On the Convergence Rate of Decomposable Submodular Function Minimization

Robert Nishihara, Stefanie Jegelka, Michael I. Jordan
Electrical Engineering and Computer Science
University of California
Berkeley, CA 94720
{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]. A set function $F\colon 2^V \to \mathbb{R}$ 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 [28], 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 can be inefficient on large-scale problems. For this reason, the development of scalable, parallelizable algorithms has been an active area of research [24, 25, 29, 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 occuring submodular functions (for example, graph cuts, hypergraph cuts, MAP inference with higher-order potentials [16, 28, 37], co-segmentation [22], certain structured-sparsity inducing functions [26], covering functions [35], and combinations thereof) can be expressed as a sum

$$F(A) = \sum\nolimits_{r=1}^{R} F_r(A) \tag{1}$$

of simple submodular functions. Recent work demonstrates that this structure offers important practical benefits [25, 29, 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 [25]. 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. [25] 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 [19]. Strongly and weakly polynomial-time combinatorial algorithms followed [32]. The current fastest running times are $O(N^5\tau_1 + N^6)$ [34] in general and $O((N^4\tau_1 + N^5)\log F_{\max})$ for integer-valued functions [23], where $F_{\max} = \max_A |F(A)|$ and τ_1 is the time required to evaluate F. Some work has addressed decomposable functions [25, 29, 35]. The running times in [29] apply to integer-valued functions and range from $O((N+R)^2 \log F_{\max})$ for cuts to $O((N+Q^2R)(N+Q^2R+QR\tau_2)\log F_{\max})$, where $Q \leq N$ is the maximal cardinality of the support of any F_r, and τ_2 is the time required to minimize a simple function. Stobbe and Krause [35] 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, 7, 11, 12, 20, 21, 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 [5, 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 κ_* (defined in Section 3.1). However, except in very limited cases, it is unclear how to compute or even bound κ_*. While it is known that $\kappa_* < \infty$ for polyhedra [5, 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 κ_* 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 κ_* along with the computability of c_F. To our knowledge, we are the first to relate κ_* 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 $s(A) = \sum_{n \in A} s_n$. The *base polytope* of F is defined as the set of all modular functions that are dominated by F and that sum to $F(V)$,

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

The *Lovász extension* $f\colon \mathbb{R}^N \to \mathbb{R}$ 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 (P1) can be relaxed to the non-smooth convex optimization problem

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

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 [30] and smoothing techniques [35], but suffers from the non-smoothness of f [25]. Alternatively, we can formulate a proximal version of the problem

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

By thresholding the optimal solution of Problem (P3) at zero, we recover the indicator vector of an optimal discrete solution [17], [2, Proposition 8.4].

Lemma 1. *[25] The dual of the right-hand side of Problem* (P3) *is the best-approximation problem*

$$\min \ \|a-b\|^2 \quad a\in\mathcal{A}, \ \ b\in\mathcal{B}, \tag{P4}$$

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

Lemma 1 implies that we can minimize a decomposable submodular function by solving Problem (P4), which means finding the closest points between the subspace $\mathcal{A}$ and the product $\mathcal{B}$ of base polytopes. Projecting onto $\mathcal{A}$ is straightforward because $\mathcal{A}$ is a subspace. Projecting onto $\mathcal{B}$ 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 $F_r - z$ [25]. We can compute these projections easily because each F_r is simple.

Throughout this paper, we use $\mathcal{A}$ and $\mathcal{B}$ to refer to the specific polyhedra defined in Lemma 1 (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 $\mathcal{B}$ depends on the submodular functions $F_1,\dots,F_R$, 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 $a_0\in\mathcal{A}$, and inductively generate two sequences via $b_k = \Pi_{\mathcal{B}} a_k$ and $a_{k+1} = \Pi_{\mathcal{A}} b_k$. Given the nature of $\mathcal{A}$ and $\mathcal{B}$, this algorithm is easy to implement and use in our setting, and it solves Problem (P4) [25]. 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 $\|a_k - a_*\| \le C_1\alpha^k$ and $\|b_k - b_*\| \le C_2\alpha^k$ for all k and for some constants C_1 and C_2. Smaller values of α are better.

