On the Convergence Rate of Decomposable Submodular Function Minimization

Robert Nishihara, Stefanie Jegelka, Michael I. Jordan
University of California, Berkeley
{rkn,stefje,jordan}@eecs.berkeley.edu

Abstract

Submodular functions describe a variety of discrete problems in machine learning, signal processing, and computer vision. However, minimizing submodular functions poses a number of algorithmic challenges. Recent work introduced an easy-to-use, parallelizable algorithm for minimizing submodular functions that decompose as the sum of “simple” submodular functions. Empirically, this algorithm performs extremely well, but no theoretical analysis was given. In this paper, we show that the algorithm converges linearly, and we provide upper and lower bounds on the rate of convergence. Our proof relies on the geometry of submodular polyhedra and draws on results from spectral graph theory.

1 Introduction

A large body of recent work demonstrates that many discrete problems in machine learning can be phrased as the optimization of a submodular set function [2, 30]. A set function \( F: 2^V \rightarrow \mathbb{R} \) over a ground set \( V \) of \( N \) elements is submodular if the inequality \( F(A) + F(B) \geq F(A \cup B) + F(A \cap B) \) holds for all subsets \( A, B \subseteq V \). Problems like clustering [33], structured sparse variable selection [1], MAP inference with higher-order potentials [27], and corpus extraction problems [31] can be reduced to the problem of submodular function minimization (SFM), that is

\[
\min_{A \subseteq V} F(A). \tag{P1}
\]

Although SFM is solvable in polynomial time, existing algorithms are often inefficient on large-scale problems. For this reason, the development of scalable, parallelizable algorithms has been an active area of research [23, 24, 28, 35]. Approaches to solving Problem (P1) are either based on combinatorial optimization or on convex optimization via the Lovász extension.

Functions that occur in practice are usually not arbitrary and frequently possess additional exploitable structure. For example, a number of submodular functions admit specialized algorithms that solve Problem (P1) very quickly. Examples include cut functions on certain kinds of graphs, concave functions of the cardinality \( |A| \), and functions counting joint ancestors in trees. We will use the term simple to refer to functions \( F \) for which we have a fast subroutine for minimizing \( F + s \), where \( s \in \mathbb{R}^N \) is any modular function. We treat these subroutines as black boxes. Many commonly occurring submodular functions (for example, graph cuts, hypergraph cuts, MAP inference with higher-order potentials [16, 27, 37], co-segmentation [21], certain structured-sparsity inducing functions [25], covering functions [35], and combinations thereof) can be expressed as a sum

\[
F(A) = \sum_{r=1}^{R} F_r(A) \tag{1}
\]

of simple submodular functions. Recent work demonstrates that this structure offers important practical benefits [24, 28, 35]. For instance, it admits iterative algorithms that minimize
each $F_r$ separately and combine the results in a straightforward manner (for example, dual decomposition).

In particular, it has been shown that the minimization of decomposable functions can be rephrased as a **best-approximation problem**, the problem of finding the closest points in two convex sets [24]. This formulation brings together SFM and classical projection methods and yields empirically fast, parallel, and easy-to-implement algorithms. In these cases, the performance of projection methods depends heavily on the specific geometry of the problem at hand and is not well understood in general. Indeed, while Jegelka et al. [24] show good empirical results, the analysis of this alternative approach to SFM was left as an open problem.

**Contributions.** In this work, we study the geometry of the submodular best-approximation problem and ground the prior empirical results in theoretical guarantees. We show that SFM via alternating projections, or block coordinate descent, converges at a **linear rate**. We show that this rate holds for the best-approximation problem, relaxations of SFM, and the original discrete problem. More importantly, we prove upper and lower bounds on the worst-case rate of convergence. Our proof relies on analyzing angles between the polyhedra associated with submodular functions and draws on results from spectral graph theory. It offers insight into the geometry of submodular polyhedra that may be beneficial beyond the analysis of projection algorithms.

**Submodular minimization.** The first polynomial-time algorithm for minimizing arbitrary submodular functions was a consequence of the ellipsoid method [18]. Strongly and weakly polynomial-time combinatorial algorithms followed [32]. The current fastest running times are $O(N^5 \tau_1 + N^6)$ [31] in general and $O((N^4 \tau_1 + N^5) \log F_{\text{max}})$ for integer-valued functions [22], where $F_{\text{max}} = \max_A |F(A)|$ and $\tau_1$ is the time required to evaluate $F$. Some work has addressed decomposable functions [24, 25, 33]. The running times in [25] apply to integer-valued functions and range from $O((N + R)^2 \log F_{\text{max}})$ for cuts to $O((N + Q^2 R)(N + Q^2 R + QR \tau_2) \log F_{\text{max}})$, where $Q \leq N$ is the maximal cardinality of the support of any $F_r$, and $\tau_2$ is the time required to minimize a simple function. Stobbe and Krause [33] use a convex optimization approach based on Nesterov’s smoothing technique. They achieve a (sublinear) convergence rate of $O(1/k)$ for the discrete SFM problem. Their results and our results do not rely on the function being integral.

**Projection methods.** Algorithms based on alternating projections between convex sets (and related methods such as the Douglas-Rachford algorithm) have been studied extensively for solving convex-feasibility and best-approximation problems [4, 5, 6, 11, 12, 13, 20, 36, 38]. See Deutsch [10] for a survey of applications. In the simple case of subspaces, the convergence of alternating projections has been characterized in terms of the Friedrichs angle $c_F$ between the subspaces $F_r$ [4, 6]. There are often good ways to compute $c_F$ (see Lemma 6), which allow us to obtain concrete linear rates of convergence for subspaces. The general case of alternating projections between arbitrary convex sets is less well understood. Bauschke and Borwein [3] give a general condition for the linear convergence of alternating projections in terms of the value $\kappa_*$ (defined in Section 3.1). However, except in very limited cases, it is unclear how to compute or even bound $\kappa_*$. While it is known that $\kappa_* < \infty$ for polyhedra [3, Corollary 5.26], the rate may be arbitrarily slow, and the challenge is to bound the linear rate away from one. We are able to give a specific uniform linear rate for the submodular polyhedra that arise in SFM.

Although both $\kappa_*$ and $c_F$ are useful quantities for understanding the convergence of projection methods, they largely have been studied independently of one another. In this work, we relate these two quantities for polyhedra, thereby obtaining some of the generality of $\kappa_*$ along with the computability of $c_F$. To our knowledge, we are the first to relate $\kappa_*$ and $c_F$ outside the case of subspaces. We feel that this connection may be useful beyond the context of submodular polyhedra.
1.1 Background

Throughout this paper, we assume that $F$ is a sum of simple submodular functions $F_1, \ldots, F_R$ and that $F(\emptyset) = 0$. Points $s \in \mathbb{R}^N$ can be identified with (modular) set functions via $s(A) = \sum_{a \in A} s_a$. The base polytope of $F$ is defined as the set of all modular functions that are dominated by $F$ and that sum to $F(V)$,

$$B(F) = \{ s \in \mathbb{R}^N | s(A) \leq F(A) \text{ for all } A \subseteq V \text{ and } s(V) = F(V) \}.$$

The Lovász extension $f : \mathbb{R}^N \to \mathbb{R}$ of $F$ can be written as the support function of the base polytope, that is $f(x) = \max_{s \in B(F)} s^\top x$. Even though $B(F)$ may have exponentially many faces, the extension $f$ can be evaluated in $O(N \log N)$ time [15]. The discrete SFM problem in Problem (P4) can be relaxed to the non-smooth convex optimization problem

$$\min_{x \in [0,1]^N} f(x) = \min_{x \in [0,1]^N} \sum_{r=1}^R f_r(x), \quad \text{(P2)}$$

where $f_r$ is the Lovász extension of $F_r$. This relaxation is exact – rounding an optimal continuous solution yields the indicator vector of an optimal discrete solution. The formulation in Problem (P2) is amenable to dual decomposition [20] and smoothing techniques [35], but suffers from the non-smoothness of $f$ [24]. Alternatively, we can formulate a proximal version of the problem

$$\min_{x \in \mathbb{R}^N} f(x) + \frac{1}{2}\|x\|^2 \equiv \min_{x \in \mathbb{R}^N} \sum_{r=1}^R (f_r(x) + \frac{1}{2R}\|x\|^2). \quad \text{(P3)}$$

By thresholding the optimal solution of Problem (P3) at zero, we also recover the indicator vector of an optimal discrete solution [2, Proposition 8.4].

**Lemma 1.** The dual of the right-hand version of Problem (P3) is the best-approximation problem

$$\min \|a - b\|^2 \quad a \in A, \ b \in B, \quad \text{(P4)}$$

where $A = \{(a_1, \ldots, a_R) \in \mathbb{R}^{NR} | \sum_{r=1}^R a_r = 0 \}$ and $B = B(F_1) \times \cdots \times B(F_R)$.

