returns the group P_t satisfying

$$\left\|\mathrm{proj}_{P_t}\left(\nabla f\left(B^{(L_{t-1})}\right)\right)\right\| \geq \beta \max_{P\in\mathcal{A}} \left\|\mathrm{proj}_P\left(\nabla f\left(B^{(L_{t-1})}\right)\right)\right\|.$$

In other words, whereas usually $\texttt{OMPSel}$ finds the atom (vector) that best explains the gradient of the previous iterate, it now finds the group (subspace) that can best explain the gradient of the previous iterate. We are now able to define θ and θ_r.

$$\begin{aligned}\theta &= \min_{\|v\|=1} \max_{P\in\mathcal{A}} \frac{|\langle v, \mathrm{proj}_P(v)\rangle|}{\|\mathrm{proj}_P(v)\|} \\ &= \min_{\|v\|=1} \max_{P\in\mathcal{A}} \|\mathrm{proj}_P(v)\|,\end{aligned}$$

and similarly,

$$\theta_r = \min_{\substack{U\subset\mathcal{A}\\|U|=r}} \min_{\substack{v\in\mathbf{span}\left(\bigcup_{P\in U} P\right)\\\|v\|=1}} \max_{P\in\mathcal{A}} \|\mathrm{proj}_P(v)\|.$$

It suffices for us to lower bound θ_r; the proof for θ is identical, setting $r = L$. Given any partition $\mathcal{P}$ of $\mathbb{R}^n$, and any subset $U \subseteq \mathcal{P}$, $|U| = r$, we consider what the minimizing vector v that attains θ_U would look like. Analogous to what we saw in the proofs for the standard basis and rank-one matrices, if we can find a vector v such that $\|\mathrm{proj}_P(v)\| = r^{-1/2}$ for all $P \in U$, then v is clearly the minimizing vector that attains θ_U, since $\|v\| = 1$ and shifting any mass around in the vector can only cause $\max_{P\in U} \|\mathrm{proj}_P(v)\|$ to increase. Let us construct such a v. Consider a vector $v = \sum_{i=1}^r c_i \mathbb{1}_{P_i}$, where $\mathbb{1}_{P_i} \in \mathbb{R}^n$ is the indicator vector corresponding to the group $P_i \in U$. We

have the constraint $\|v\| = 1$, and the property $\left\|\text{proj}_{P_i}(v)\right\| = r^{-1/2}$ for all i. In other words, a vector v attaining θ_U would satisfy

$$\sum_{i=1}^{r} c_i^2 \, |P_i| = 1$$
$$|c_i| \sqrt{|P_i|} = r^{-1/2} \quad i = 1, \ldots, r.$$

A possible solution is $c_i = (r \, |P_i|)^{-1/2}$, $i = 1, \ldots, r$. Substituting v into our expression for θ and θ_r, we get $\theta = n^{-1/2}$ and $\theta_r = r^{-1/2}$, and these lower bounds are tight our construction.

2-Ortho Basis

In the signal processing and dictionary learning communities, the 2-ortho basis refers to the union of the standard and Fourier orthogonal bases [15, 37]. We will consider a general union of two orthogonal bases, w.l.o.g. the union of the standard basis and an arbitrary orthogonal basis. Let us denote this atomic set

$$\mathcal{A} := \{e_i\}_{i=1}^n \cup \{\psi_j\}_{j=1}^n \subset \mathbb{R}^n.$$

A popular, albeit crude, measure of distance between bases is the "mutual coherence". For example, we have for the standard-union-Fourier,

$$\mu(\mathcal{A}) := \max_{i,j} \left| e_i^\top \psi_j \right| = n^{-1/2}.$$

We note that $n^{-1/2}$ is the lower bound on the mutual coherence of two orthogonal bases. We also recall that the (non-sparse) atomic condition number θ is monotonically increasing with respect to adding more elements. One might therefore hope that there is some link between the mutual coherence and a factor of "improvement" to θ. However, we will show that in general, the atomic condition number θ cannot be improved above $n^{-1/2}$ by example.

We consider the extremal case of the Hadamard basis (matrix), which is an orthogonal basis consisting of vectors in $\{\pm 1\}^n$. To be sure, Hadamard bases do not exist for every dimension (consider any odd dimension greater than 1), but there exist infinitely many. We consider a special infinite subfamily of the Hadamard bases: the regular Hadamard bases. Regular Hadamard bases are simply Hadamard matrices whose row and column sums are all equal, which restricts the dimension of the basis to square numbers. In particular, if the dimension of the regular Hadamard basis/matrix is $n = 4u^2$, then the row and column sums are all equal to $\pm 2u$, which further implies that each column has $2u^2 \pm u$ positive entries and $2u^2 \mp u$ negative entries. It is simple to see that any Hadamard basis attains the minimal $n^{-1/2}$ mutual coherence with the standard basis. We now consider the scaled all-ones vector $n^{-1/2}\mathbb{1}$, which we recall attains θ when $\mathcal{A}$ is the standard basis. For any member v of the regular Hadamard basis, its inner product with the all-ones vector is also

$$\frac{|\langle v, \mathbb{1} \rangle|}{\|v\| \, \|\mathbb{1}\|} = \frac{1}{n} 2u = \frac{1}{n} \sqrt{n} = n^{-1/2}.$$