Analysis: Intuition. We will provide a detailed analysis of the convergence of AP for the polyhedra $\mathcal{A}$ and $\mathcal{B}$. 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

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

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

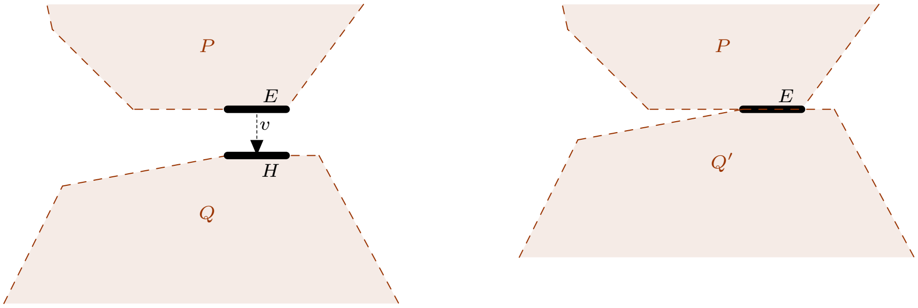

Figure 1: The optimal sets E, H in Equation (4), the vector v, and the shifted polyhedron Q'.

faces of P and Q and showing that the convergence rate of AP between P and Q is at worst $\max_{x,y} c_F(\text{aff}_0(P_x), \text{aff}_0(Q_y))^2$, where $\text{aff}(C)$ is the affine hull of C and $\text{aff}_0(C) = \text{aff}(C) - c$ for some $c \in C$. The faces $\{P_x\}_{x \in \mathbb{R}^D}$ of P are defined as the nonempty maximizers of linear functions over P, that is

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

While we look at angles between pairs of faces, we remark that Deutsch and Hundal [13] 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 polyhedra $\mathcal{A}$ and $\mathcal{B}$, 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 $1 - \frac{1}{N^2R^2}$ on the rate, stated in Theorem 12.

In Theorem 14, we show a lower bound of $1 - \frac{2\pi^2}{N^2 R}$ 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 $\mathcal{A}$ and $\mathcal{B}$. The results in this section are proved in Appendix A.

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

$$E = \{p \in P \,|\, d(p, Q) = d(P, Q)\} \qquad H = \{q \in Q \,|\, d(q, P) = d(Q, P)\}, \tag{4}$$

and let $v = \Pi_{Q-P} 0$ (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 $(e, e + v) \in E \times H$ are solutions to the best-approximation problem between P and Q. In our analysis, we will mostly study the translated version $Q' = Q - v$ of Q that intersects P at E.

For $x \in \mathbb{R}^D \backslash E$, the function κ 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 κ 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 κ is unbounded. The maximum $\kappa_* = \sup_{x \in (P \cup Q') \backslash E} \kappa(x)$ is useful for bounding 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 $1 - \frac{1}{\kappa_*^2}$. Specifically,*

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

3.2 Relating κ_* to the Angles Between Faces of the Polyhedra

In this section, we consider the case of polyhedra P and Q, and we bound κ_* in terms of the angles between pairs of their faces. In Lemma 3, we show that κ 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 those 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 κ at some later point in the resulting sequence.

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

We can now bound κ by angles between faces of P and Q.

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

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

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

Note that $\mathrm{aff}_0(Q_y) = \mathrm{aff}_0(Q'_y)$. Proposition 4 immediately gives us the following corollary.