Lemma [1] is shown in [24]. It implies that we can minimize a decomposable submodular function by solving Problem (P4), which means finding the closest points between the subspace $A$ and the product $B$ of base polytopes. Projecting onto $A$ is straightforward because $A$ is a subspace. Projecting onto $B$ amounts to projecting onto each $B(F_r)$ separately. The projection $\Pi_{B(F_r)} z$ of a point $z$ onto $B(F_r)$ may be solved by minimizing $F_r - z$ [24]. We can compute these projections easily because each $F_r$ is simple.

Throughout this paper, we use $A$ and $B$ to refer to the specific polyhedra defined in Lemma [11] (which live in $\mathbb{R}^{NR}$) and $P$ and $Q$ to refer to general polyhedra (sometimes arbitrary convex sets) in $\mathbb{R}^D$. Note that the polyhedron $B$ depends on the submodular functions $F_1, \ldots, F_R$, but we omit the dependence to simplify our notation. Our bound will be uniform over all submodular functions.

2 Algorithm and Idea of Analysis

A popular class of algorithms for solving best-approximation problems are projection methods [5]. The most straightforward approach uses alternating projections (AP) or block coordinate descent. Start with any point $a_0 \in A$, and inductively generate two sequences via $b_k = \Pi_{B(a_k)}$ and $a_{k+1} = \Pi_{A(b_k)}$. Given the nature of $A$ and $B$, this algorithm is easy to implement and use in our setting. Jegelka et al. [24] propose to solve Problem (P4) with AP between $A$ and $B$. This is the algorithm that we will analyze.

The sequence $(a_k, b_k)$ will eventually converge to an optimal pair $(a_*, b_*)$. We say that AP converges linearly with rate $\alpha < 1$ if $\|a_k - a_*\| \leq C_1 \alpha^k$ and $\|b_k - b_*\| \leq C_2 \alpha^k$ for all $k$ and for some constants $C_1$ and $C_2$. Smaller values of $\alpha$ are better.
Analysis: Intuition. We will provide a detailed analysis of the convergence of AP for the polyhedra $A$ and $B$. To motivate our approach, we first provide some intuition with the following much-simplified setup. Let $U$ and $V$ be one-dimensional subspaces spanned by the unit vectors $u$ and $v$ respectively. In this case, it is known that AP converges linearly with rate $\cos^2 \theta$, where $\theta \in [0, \frac{\pi}{2}]$ is the angle such that $\cos \theta = u^\top v$. The smaller the angle, the slower the rate of convergence. For subspaces $U$ and $V$ of higher dimension, the relevant generalization of the “angle” between the subspaces is the Friedrichs angle [11], Definition 9.4, whose cosine is given by

$$c_F(U, V) = \sup \left\{ u^\top v \mid u \in U \cap (U \cap V)^\perp, v \in V \cap (U \cap V)^\perp, \|u\| \leq 1, \|v\| \leq 1 \right\}. \quad (2)$$

In finite dimensions, $c_F(U, V) < 1$. In general, when $U$ and $V$ are subspaces of arbitrary dimension, AP will converge linearly with rate $c_F(U, V)^2$ [11, Theorem 9.8]. If $U$ and $V$ are affine spaces, AP still converges linearly with rate $c_F(U - u, V - v)^2$, where $u \in U$ and $v \in V$.

We are interested in rates for polyhedra $P$ and $Q$, which we define as the intersection of finitely many halfspaces. We generalize the preceding results by considering all pairs $(P_x, Q_y)$ of faces of $P$ and $Q$ and showing that the convergence rate of AP between $P$ and $Q$ is at worst $\max_{x,y} c_F(\text{aff}(P_x), \text{aff}(Q_y))^2$, where $\text{aff}(C)$ is the affine hull of $C$ and $\text{aff}(C) = \text{aff}(C) - c$ for some $c \in C$. The faces $\{P_x\}_{x \in \mathbb{R}^d}$ of $P$ are defined as the maximizers of linear functions over $P$, that is

$$P_x = \arg \max_{p \in P} x^\top p. \quad (3)$$

While we look at angles between pairs of faces, we remark that Deutsch and Hundal [12] consider a different generalization of the “angle” between arbitrary convex sets.

Roadmap of the Analysis. Our analysis has two main parts. First, we relate the convergence rate of AP between polyhedra $P$ and $Q$ to the angles between the faces of $P$ and $Q$. To do so, we give a general condition under which AP converges linearly (Theorem 2), which we show depends on the angles between the faces of $P$ and $Q$ (Corollary 5) in the polyhedral case. Second, we specialize to the polyhedral $A$ and $B$, and we equate the angles with eigenvalues of certain matrices and use tools from spectral graph theory to bound the relevant eigenvalues in terms of the conductance of a specific graph. This yields a worst-case bound of $1 - \frac{1}{N^2 R^2}$ on the rate, stated in Theorem 12.

In Theorem 14 we show a lower bound of $1 - \frac{2\pi^2}{N^2 R^2}$ on the worst-case convergence rate.

3 The Upper Bound

We first derive an upper bound on the rate of convergence of AP between the polyhedra $A$ and $B$.

3.1 A Condition for Linear Convergence

We begin with a condition under which AP between two closed convex sets $P$ and $Q$ converges linearly. This result is similar to that of Bauschke and Borwein [3, Corollary 3.14], but the rate we achieve is twice as fast and relies on slightly weaker assumptions.

We will need a few definitions from Bauschke and Borwein [3]. Let $d(K_1, K_2) = \inf \{\|k_1 - k_2\| : k_1 \in K_1, k_2 \in K_2 \}$ be the distance between sets $K_1$ and $K_2$. Define the sets of “closest points” as

$$E = \{p \in P \mid d(p, Q) = d(P, Q)\} \quad \text{and} \quad H = \{q \in Q \mid d(q, P) = d(Q, P)\}, \quad (4)$$

and let $v = \Pi_{Q \setminus P} 0$ (see Figure 1). Note that $H = E + v$, and when $P \cap Q \neq \emptyset$ we have $v = 0$ and $E = H = P \cap Q$. Therefore, we can think of the pair $(E, H)$ as a generalization of the intersection $P \cap Q$ to the setting where $P$ and $Q$ do not intersect. Pairs of points
For $x \in \mathbb{R}^D \setminus E$, the function $\kappa$ relates the distance to $E$ with the distances to $P$ and $Q'$,

$$\kappa(x) = \frac{d(x, E)}{\max\{d(x, P), d(x, Q')\}}.$$ 

If $\kappa$ is bounded, then whenever $x$ is close to both $P$ and $Q'$, it must also be close to their intersection. If, for example, $D \geq 2$ and $P$ and $Q$ are balls of radius one whose centers are separated by distance exactly two, then $\kappa$ is unbounded. The maximum $\kappa_* = \sup_{x \in (P \cup Q') \setminus E} \kappa(x)$ is intimately related to the convergence rate.

**Theorem 2.** Let $P$ and $Q$ be convex sets, and suppose that $\kappa_* < \infty$. Then AP between $P$ and $Q$ converges linearly with rate $1 - \frac{1}{\kappa_*}$. Specifically,

$$\|p_k - p_*\| \leq 2\|p_0 - p_*\|(1 - \frac{1}{\kappa_*})^k \quad \text{and} \quad \|q_k - q_*\| \leq 2\|q_0 - q_*\|(1 - \frac{1}{\kappa_*})^k.$$ 

We prove Theorem 2 in Appendix A.1.

### 3.2 Relating $\kappa_*$ to the Angles Between Faces of the Polyhedra

In this section, we consider the case of polyhedra $P$ and $Q$, and we bound $\kappa_*$ in terms of the angles between pairs of their faces. In Lemma 3 we show that $\kappa$ is nondecreasing along the sequence of points generated by AP between $P$ and $Q'$. We treat points $p$ for which $\kappa(p) = 1$ separately because these are the points from which AP between $P$ and $Q'$ converges in one step. This lemma enables us to bound $\kappa(p)$ by initializing AP at $p$ and bounding $\kappa$ at some later point in the resulting sequence.

**Lemma 3.** For any $p \in P \setminus E$, either $\kappa(p) = 1$ or $1 < \kappa(p) \leq \kappa(\Pi_Q p)$. Similarly, for any $q \in Q' \setminus E$, either $\kappa(q) = 1$ or $1 < \kappa(q) \leq \kappa(\Pi_P q)$.

We prove Lemma 3 in Appendix A.3. We can now bound $\kappa$ by the angles between the faces of $P$ and $Q$.