In other words, the all-ones vector *also* attains θ for the regular Hadamard basis. Therefore, θ of the union of the standard and Hadamard basis is still $n^{-1/2}$. This means that in full generality, there may be no relation between the mutual coherence of two orthogonal bases $\mu(\mathcal{A})$ and the atomic condition number θ.

It remains to be seen what happens to the sparse atomic condition number θ_r. We have seen that arbitrarily adding atoms to the standard basis, though improving θ, may instantaneously cause θ_r to shrink from $r^{-1/2}$ to $n^{-1/2}$. Therefore, the more may not be the merrier when it comes to the sparse atomic condition number. However, we note that the example we used to demonstrate corrupting the θ_r value, where we added $e_1 + \varepsilon \mathbb{1}$ to $\{e_i\}_{i=1}^n$, would have a mutual coherence arbitrarily close to 1 (since we are essentially adding a slightly perturbed member of the standard basis). However, what happens when we can guarantee two orthogonal bases are a certain angle away from each other? We will show that when the sparsity is below a certain level $r \leq \mu(\mathcal{A})^{-1}$, then the sparse atomic condition number is lower bounded by $\Omega\left(r^{-1/2}\right)$. In other words, under sufficient sparsity, the sparse condition number of a 2-ortho basis is on the same order as just one orthogonal basis.

Let $\{e_i\}_{i=1}^n$ be an orthogonal basis (w.l.o.g. the standard basis) and $\{\psi_j\}_{j=1}^n$ be another orthogonal basis, with mutual coherence μ. Consider any vector

$$x = \sum_{i \in I} c_i e_i + \sum_{j \in J} d_j \psi_j \neq 0$$

where $|I| = r_1$, $|J| = r_2$, $r_1 + r_2 \leq \mu^{-1}$. For our purposes, we can assume $\|x\| = 1$. We may also assume $r_1, r_2 > 0$, since if either one is 0, then we are reduced to computing the sparse atomic number of a single orthogonal basis. From the AM-GM inequality, we have $\sqrt{r_1 r_2} \leq (r_1 + r_2)/2 \leq \mu^{-1}/2$. We also have

$$\|x\|^2 = \|c\|^2 + \sum_{i \in I} \sum_{j \in J} c_i d_j \langle e_i, \psi_j \rangle + \|d\|^2,$$

where c and d are the coefficient vectors from x. From the definition of mutual coherence, we have $-\mu^{-1} \leq \langle e_i, \psi_j \rangle \leq \mu^{-1}$. Thus, we can make a crude upper and lower bound on $\|x\|^2$ with respect to $\|c\|$ and $\|d\|$:

$$\begin{aligned}
\|x\|^2 &\leq \|c\|^2 + \|d\|^2 + \|c\|_1 \|d\|_1 \mu \\
&\leq \|c\|^2 + \|d\|^2 + \sqrt{r_1 r_2} \|c\| \|d\| \mu \\
&\leq \|c\|^2 + \|d\|^2 + \frac{1}{2} \|c\| \|d\| \\
&\leq \frac{5}{4} \left(\|c\|^2 + \|d\|^2 \right), \\
\|x\|^2 &\geq \|c\|^2 + \|d\|^2 - \|c\|_1 \|d\|_1 \mu \\
&\geq \frac{3}{4} \left(\|c\|^2 + \|d\|^2 \right).
\end{aligned}$$

Without loss of generality, let $\|c\| \geq \|d\|$. Hence, $\|c\| \geq \sqrt{\frac{4}{10}} \|x\|$. Let $c_p = \arg\max_{i \in I} |c_i|$. Letting

$\|x\| = 1$, we have

$$\begin{aligned}
\frac{|\langle x, c_p\rangle|}{\|x\|\,\|c_p\|} &\geq |c_p| - \mu \sum_{j\in J} |d_j| \\
&\geq \|c\| / \sqrt{r_1} - \sqrt{r_2}\mu \|d\| \\
&\geq \|c\| / \sqrt{r_1} - \|d\| / (2\sqrt{r_1}) \quad (\text{since } \sqrt{r_1 r_2} \leq \mu^{-1}/2) \\
&\geq \|c\| / (2\sqrt{r_1}) \\
&\geq \|c\| / \left(2\sqrt{r}\right) \\
&\geq \left(\sqrt{4/10}\,\|x\|\right) \frac{1}{2} r^{-1/2} \\
&\geq 10^{-1/2} r^{-1/2} \\
&= \Omega\left(r^{-1/2}\right).
\end{aligned}$$

Therefore, for any $r \leq \mu^{-1}$, we have

$$\begin{aligned}
\theta_r &:= \min_{\substack{U\subset\mathcal{A} \\ |U|\leq r}} \min_{v\in \mathbf{span}(U)} \max_{a\in\mathcal{A}} \frac{|\langle v, a\rangle|}{\|a\|} \\
&= \Omega\left(r^{-1/2}\right).
\end{aligned}$$

We note that μ^{-1} is at maximum $\sqrt{n}$, which is attained by the classic standard-union-Fourier basis or the standard-union-Hadamard basis we discussed.