**Proposition 4.** If $P$ and $Q$ are polyhedra and $p \in P \setminus E$, then there exist faces $P_x$ and $Q_y$ such that

$$1 - \frac{1}{\kappa(p)^2} \leq c_F(\text{aff}_0(P_x), \text{aff}_0(Q_y))^2.$$ 

The analogous statement holds when we replace $p \in P \setminus E$ with $q \in Q' \setminus E$.

Note that $\text{aff}_0(Q_y) = \text{aff}_0(Q'_y)$. We prove Proposition 4 in Appendix A.4. Proposition 4 immediately gives us the following corollary.

**Corollary 5.** If $P$ and $Q$ are polyhedra, then

$$1 - \frac{1}{\kappa_*^2} \leq \max_{x, y \in \mathbb{R}^D} c_F(\text{aff}_0(P_x), \text{aff}_0(Q_y))^2.$$
3.3 Angles Between Subspaces and Singular Values

Corollary 5 leaves us with the task of bounding the Friedrichs angle. To do so, we first relate the Friedrichs angle to the singular values of certain matrices in Lemma 6. We then specialize this to base polyhedra of submodular functions. For convenience, we prove Lemma 6 in Appendix A.4, though this result is implicit in the characterization of principal angles between subspaces given in Knyazev and Argentati 26, Section 1. Ideas connecting angles between subspaces and eigenvalues are also used by Diaconis et al. 14.

Lemma 6. Let $S$ and $T$ be matrices with orthonormal rows and with equal numbers of columns. If all of the singular values of $ST^\top$ equal one, then $c_F(\text{null}(S), \text{null}(T)) = 0$. Otherwise, $c_F(\text{null}(S), \text{null}(T))$ is equal to the largest singular value of $ST^\top$ that is less than one.

Faces of relevant polyhedra. Let $A_x$ and $B_y$ be faces of the polyhedra $A$ and $B$ from Lemma 11. Since $A$ is a vector space, its only nonempty face is $A_x = A$. Hence, $A_x = \text{null}(S)$, where $S$ is the $N \times NR$ matrix

$$S = \frac{1}{\sqrt{R}} \begin{pmatrix} I_N & \cdots & I_N \\ \text{repeated } R \text{ times} \end{pmatrix}. \quad (5)$$

Here, $I_N$ denotes the $N \times N$ identity matrix. The matrix for $\text{aff}_0(B_y)$ requires a bit more elaboration. Since $B$ is a Cartesian product, we have $B_y = B(F_1)_{y_1} \times \cdots \times B(F_R)_{y_R}$, where $y = (y_1, \ldots, y_R)$ and $B(F_r)$ is a face of $B(F_r)$. To proceed, we use the following characterization of faces of base polytopes [2, Proposition 4.7].

Proposition 7. Let $F$ be a submodular function, and let $B(F)_x$ be a face of $B(F)$. Then there exists a partition of $V$ into disjoint sets $A_1, \ldots, A_M$ such that

$$\text{aff}(B(F)_x) = \bigcap_{m=1}^M \{ s \in \mathbb{R}^N | s(A_1 \cup \cdots \cup A_m) = F(A_1 \cup \cdots \cup A_m) \}.$$ 

The following corollary is immediate.

Corollary 8. Define $F$, $B(F)_x$, and $A_1, \ldots, A_M$ as in Proposition 7. Then

$$\text{aff}_0(B(F)_x) = \bigcap_{m=1}^M \{ s \in \mathbb{R}^N | s(A_1 \cup \cdots \cup A_m) = 0 \}.$$ 

Corollary 8 implies that, for each $F_r$, there exists a partition of $V$ into disjoint sets $A_{r1}, \ldots, A_{rM_r}$ such that

$$\text{aff}_0(B_y) = \bigcap_{r=1}^R \bigcap_{m=1}^{M_r} \{ (s_1, \ldots, s_R) \in \mathbb{R}^{NR} \mid s_r(A_{r1} \cup \cdots \cup A_{rM_r}) = 0 \}.$$ \hspace{1cm} (6)

In other words, we can write $\text{aff}_0(B_y)$ as the nullspace of either of the matrices

$$T' = \begin{pmatrix} 1_{A_{11}}^\top \\ \vdots \\ 1_{A_{11} \cup \cdots \cup A_{1M_1}}^\top \\ \vdots \\ 1_{A_{R1}}^\top \\ \vdots \\ 1_{A_{R1} \cup \cdots \cup A_{RM_R}}^\top \end{pmatrix} \quad \text{or} \quad T = \begin{pmatrix} 1_{A_{11}}^\top \sqrt{|A_{11}|} \\ \vdots \\ 1_{A_{1M_1}}^\top \sqrt{|A_{1M_1}|} \\ \vdots \\ 1_{A_{R1}}^\top \sqrt{|A_{R1}|} \\ \vdots \\ 1_{A_{RM_R}}^\top \sqrt{|A_{RM_R}|} \end{pmatrix},$$
Lemma 10. For (stated in Appendix A.6) by bounding the Cheeger constant $h$ rows and columns are indexed by $(r, m)$ corresponding to row $(r, m)$ and column $(r, m)$ equals

$$ST^\top = \frac{1}{\sqrt{R}} \left( \begin{array}{cccc} \frac{1_{A_{11}}}{\sqrt{|A_{11}|}} & \cdots & \frac{1_{A_{1M}}}{\sqrt{|A_{1M}|}} & \cdots \frac{1_{A_{R1}}}{\sqrt{|A_{R1}|}} & \cdots & \frac{1_{AR_{RM}}}{\sqrt{|AR_{RM}|}} \end{array} \right)$$

that is less than one. We rephrase this conclusion in the following remark.

Remark 9. The largest eigenvalue of $(ST^\top)^\top (ST^\top)$ less than one equals $c_F(\text{aff}(A_x), \text{aff}_0(B_y))^2$.

Let $M_{\text{all}} = M_1 + \cdots + M_R$. Then $(ST^\top)^\top (ST^\top)$ is the $M_{\text{all}} \times M_{\text{all}}$ square matrix whose rows and columns are indexed by $(r, m)$ with $1 \leq r \leq R$ and $1 \leq m \leq M_r$ and whose entry corresponding to row $(r_1, m_1)$ and column $(r_2, m_2)$ equals

$$\frac{1}{R} \sqrt{|A_{r_1m_1}|} \frac{1_{A_{r_1m_1}}} |A_{r_2m_2}| = \frac{1}{R} \sqrt{|A_{r_1m_1}|} \frac{1_{A_{r_1m_1}}} |A_{r_2m_2}|$$

3.4 Bounding the Relevant Eigenvalues

It remains to bound the largest eigenvalue of $(ST^\top)^\top (ST^\top)$ that is less than one. To do so, we view the matrix as the symmetric normalized Laplacian of a particular weighted graph.

Let $G$ be the graph whose vertices are indexed by $(r, m)$ with $1 \leq r \leq R$ and $1 \leq m \leq M_r$. Let the edge between vertices $(r_1, m_1)$ and $(r_2, m_2)$ have weight $|A_{r_1m_1} \cap A_{r_2m_2}|$. We may assume that $G$ is connected (the analysis in this case subsumes the analysis in the general case). The symmetric normalized Laplacian $\mathcal{L}$ of this graph is closely related to our matrix of interest,

$$(ST^\top)^\top (ST^\top) = I - \frac{R-1}{R} \mathcal{L}. \quad (7)$$

Hence, the largest eigenvalue of $(ST^\top)^\top (ST^\top)$ that is less than one can be determined from the smallest nonzero eigenvalue $\lambda_2(\mathcal{L})$ of $\mathcal{L}$. We bound $\lambda_2(\mathcal{L})$ via Cheeger’s inequality (stated in Appendix A.3) by bounding the Cheeger constant $h_G$ of $G$.

Lemma 10. For $R \geq 2$, we have $h_G \geq \frac{2}{\sqrt{R}}$ and hence $\lambda_2(\mathcal{L}) \geq \frac{2}{\sqrt{R}}$.

We prove Lemma 10 in Appendix A.3. Combining Remark 9, Equation (7), and Lemma 10 we obtain the following bound on the Friedichs angle.

Proposition 11. Assuming that $R \geq 2$, we have

$$c_F(\text{aff}(A_x), \text{aff}_0(B_y))^2 \leq 1 - \frac{R-1}{R} - \frac{2}{\sqrt{R}} \leq 1 - \frac{1}{\sqrt{R}}.$$  

Together with Theorem 2 and Corollary 5, Proposition 11 implies the final bound on the rate.

Theorem 12. The AP algorithm for Problem (P4) converges linearly with rate $1 - \frac{1}{\sqrt{R}}$.

4 A Lower Bound

To prove the tightness of Theorem 12 we construct a submodular function and a decomposition that leads to a slow rate. We make use of Lemma 10 which describes the exact rate in the subspace case, and Corollary 20 which allows us to reduce our example to the subspace case. Both results are shown in Appendix B.1. For our worst-case example below, the empirical convergence rate matches the theoretical lower bound of Theorem 14 (Appendix B.3).

For each $x, y \in V$, define the submodular function $G_{xy}$ on $V$ to be the cut function on the graph with the single edge $(x, y)$

$$G_{xy} = \begin{cases} 1 & \text{if } |A \cap \{x, y\}| = 1 \\ 0 & \text{otherwise} \end{cases}$$
Now, let \( N \) be even and \( R \geq 2 \) and define the submodular function \( F^{lb} = F^{lb}_1 + \cdots + F^{lb}_R \), where
\[
F^{lb}_1 = G_{12} + G_{34} + \cdots + G_{(N-1)N} \quad F^{lb}_2 = G_{23} + G_{45} + \cdots + G_{N1}
\]
and \( F^{lb}_r = 0 \) for all \( r \geq 3 \). If we restrict the support of \( G_{xy} \) to \( \{x, y\} \), then \( B(G_{xy}) = \{(s, -s) \mid s \in [-1, 1]\} \). Therefore, the product \( B \) of base polytopes for this problem, which we refer to as \( B^{lb} \), is
\[
B^{lb} = \left\{ (s_1, -s_1, \ldots, s_N, -s_N, -t_N, t_N, -t_1, t_1, \ldots, t_N, 0, \ldots, 0, 0, \ldots, 0) \mid s_i, t_j \in [-1, 1] \right\}.
\]

In Figure 3, we illustrate several runs of AP between \( A \) and \( B^{lb} \).

**Lemma 13.** The Friedricks angle between \( A \) and \( \text{aff}(B^{lb}) \) satisfies
\[
\cos(\pi A, \text{aff}(B^{lb})) \geq 1 - \frac{1}{N} \left(1 - \cos\left(\frac{2\pi}{N}\right)\right).
\]

Lemma 13 proved in Appendix B.2 leads to the following lower bound on the worst-case rate.

**Theorem 14.** The worst-case convergence rate of AP between the polyhedra \( A \) and \( B \) from Lemma 13 is at least \( 1 - \frac{2\pi^2}{N^2 R} \).

**Proof.** From Corollary 21 and Lemma 13, we see that there is some nonzero \( a_0 \in A \) such that the sequences \( \{a_k\}_{k \geq 0} \) and \( \{b_k\}_{k \geq 0} \) generated by AP between \( A \) and \( B^{lb} \) satisfy
\[
\|a_k\| = (1 - \frac{1}{R}(1 - \cos\left(\frac{2\pi}{N}\right)))^k \|a_0\| \quad \text{and} \quad \|b_k\| = (1 - \frac{1}{R}(1 - \cos\left(\frac{2\pi}{N}\right)))^k \|b_0\|.
\]
Lastly, the inequality \( 1 - \cos(x) \leq \frac{1}{2}x^2 \) gives the lower bound. \( \square \)

## 5 Convergence of the Primal Objective

We have shown that AP produces a sequence of points \( \{a_k\}_{k \geq 0} \) and \( \{b_k\}_{k \geq 0} \) in \( \mathbb{R}^{NR} \) such that \( (a_k, b_k) \to (a_*, b_*) \) linearly, where \( (a_*, b_*) \) minimizes the objective in Problem P2. In this section, we show that this result also implies the linear convergence of the objective in Problem P3 and of the original discrete objective in Problem P1.

Define the matrix \( \Gamma = -R^{1/2}S \), where \( S \) is the matrix defined in Equation 13. Multiplication by \( \Gamma \) maps a vector \( (w_1, \ldots, w_N) \) to \( -\sum w_r \), where \( w_r \in \mathbb{R}^N \) for each \( r \). Set \( x_k = \Gamma b_k \) and \( x_* = \Gamma b_* \). As shown in Jegelka et al. [24], Problem P3 is minimized by \( x_* \).

**Proposition 15.** We have \( f(x_k) + \frac{1}{2}\|x_k\|^2 \to f(x_*) + \frac{1}{2}\|x_*\|^2 \) linearly with rate \( 1 - \frac{1}{N^2 R} \).

We prove Proposition 15 in Appendix C.1. This linear rate of convergence translates into the following linear rate for the original discrete problem. We prove Theorem 16 in Appendix C.2.

**Theorem 16.** Choose \( A_* \in \arg\min_{A \subseteq V} F(A) \). Let \( A_k \) be the suplevel set of \( x_k \) with smallest value of \( F \). Then \( F(A_k) \to F(A_*) \) linearly with rate \( 1 - \frac{1}{2N^2 R} \).

## 6 Discussion

In this work, we analyze projection methods for SFM and give upper and lower bounds on the linear rate of convergence. This means that the number of iterations required for an accuracy of \( \epsilon \) is logarithmic in \( 1/\epsilon \), not linear as in previous work [33]. Our rate is uniform over all submodular functions. Moreover, our proof highlights how the number \( R \) of components and the facial structure of \( B \) affect the convergence rate. These insights may serve as guidelines when working with projection algorithms and aid in the analysis.
of special cases. For example, reducing $R$ is often possible. Any collection of $F_r$ that have disjoint support, such as the cut functions corresponding to the rows or columns of a grid graph, can be grouped together as one component without making the projection harder.

Our analysis also shows the effects of additional properties of $F$. For example, suppose that $F$ is separable, that is, $F(V) = F(S) + F(V \setminus S)$ for some nonempty $S \subseteq V$. Then the subsets $A_{rm} \subseteq V$ defining the relevant faces of $B$ satisfy either $A_{rm} \subseteq S$ or $A_{rm} \subseteq S^c$. This makes $G$ in Section 3.4 disconnected, and as a result, the $N$ in Theorem 12 gets replaced by $\max\{|S|, |S^c|\}$ for an improved rate. This applies without the user needing to know $S$ when running the algorithm.

A number of future directions suggest themselves. We addressed AP, but Jegelka et al. [24] also considered the related Douglas-Rachford (DR) algorithm. DR between subspaces converges linearly with rate $c_F$, as opposed to $c_{F^2}$ for AP. We suspect that our approach may be modified to analyze DR between polyhedra. Further questions include the extension to multiple polyhedra.

Acknowledgments

We would like to thank Mădălina Persu for suggesting the use of Cheeger’s inequality. This research is supported in part by NSF CISE Expeditions Award CCF-1139158, LBNL Award 7076018, and DARPA XData Award FA8750-12-2-0331, and gifts from Amazon Web Services, Google, SAP, The Thomas and Stacey Siebel Foundation, Apple, C3Energy, Cisco, Cloudera, EMC, Ericsson, Facebook, GameOnTalis, Guavus, HP, Huawei, Intel, Microsoft, NetApp, Pivotal, Splunk, Virdata, VMware, WANdisco, and Yahoo!. This material is based upon work supported in part by the Office of Naval Research under grant number N00014-11-1-0688. This material is based upon work supported in part by the US ARL and the US ARO under contract/grant number W911NF-11-1-0391.

References

[1] F. Bach. Structured sparsity-inducing norms through submodular functions. In Advances in Neural Information Processing Systems, 2011.
[2] F. Bach. Learning with submodular functions: A convex optimization perspective. Foundations and Trends in Machine Learning, 6(2-3):145–373, 2013.
[3] H. H. Bauschke and J. M. Borwein. On the convergence of von Neumann’s alternating projection algorithm for two sets. Set-Valued Analysis, 1(2):185–212, 1993.
[4] H. H. Bauschke and J. M. Borwein. Dykstra’s alternating projection algorithm for two sets. Journal of Approximation Theory, 79(3):418–443, 1994.
[5] H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. SIAM Review, 38(3):367–426, 1996.
[6] H. H. Bauschke, J. Cruz, H. M. Phan, and X. Wang. The rate of linear convergence of the Douglas-Rachford algorithm for subspaces is the cosine of the Friedrichs angle. arXiv:1309.4709, 2013.
[7] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. SIAM Journal on Optimization, 23(4):2037–2060, 2013.
[8] J. V. Burke and J. J. Moré. On the identification of active constraints. SIAM Journal on Numerical Analysis, 25(5):1197–1211, 1988.
[9] F. R. Chung. Spectral Graph Theory. American Mathematical Society, 1997.
[10] F. Deutsch. The method of alternating orthogonal projections. In Approximation Theory, Spline Functions and Applications, pages 105–121. Springer, 1992.
[11] F. Deutsch. Best Approximation in Inner Product Spaces, volume 7. Springer, 2001.
[12] F. Deutsch and H. Hundal. The rate of convergence of Dykstra’s cyclic projections algorithm: The polyhedral case. Numerical Functional Analysis and Optimization, 15(5-6):537–565, 1994.
[13] F. Deutsch and H. Hundal. The rate of convergence for the cyclic projections algorithm ii: angles between convex sets. Journal of Approximation Theory, 142(1):36–55, 2006.
[14] P. Diaconis, K. Khare, and L. Saloff-Coste. Stochastic alternating projections. *Illinois Journal of Mathematics*, 54(3):963–979, 2010.