Binary sign vectors

Given vector v, it is clear that the atom $a \in \{\pm 1\}^n$ that maximizes $\frac{\langle v,a\rangle}{\|a\|}$ is $a = \text{sign}(v)$, where $\text{sign}(v)$ is the binary sign vector whose entries are the signs of the entries of v. Therefore, we have

$$\begin{aligned}
\theta &= \min_{\|v\|=1} \max_{a\in\mathcal{A}} \frac{|\langle v, a\rangle|}{\|a\|} \\
&= \min_{\|v\|=1} n^{-1/2} \langle v,\ \text{sign}(v)\rangle \\
&= \min_{\|v\|=1} n^{-1/2} \|v\|_1 \\
&= n^{-1/2},
\end{aligned}$$

where the last line comes from the ℓ^1-ℓ^2 equivalence of norms inequality, which is attained by $v = e_i$ for any standard basis vector e_i. We observe that θ_r cannot be any larger than $n^{-1/2}$ by a simple example. Let U be any set containing a_1, a_2, which are two sign vectors that are identical apart from one entry where the sign is flipped, without loss of generality the first entry. Take $x = \frac{1}{2}a_1 + \frac{1}{2}a_2 = e_1$, which is in $\mathbf{span}\,(U)$. As we saw earlier, e_1 is a vector that attains θ, and therefore $\theta_r = \theta = n^{-1/2}$.

Orthogonal Matrices

The orthogonal matrices are an interesting atomic set where the θ and θ_r values are precisely identical. Let $\|\cdot\|_F$ denote the Frobenius norm of a matrix, and $\|\cdot\|_*$ denote the nuclear norm (or

trace norm). Given any matrix M, $\|M\|_F = 1$ and its singular value decomposition $M = U\Sigma V^\top$, we know that the closest orthogonal matrix in Frobenius norm to M is the matrix $Q = UV^\top$ [38]. From the identity

$$\begin{aligned}\|M - Q\|_F^2 &= \|M\|_F^2 + \|Q\|_F^2 - 2\langle M, Q\rangle \\ &= \|M\|_F^2 + n - 2\langle M, Q\rangle\end{aligned}$$

we observe the optimal solutions of the following two problems are identical

$$\min_{Q:\, Q^\top Q = I_n} \|M - Q\|_F^2 \equiv \max_{Q:\, Q^\top Q = I_n} \langle M, Q\rangle .$$

We can now establish the lower bound for θ

$$\begin{aligned}\theta &:= \min_{\|M\|_F = 1} \max_{Q \in \mathcal{A}} \frac{|\langle M, Q\rangle|}{\|Q\|} \\ &= \min_{\|M\|_F = 1} n^{-1/2} \mathrm{tr}\left((UV^\top)^\top M\right) \\ &= \min_{\|M\|_F = 1} n^{-1/2} \mathrm{tr}\left(\Sigma\right) \\ &= \min_{\|M\|_F = 1} n^{-1/2} \|M\|_* .\end{aligned}$$

Since the nuclear norm and Frobenius norms are special cases of Schatten p-norms, with $p = 1, 2$ respectively, which are defined as p-norms on the singular values of a matrix, we have from the equivalence of norms: $\|M\|_* \geq \|M\|_F$, which gets us the lower bound on θ:

$$\begin{aligned}\theta &= \min_{\|M\|_F = 1} n^{-1/2} \|M\|_* \\ &\geq \min_{\|M\|_F = 1} n^{-1/2} \|M\|_F \\ &= n^{-1/2}.\end{aligned}$$

It is simple to see that θ_r cannot be better than $n^{-1/2}$ by considering the following example. Let $r = 2$ and $U = \{I_n, Q\}$, where

$$Q = \begin{bmatrix} ((d-1)/d)^{-1/2} & -d^{-1/2} & \\ d^{-1/2} & ((d-1)/d)^{-1/2} & \\ & & I_{n-2} \end{bmatrix}.$$

and we set $M = 2\sqrt{1 - \sqrt{\frac{d-1}{d}}}\,(I_n - Q)$. Letting d be arbitrarily large, we see that θ_r can be made arbitrarily close to θ for any $r \geq 2$.