[15] J. Edmonds. *Combinatorial Structures and Their Applications*, chapter Submodular Functions, Matroids and Certain Polyhedra, pages 69–87. Gordon and Breach, 1970.

[16] A. Fix, T. Joachims, S. Park, and R. Zabih. Structured learning of sum-of-submodular higher order energy functions. In *Int. Conference on Computer Vision (ICCV)*, 2013.

[17] R. M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006.

[18] M. Grötschel, L. Lovász, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

[19] L. Gubin, B. Polyak, and E. Raik. The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, 7(6):1–24, 1967.

[20] I. Halperin. The product of projection operators. *Acta Sci. Math. (Szeged)*, 23:96–99, 1962.

[21] D. Hochbaum and V. Singh. An efficient algorithm for co-segmentation. In *Int. Conference on Computer Vision (ICCV)*, 2009.

[22] S. Iwata. A faster scaling algorithm for minimizing submodular functions. *SIAM J. on Computing*, 32:833–840, 2003.

[23] S. Jegelka, H. Lin, and J. Bilmes. On fast approximate submodular minimization. In *Advances in Neural Information Processing Systems*, 2011.

[24] S. Jegelka, F. Bach, and S. Sra. Reflection methods for user-friendly submodular optimization. In *Advances in Neural Information Processing Systems*, pages 1313–1321, 2013.

[25] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *JMLR*, page 22972334, 2011.

[26] A. V. Knyazev and M. E. Argentati. Principal angles between subspaces in an A-based scalar product: algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040, 2002.

[27] P. Kohli, L. Ladicky, and P. Torr. Robust higher order potentials for enforcing label consistency. *Int. Journal of Computer Vision*, 82, 2009.

[28] V. Kolmogorov. Minimizing a sum of submodular functions. *Discrete Applied Mathematics*, 160(15):2246–2258, 2012.

[29] N. Komodakis, N. Paragios, and G. Tziritas. MRF energy minimization and beyond via dual decomposition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2011.

[30] A. Krause and S. Jegelka. Submodularity in machine learning – new directions. *ICML* Tutorial, 2013.

[31] H. Lin and J. Bilmes. Optimal selection of limited vocabulary speech corpora. In *Proc. Interspeech*, 2011.

[32] S. McCormick. *Handbook on Discrete Optimization*, chapter Submodular Function Minimization, pages 321–391. Elsevier, 2006.

[33] M. Narasimhan and J. Bilmes. Local search for balanced submodular clusterings. In *IJCAI*, pages 981–986, 2007.

[34] J. Orlin. A faster strongly polynomial time algorithm for submodular function minimization. *Math. Programming*, 118:237–251, 2009.

[35] P. Stobbe and A. Krause. Efficient minimization of decomposable submodular functions. In *Advances in Neural Information Processing Systems*, 2010.

[36] P. Tseng. Alternating projection-proximal methods for convex programming and variational inequalities. *SIAM Journal on Optimization*, 7(4):951–965, 1997.

[37] S. Vicente, V. Kolmogorov, and C. Rother. Joint optimization of segmentation and appearance models. In *Int. Conference on Computer Vision (ICCV)*, 2009.

[38] J. Von Neumann. *Functional Operators: The Geometry of Orthogonal Spaces*. Princeton University Press, 1950.
A Upper Bound Results

A.1 Proof of Theorem 2

For the proof of this theorem, we will need the fact that projection maps are firmly nonexpansive, that is, for a closed convex nonempty subset $C \subseteq \mathbb{R}^D$, we have

$$\|\Pi_C x - \Pi_C y\|^2 + \|(x - \Pi_C x) - (y - \Pi_C y)\|^2 \leq \|x - y\|^2$$

for all $x, y \in \mathbb{R}^D$. Now, suppose that $\kappa_* < \infty$. Let $e = \Pi_E p_k$ and note that $v = \Pi_Q e - e$ and that $\Pi_Q e \in H$. We have

$$\kappa_*^{-2}d(p_k, E)^2 \leq d(p_k, Q')^2$$

$$\leq \|p_k - \Pi_Q p_k + v\|^2$$

$$\leq \|(p_k - \Pi_Q p_k) - (e - \Pi_Q e)\|^2$$

$$\leq \|p_k - e\|^2 - \|\Pi_Q p_k - \Pi_Q e\|^2$$

$$\leq d(p_k, E)^2 - d(q_k, H)^2.$$  

It follows that $d(q_k, H) \leq (1 - \kappa_*^{-2})^{1/2}d(p_k, E)$. Similarly, we have $d(p_{k+1}, E) \leq (1 - \kappa_*^{-2})^{1/2}d(q_k, H)$. Combining these and inducting, we see that

$$d(p_k, E) \leq (1 - \kappa_*^{-2})^kd(p_0, E)$$

$$d(q_k, H) \leq (1 - \kappa_*^{-2})^kd(q_0, H).$$

As shown in Bauschke and Borwein [3, Theorem 3.3], the above implies that $p_k \to p_\ast \in E$ and $q_k \to q_\ast \in H$ and that

$$\|p_k - p_\ast\| \leq 2\|p_0 - p_\ast\|(1 - \kappa_*^{-2})^k$$

$$\|q_k - q_\ast\| \leq 2\|q_0 - q_\ast\|(1 - \kappa_*^{-2})^k.$$  

A.2 Connection Between $\kappa$ and $c_F$ in the Subspace Case

In this section, we introduce a simple lemma connecting $\kappa$ and $c_F$ in the case of subspaces $U$ and $V$. We will use this lemma in several subsequent proofs.

Lemma 17. Let $U$ and $V$ be subspaces and suppose $u \in U \cap (U \cap V)^\perp$ and that $u \neq 0$. Then

(a) $\|\Pi_V u\| \leq c_F(U, V)\|u\|$  
(b) $\kappa(u) \leq (1 - c_F(U, V)^2)^{-1/2}$  
(c) $\kappa(u) = (1 - c_F(U, V)^2)^{-1/2}$ if and only if $\|\Pi_V u\| = c_F(U, V)\|u\|.$

Proof. Part [(a)] follows from the definition of $c_F$. Indeed,

$$c_F(U, V) \geq \frac{u^\top(\Pi_V u)}{\|u\|\|\Pi_V u\|} = \frac{\|\Pi_V u\|^2}{\|u\|\|\Pi_V u\|} = \frac{\|\Pi_V u\|}{\|u\|}.$$  

Part [(b)] follows from Part [(a)] and the observation that $\kappa(u) = (1 - \|\Pi_V u\|^2/\|u\|^2)^{-1/2}$. Part [(c)] follows from the same observation. \qed

A.3 Proof of Lemma 3

It suffices to prove the statement for $p \in P \setminus E$. For $p \in P \setminus E$, define $q = \Pi_Q p$, $e = \Pi_E q$, and $p'' = \Pi_{[p, e]}q$, where $[p, e]$ denotes the line segment between $p$ and $e$ (which is contained in $P$ by convexity). See Figure 2 for a graphical depiction. If $q \in E$, then $\kappa(p) = 1$. So we may assume that $q \notin E$ which also implies that $d(p'', E) > 0$ and $d(\Pi_{pq}, E) > 0$. We have

$$\kappa(p) = \frac{d(p, E)}{d(p, Q')} \leq \frac{\|p - e\|}{\|p - q\|} \leq \frac{\|q - e\|}{\|q - p''\|} \leq \frac{d(q, E)}{d(q, P')} = \kappa(q).$$  

(8)
Figure 2: Illustration of the proof of Lemma 3.

The first inequality holds because \( d(p, E) \leq \|p - e\| \) and \( d(p, Q') = \|p - q\| \). The middle inequality holds because the area of the triangle with vertices \( p, q, \) and \( e \) can be expressed as both \( \frac{1}{2} \|p - e\| \|q - p''\| \) and \( \frac{1}{2} \|p - q\| \|q - e\| \sin \theta \), where \( \theta \) is the angle between vectors \( p - q \) and \( e - q \), so

\[
\|p - e\| \|q - p''\| = \|p - q\| \|q - e\| \sin \theta \leq \|p - q\| \|q - e\|.
\]

The third inequality holds because \( \|q - e\| = d(q, E) \) and \( \|q - p''\| \geq d(q, P) \). The chain of inequalities in Equation (8) prove the lemma.

### A.4 Proof of Proposition 4

Suppose that \( p \in P \setminus E \) (the case \( q \in Q' \setminus E \) is the same), and let \( e = \Pi_{EP} p \). If \( \kappa(p) = 1 \), the statement is evident, so we may assume that \( \kappa(p) > 1 \). We will construct sequences of polyhedra

\[
P \supseteq P_1 \supseteq \cdots \supseteq P_j \\
Q' \supseteq Q'_1 \supseteq \cdots \supseteq Q'_j,
\]

where \( P_{j+1} \) is a face of \( P_j \) and \( Q'_{j+1} \) is a face of \( Q'_j \) for \( 1 \leq j \leq J - 1 \). Either \( \dim(\text{aff}(P_{j+1})) < \dim(\text{aff}(P_j)) \) or \( \dim(\text{aff}(Q'_{j+1})) < \dim(\text{aff}(Q'_j)) \) will hold. We will further define \( E_j = P_j \cap Q'_j \), which will contain \( e \), so that we can define

\[
\kappa_j(x) = \frac{d(x, E_j)}{\max\{d(x, P_j), d(x, Q'_j)\}}
\]

for \( x \in \mathbb{R}^D \setminus E_j \) (this is just the function \( \kappa \) defined for the polyhedra \( P_j \) and \( Q'_j \)). Our construction will yield points \( p_j \in P_j \), and \( q_j \in Q'_j \) such that \( p_j \in \text{relint}(P_j) \setminus E_j \), \( q_j \in \text{relint}(Q'_j) \setminus E_j \), and \( q_j = \Pi_{Q'_j} p_j \) for each \( j \). Furthermore, we will have

\[
\kappa(p) \leq \kappa_1(p_1) \leq \cdots \leq \kappa_J(p_J).
\]

Now we describe the construction. For any \( t \in [0, 1] \), define \( p^t = (1 - t)p + te \) to be the point obtained by moving \( p \) by the appropriate amount toward \( e \). Note that \( t \mapsto \kappa(p^t) \) is a nondecreasing function on the interval \([0, 1]\). Choose \( \varepsilon > 0 \) sufficiently small so that every face of either \( P \) or \( Q' \) that intersects \( B_\varepsilon(e) \), the ball of radius \( \varepsilon \) centered on \( e \), necessarily contains \( e \). Now choose \( 0 \leq t_0 < 1 \) sufficiently close to 1 so that \( \|p^{t_0} - e\| < \varepsilon \). It follows
that $e$ is contained in the face of $P$ whose relative interior contains $p^{1o}$. It further follows that $e$ is contained in the face of $Q'$ whose relative interior contains $\Pi Q^o p^{1o}$ because
\[
\|\Pi Q^o p^{1o} - e\| = \|\Pi Q^o p^{1o} - \Pi Q^o e\| \leq \|p^{1o} - e\| < \epsilon.
\]
To initialize the construction, set
\[
p_1 = p^{1o},
q_1 = \Pi Q^o p^{1o},
\]
and let $P_1$ and $Q'_1$ be the unique faces of $P$ and $Q'$ respectively such that $p_1 \in \text{relint}(P_1)$ and $q_1 \in \text{relint}(Q'_1)$ (the relative interiors of the faces of a polyhedron partition that polyhedron [8, Theorem 2.2]). Note that $q_1 \notin E$ because $\kappa(p_1) \geq \kappa(p) > 1$. Note that $e \in E_1 = P_1 \cap Q'_1$ so that
\[
\kappa(p) \leq \kappa(p_1) = \frac{d(p_1, E)}{d(p_1, Q')} = \frac{\|p_1 - e\|}{\|p_1 - q_1\|} = \frac{d(p_1, E_1)}{d(p_1, Q'_1)} = \kappa_1(p_1).
\]
Now, inductively assume that we have defined $P_j$, $Q'_j$, $p_j$, and $q_j$ satisfying the stated properties. Generate the sequences \{x_k\}_{k \geq 0} and \{y_k\}_{k \geq 0} with $x_k \in P_j$ and $y_k \in Q'_j$ by running AP between the polyhedra $P_j$ and $Q'_j$ initialized with $x_0 = p_j$. There are two possibilities, either $x_k \in \text{relint}(P_j)$ and $y_k \in \text{relint}(Q'_j)$ for every $k$, or there is some $k$ for which either $x_k \notin \text{relint}(P_j)$ or $y_k \notin \text{relint}(Q'_j)$. Note that AP will not terminate after a finite number of steps.

Suppose that $x_k \in \text{relint}(P_j)$ and $y_k \in \text{relint}(Q'_j)$ for every $k$. Then set $J = j$ and terminate the procedure. Otherwise, choose $k'$ such that either $x_{k'} \notin \text{relint}(P_j)$ or $y_{k'} \notin \text{relint}(Q'_j)$. Now set $p_{j+1} = x_{k'}$ and $q_{j+1} = y_{k'}$. Let $P_{j+1}$ and $Q'_{j+1}$ be the unique faces of $P_j$ and $Q'_j$ respectively such that $p_{j+1} \in \text{relint}(P_{j+1})$ and $q_{j+1} \in \text{relint}(Q'_{j+1})$. Note that $p_{j+1}, q_{j+1} \notin E_{j+1} = P_{j+1} \cap Q'_{j+1}$ and $e \in E_{j+1}$. We have
\[
\kappa_j(p_j) < \kappa_j(p_{j+1}) = \frac{d(p_{j+1}, E_j)}{d(p_{j+1}, Q'_j)} = \frac{d(p_{j+1}, E_j)}{\|p_{j+1} - q_{j+1}\|} \leq \frac{d(p_{j+1}, E_{j+1})}{d(p_{j+1}, Q'_{j+1})} = \kappa_{j+1}(p_{j+1}).
\]
The preceding work shows the inductive step. Note that if $P_{j+1} \neq P_j$ then $\dim(\text{aff}(P_{j+1})) < \dim(\text{aff}(P_j))$ and if $Q'_{j+1} \neq Q'_j$ then $\dim(\text{aff}(Q'_{j+1})) < \dim(\text{aff}(Q'_j))$. One of these will hold, so the induction will terminate after a finite number of steps.

We have produced the sequence in Equation (9) and we have created $p_j$, $P_j$, and $Q'_j$ such that AP between $P_j$ and $Q'_j$, when initialized at $p_j$, generates the same sequence of points as AP between $\text{aff}(P_j)$ and $\text{aff}(Q'_j)$. Using this fact, along with Deutsch and Hundal [12, Theorem 9.3], we see that $\Pi_{\text{aff}(P_j)} \cap \text{aff}(Q'_j) p_j \in E_j$. Using this, along with Lemma [14][1b], we see that
\[
\kappa_j(p_j) \leq (1 - c_F(\text{aff}_0(P_j), \text{aff}_0(Q'_j))^2)^{-1/2}.
\]
Equations (10) and (9) prove the result. Note that $P_j$ and $Q'_j$ are faces of $P$ and $Q'$ respectively. We can switch between faces of $Q'$ and faces of $Q$ because doing so amounts to translating by $v$ which does not affect the angles.

### A.5 Proof of Lemma [6]

We have
\[
c_F(\text{null}(S), \text{null}(T)) = c_F(\text{range}(S^T)^\perp, \text{range}(T^T)^\perp)
= c_F(\text{range}(S^T), \text{range}(T^T)),
\]
where the first equality uses the fact that $\text{null}(W) = \text{range}(W^T)^\perp$ for matrices $W$, and the second equality uses the fact that $c_F(U^\perp, V^\perp) = c_F(U, V)$ for subspaces $U$ and $V$ [6, Fact 2.3].

Let $S^T$ and $T^T$ have dimensions $D \times J$ and $D \times K$ respectively, and let $X$ and $Y$ be the subspaces spanned by the columns of $S^T$ and $T^T$ respectively. Without loss of generality,
assume that \( J \leq K \). Let \( \sigma_1 \geq \cdots \geq \sigma_J \) be the singular values of \( ST^T \) with corresponding left singular vectors \( u_1, \ldots, u_J \) and right singular vectors \( v_1, \ldots, v_J \). Let \( x_j = S^T u_j \) and let \( y_j = T^T v_j \) for \( 1 \leq j \leq J \). By definition, we can write

\[
\sigma_j = \max_{u,v} \{ u^T ST^T v \mid u \perp \text{span}(u_1, \ldots, u_{j-1}), v \perp \text{span}(v_1, \ldots, v_{j-1}), \|u\| = 1, \|v\| = 1 \}.
\]

Since the \( \{ u_j \} \) are orthonormal, so are the \( \{ x_j \} \). Similarly, since the \( \{ v_j \} \) are orthonormal, so are the \( \{ y_j \} \). Suppose that all of the singular values of \( ST^T \) equal one. Then we must have \( x_j = y_j \) for each \( j \), which implies that \( X \subseteq Y \), and so \( c_F(X, Y) = 0 \).

Now suppose that \( \sigma_1 = \cdots = \sigma_{\ell} = 1 \), and \( \sigma_{\ell+1} \neq 1 \). It follows that

\[
X \cap Y = \text{span}(x_1, \ldots, x_\ell) = \text{span}(y_1, \ldots, y_\ell),
\]

and so

\[
\sigma_{\ell+1} = \sup_{u,v} \{ u^T ST^T v \mid u \in \text{span}(u_1, \ldots, u_\ell)^\perp, v \in \text{span}(v_1, \ldots, v_\ell)^\perp, \|u\| = 1, \|v\| = 1 \}
\]

\[
= \sup_{x,y} \{ x^T y \mid x \in X \cap (X \cap Y)^\perp, y \in Y \cap (X \cap Y)^\perp, \|x\| = 1, \|y\| = 1 \}
\]

\[
= c_F(X, Y).
\]

### A.6 Cheeger’s Inequality

For an overview of spectral graph theory, see Chung [9]. We state Cheeger’s inequality below.

Let \( G \) be a weighted, connected graph with vertex set \( V_G \) and edge weights \((w_{ij}), i,j \in V_G\). Define the weighted degree of a vertex \( i \) to be \( \delta_i = \sum_{j \neq i} w_{ij} \), define the volume of a subset of vertices to be the sum of their weighted degrees, \( \text{vol}(U) = \sum_{i \in U} \delta_i \), and define the size of the cut between \( U \) and its complement \( U^c \) to be the sum of the weights of the edges between \( U \) and \( U^c \),

\[
|E(U, U^c)| = \sum_{i \in U, j \in U^c} w_{ij}.
\]

The Cheeger constant is defined as

\[
h_G = \min_{\emptyset \neq U \subseteq V_G} \frac{|E(U, U^c)|}{\text{min}(\text{vol}(U), \text{vol}(U^c))}.
\]

Let \( L \) be the unnormalized Laplacian of \( G \), i.e. the \( |V_G| \times |V_G| \) matrix whose entries are defined by

\[
L_{ij} = \begin{cases} 
-w_{ij} & \text{if } i \neq j \\
\delta_i & \text{otherwise}
\end{cases}
\]

Let \( D \) be the \( |V_G| \times |V_G| \) diagonal matrix defined by \( D_{ii} = \delta_i \). Then \( \mathcal{L} = D^{-1/2}LD^{-1/2} \) is the normalized Laplacian. Let \( \lambda_2(\mathcal{L}) \) denote the second smallest eigenvalue of \( \mathcal{L} \) (since \( G \) is connected, there will be exactly one eigenvalue equal to zero).

**Theorem 18** (Cheeger’s inequality). We have \( \lambda_2(\mathcal{L}) \geq \frac{h_G^2}{2} \).
A.7 Proof of Lemma 10

Proof. We have
\[
\min(\text{vol}(U), \text{vol}(U^c)) \leq \frac{1}{2} \text{vol}(V_G)
\]
\[
= \frac{1}{2} \sum_{(r,m)} \left( \sum_{(r',m') \neq (r,m)} |A_{rm} \cap A_{r'm'}| \right)
\]
\[
= \frac{1}{2} \sum_{(r,m)} (R - 1)|A_{rm}|
\]
\[
= \frac{1}{2} NR(R - 1).
\]

Since \( G \) is connected, for any nonempty set \( U \subseteq V_G \), there must be some element \( v \in V \) (here \( V \) is the ground set of our submodular function \( F \), not the set of vertices \( V_G \)) such that \( v \in A_{r_1m_1} \cap A_{r_2m_2} \) for some \((r_1,m_1) \in U \) and \((r_2,m_2) \in U^c \). Suppose that \( v \) appears in \( k \) of the subsets of \( V \) indexed by elements of \( U \) and in \( R - k \) of the subsets of \( V \) indexed by elements of \( U^c \). Then
\[
|E(U,U^c)| \geq k(R - k) \geq R - 1.
\]

It follows that
\[
h_G \geq \frac{R - 1}{2NR(R - 1)} = \frac{2}{NR}.
\]

It follows from Theorem 18 that \( \lambda_2(\mathcal{L}) \geq \frac{2}{NR^2} \).

B Results for the Lower Bound

B.1 Some Helpful Results

In Lemma 19, we show how AP between subspaces \( U \) and \( V \) can be initialized to exactly achieve the worst-case rate of convergence. Then in Corollary 20 we show that if subsets \( U' \) and \( V' \) look like subspaces \( U \) and \( V \) near the origin, we can initialize AP between \( U' \) and \( V' \) to achieve the same worst-case rate of convergence.

Lemma 19. Let \( U \) and \( V \) be subspaces with \( U \not\subseteq V \) and \( V \not\subseteq U \). Then there exists some nonzero point \( u_0 \in U \cap (U \cap V)^\perp \) such that when we initialize AP at \( u_0 \), the resulting sequences \( \{u_k\}_{k \geq 0} \) and \( \{v_k\}_{k \geq 0} \) satisfy
\[
\|u_k\| = c_F(U,V)^{2k}\|u_0\|
n\|v_k\| = c_F(U,V)^{2k}\|v_0\|.
\]

Proof. Find \( u_* \in U \cap (U \cap V)^\perp \) and \( v_* \in V \cap (U \cap V)^\perp \) with \( \|u_*\| = 1 \) and \( \|v_*\| = 1 \) such that \( u_0^\top v_* = c_F(U,V) \), which we can do by compactness. By Lemma 14(a)
\[
c_F(U,V) = \Pi_V u_* = \Pi_V u_* \leq \Pi_V u_* \leq c_F(U,V).
\]

Set \( u_0 = u_* \) and generate the sequences \( \{u_k\}_{k \geq 0} \) and \( \{v_k\}_{k \geq 0} \) via \( \text{AP} \). Since \( \|\Pi_V u_0\| = c_F(U,V) \), Lemma 14(c) implies that \( \kappa(u_0) = (1 - c_F(U,V)^2)^{-1/2} \). Since \( \kappa \) attains its maximum at \( u_0 \), Lemma 3 implies that \( \kappa \) attains the same value at every element of the sequences \( \{u_k\}_{k \geq 0} \) and \( \{v_k\}_{k \geq 0} \). Therefore, Lemma 14(c) implies that \( \|\Pi_V u_k\| = c_F(U,V)\|u_k\| \) and \( \|\Pi_U v_k\| = c_F(U,V)\|v_k\| \) for all \( k \). This proves the lemma.
**Corollary 20.** let $U$ and $V$ be subspaces with $U \not\subseteq V$ and $V \not\subseteq U$. Let $U' \subseteq U$ and $V' \subseteq V$ be subsets such that $U' \cap B_1(0) = U \cap B_1(0)$ and $V' \cap B_1(0) = V \cap B_1(0)$ for some $\epsilon > 0$. Then there is a point $u'_0 \in U'$ such that the sequences $\{u'_k\}_{k \geq 0}$ and $\{v'_k\}_{k \geq 0}$ generated by $A P$ between $U'$ and $V'$ satisfy

\[
\|u'_k\| = c_F(U, V)^{2k}\|u'_0\| \\
\|v'_k\| = c_F(U, V)^{2k}\|v'_0\|.
\]

**Proof.** Use Lemma 19 to choose some nonzero $u_0 \in U \cap (U \cap V) \perp$ satisfying this property. Then set $u'_0 = \frac{u_0}{\|u_0\|}u_0$.

**B.2 Proof of Lemma 13**

Observe that we can write

$\text{aff}(B^l) = \{(s_1, -s_1, \ldots, s_T, -s_T, t_1, -t_1, \ldots, t_T, 0, \ldots, 0, \ldots, 0) \mid s_i, t_j \in \mathbb{R}\}$.

We can write $\text{aff}(B^l)$ as the nullspace of the matrix

$T_{lb} = \begin{pmatrix} T_{lb,1} & T_{lb,2} \\ I_N & \ddots & I_N \end{pmatrix},$

where the $N \times N$ identity matrix $I_N$ is repeated $R - 2$ times and where $T_{lb,1}$ and $T_{lb,2}$ are the $N^2 \times N$ matrices

$T_{lb,1} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 1 & \cdots \\ 1 & 1 \\ \vdots \\ 1 & 1 \end{pmatrix}, \\
T_{lb,2} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 1 & \cdots \\ 1 & 1 \end{pmatrix}.$

Recall that we can write $A$ as the nullspace of the matrix $S$ defined in Equation (5). It follows from Lemma 13 that $c_F(A, \text{aff}(B^l))$ equals the largest singular value of $ST_{lb}^\top$ that is less than one. We have

$ST_{lb}^\top = \frac{1}{\sqrt{R}} \begin{pmatrix} T_{lb,1}^\top & T_{lb,2}^\top & I_N & \cdots & I_N \end{pmatrix}$.

We can permute the columns of $ST_{lb}^\top$ without changing the singular values, so $c_F(A, \text{aff}(B^l))$ equals the largest singular value of

$\frac{1}{\sqrt{R}} \begin{pmatrix} T_{lb,0}^\top & I_N & \cdots & I_N \end{pmatrix}$,

that is less than one, where $T_{lb,0}$ is the $N \times N$ circulant matrix

$T_{lb,0} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 1 & \cdots \\ 1 & 1 & 1 & \cdots \\ \vdots & \vdots & \ddots & \ddots \\ 1 & 1 & 1 & \cdots \end{pmatrix}.$

Therefore, $c_F(A, \text{aff}(B^l))^2$ equals the largest eigenvalue of

$\frac{1}{R} \begin{pmatrix} T_{lb,0}^\top & I_N & \cdots & I_N \end{pmatrix} \begin{pmatrix} T_{lb,0}^\top & I_N & \cdots & I_N \end{pmatrix}^\top = \frac{1}{R} (T_{lb,0}^\top T_{lb,0} + (R - 2)I_N)$

that is less than one. Therefore, it suffices to examine the $N \times N$ circulant matrix

$T_{lb,0}^\top T_{lb,0} = \frac{1}{2} \begin{pmatrix} 2 & 1 & 1 & \cdots \\ 1 & 2 & 1 & \cdots \\ \vdots & \vdots & \ddots & \ddots \\ 1 & 1 & 2 & \cdots \end{pmatrix}.$
Successive ratios in lower bound example

Figure 3: We run five trials of AP between \( A \) and \( B^{lb} \) with random initializations, where \( N = 10 \) and \( R = 10 \). For each trial, we plot the ratios \( d(a_{k+1}, E)/d(a_k, E) \), where \( E = A \cap B^{lb} \) is the optimal set. The red line shows the theoretical lower bound of \( 1 - \frac{1}{R}(1 - \cos(\frac{2\pi}{N})) \) on the worst-case rate of convergence.

The eigenvalues of \( T_{lb,0}^\top T_{lb,0} \) are given by \( \lambda_j = 1 + \cos(\frac{2\pi j}{N}) \) for \( 0 \leq j \leq N - 1 \) (see Gray [17, Section 3.1] for a derivation). Therefore,

\[
c_F(A, \text{aff}(B^{lb}))^2 = 1 - \frac{1}{R}(1 - \cos(\frac{2\pi}{N})).
\]

**B.3 Lower Bound Illustration**

The proof of Theorem 14 shows that there is some \( a_0 \in A \) such that when we initialize AP between \( A \) and \( B^{lb} \) at \( a_0 \), we generate a sequence \( \{a_k\}_{k \geq 0} \) satisfying

\[
d(a_k, E) = (1 - \frac{1}{R}(1 - \cos(\frac{2\pi}{N})))^k d(a_0, E),
\]

where \( E = A \cap B^{lb} \) is the optimal set. In Figure 3, we plot the theoretical bound in red, and in blue the successive ratios \( d(a_{k+1}, E)/d(a_k, E) \) for five runs of AP between \( A \) and \( B^{lb} \) with random initializations. Had we initialized AP at \( a_0 \), the successive ratios would exactly equal \( 1 - \frac{1}{R}(1 - \cos(\frac{2\pi}{N})) \). The plot of these ratios would coincide with the red line in Figure 3.

Figure 3 illustrates that the empirical behavior of AP between \( A \) and \( B^{lb} \) is often similar to the worst-case behavior, even when the initialization is random. When we initialize AP randomly, the successive ratios appear to increase to the lower bound and then remain constant. Figure 3 shows the case \( N = 10 \) and \( R = 10 \), but the plot looks similar for other \( N \) and \( R \).

We also note that the graph corresponding to our lower bound example actually achieves a Cheeger constant similar to the one used in Lemma 10.
C Results for Convergence of the Primal and Discrete Problems

C.1 Proof of Proposition 15

First, suppose that \( s \in B(F) \). Let \( A = \{n \in V \mid s_n \geq 0\} \) be the set of indices on which \( s \) is nonnegative. Then we have

\[
\|s\| \leq \|s\|_1 = 2s(A) - s(V) \leq 3F_{\text{max}}. \tag{11}
\]

Now, we show that \( f(x_k) + \frac{1}{2}\|x_k\|^2 \) converges to \( f(x_*) + \frac{1}{2}\|x_*\|^2 \) linearly with rate \( 1 - \frac{1}{NR^2} \).

We will use Equation (11) to bound the norms of \( x_k \) and \( x_* \), both of which lie in \(-B(F)\).

We will also use the fact that \( \|x_k - x_*\| \leq \|\mathcal{L}\|\|b_k - b_*\| \leq \sqrt{R}\|b_k - b_*\| \).

Finally, we will use the proof of Theorem 12 to bound \( b_k - b_* \). First, we bound the difference between the squared norms using convexity. We have

\[
\frac{1}{2}\|x_k\|^2 - \frac{1}{2}\|x_*\|^2 \leq x_k^T(x_k - x_*) \leq \|x_k\||x_k - x_*\| \leq 3F_{\text{max}}\sqrt{R}\|b_k - b_*\| \leq 6F_{\text{max}}\sqrt{R}\|b_0 - b_*\|(1 - \frac{1}{NR^2})^k. \tag{12}
\]

Next, we bound the difference in Lovász extensions. Choose \( s \in \arg \max_{s \in B(F)} s^T x_k \). Then

\[
f(x_k) - f(x_*) \leq s^T(x_k - x_*) \leq \|s\||x_k - x_*\| \leq 3F_{\text{max}}\sqrt{R}\|b_k - b_*\| \leq 6F_{\text{max}}\sqrt{R}\|b_0 - b_*\|(1 - \frac{1}{NR^2})^k. \tag{13}
\]

Combining the bounds (12) and (13), we find that

\[
(f(x_k) + \frac{1}{2}\|x_k\|^2) - (f(x_*) + \frac{1}{2}\|x_*\|^2) \leq 12F_{\text{max}}\sqrt{R}\|b_0 - b_*\|(1 - \frac{1}{NR^2})^k. \tag{14}
\]

C.2 Proof of Theorem 16

By definition, \( A_k \) is the set of the form \( \{n \in V \mid (x_k)_n \geq c\} \) for some constant \( c \) with smallest value of \( F(\{n \in V \mid (x_k)_n \geq c\}) \).

Let \((w_*, s_*) \in \mathbb{R}^N \times B(F)\) be a primal-dual optimal pair for the left-hand version of Problem (P3). The dual of this minimization problem is the projection problem \( \min_{s \in B(F)} \frac{1}{2}\|s\|^2 \).

From [2, Proposition 10.5], we see that

\[
F(A_k) - F(A_*) \leq F(A_k) - (s_*)^T(V)
\leq \sqrt{\frac{N}{2}}(f(x_k) + \frac{1}{2}\|x_k\|^2 - (f(x_*) + \frac{1}{2}\|x_*\|^2))
\leq \sqrt{6F_{\text{max}}NR^1/2}\|b_0 - b_*\|(1 - \frac{1}{NR^2})^{k/2}
\leq \sqrt{6F_{\text{max}}NR^1/2}\|b_0 - b_*\|(1 - \frac{1}{2NR^2})^k,
\]

where the third inequality uses the proof of Proposition 15. The second inequality relies on Bach [2, Proposition 10.5], which states that a duality gap of \( \epsilon \) for the left-hand version of Problem (P3) turns into a duality gap of \( \sqrt{N}\epsilon/2 \) for the original discrete problem. If our algorithm converged with rate \( \frac{1}{2} \), this would translate to a rate of \( \frac{1}{\sqrt{k}} \) for the discrete problem. But fortunately, our algorithm converges linearly, and taking a square root preserves linear convergence.
Running times. Theorem 16 implies that the number of iterations required for an accuracy of $\epsilon$ is at most

$$2N^2R^2 \log \left( \frac{\sqrt{6F_{\max N}R^{1/2}b_0 - b_*}}{\epsilon} \right).$$

Each iteration involves minimizing each of the $F_r$ separately.