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

$$1 - \frac{1}{\kappa_*^2} \leq \max_{x, y \in \mathbb{R}^D} c_F(\mathrm{aff}_0(P_x), \mathrm{aff}_0(Q_y))^2.$$

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.5, though this result is implicit in the characterization of principal angles between subspaces given in [27, 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 $c_F(\mathrm{null}(S), \mathrm{null}(T)) = 0$. Otherwise, $c_F(\mathrm{null}(S), \mathrm{null}(T))$ is equal to the largest singular value of $ST^\top$ that is less than one.*

Faces of relevant polyhedra. Let $\mathcal{A}_x$ and $\mathcal{B}_y$ be faces of the polyhedra $\mathcal{A}$ and $\mathcal{B}$ from Lemma 1. Since $\mathcal{A}$ is a vector space, its only nonempty face is $\mathcal{A}_x = \mathcal{A}$. Hence, $\mathcal{A}_x = \mathrm{null}(S)$, where S is an $N \times NR$ matrix of $N \times N$ identity matrices I_N:

$$S = \frac{1}{\sqrt{R}} \Big(\underbrace{I_N \quad \cdots \quad I_N}_{\text{repeated } R \text{ times}} \Big). \tag{5}$$

The matrix for $\mathrm{aff}_0(\mathcal{B}_y)$ requires a bit more elaboration. Since $\mathcal{B}$ is a Cartesian product, we have $\mathcal{B}_y = B(F_1)_{y_1} \times \cdots \times B(F_R)_{y_R}$, where $y = (y_1, \ldots, y_R)$ and $B(F_r)_{y_r}$ 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*

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

The following corollary is immediate.

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

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

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

$$\operatorname{aff}_0(\mathcal{B}_y) = \bigcap_{r=1}^{R} \bigcap_{m=1}^{M_r} \{(s_1, \ldots, s_R) \in \mathbb{R}^{NR} \mid s_r(A_{r1} \cup \cdots \cup A_{rm}) = 0\}. \tag{6}$$

In other words, we can write $\operatorname{aff}_0(\mathcal{B}_y)$ as the nullspace of either of the matrices

$$T' = \begin{pmatrix} 1_{A_{11}}^\top & & \\ \vdots & & \\ 1_{A_{11}\cup\cdots\cup A_{1M_1}}^\top & & \\ & \ddots & \\ & & 1_{A_{R1}}^\top \\ & & \vdots \\ & & 1_{A_{R1}\cup\cdots\cup A_{RM_R}}^\top \end{pmatrix} \text{ or } T = \begin{pmatrix} \frac{1_{A_{11}}^\top}{\sqrt{|A_{11}|}} & & \\ \vdots & & \\ \frac{1_{A_{1M_1}}^\top}{\sqrt{|A_{1M_1}|}} & & \\ & \ddots & \\ & & \frac{1_{A_{R1}}^\top}{\sqrt{|A_{R1}|}} \\ & & \vdots \\ & & \frac{1_{A_{RM_R}}^\top}{\sqrt{|A_{RM_R}|}} \end{pmatrix},$$

where 1_A is the indicator vector of $A \subseteq V$. For T', this follows directly from Equation (6). T can be obtained from T' via left multiplication by an invertible matrix, so T and T' have the same nullspace. Lemma 6 then implies that $c_F(\operatorname{aff}_0(\mathcal{A}_x), \operatorname{aff}_0(\mathcal{B}_y))$ equals the largest singular value of

$$ST^\top = \frac{1}{\sqrt{R}} \begin{pmatrix} \frac{1_{A_{11}}}{\sqrt{|A_{11}|}} & \cdots & \frac{1_{A_{1M_1}}}{\sqrt{|A_{1M_1}|}} & \cdots & \frac{1_{A_{R1}}}{\sqrt{|A_{R1}|}} & \cdots & \frac{1_{A_{RM_R}}}{\sqrt{|A_{RM_R}|}} \end{pmatrix}$$

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 $c_F(\operatorname{aff}_0(\mathcal{A}_x), \operatorname{aff}_0(\mathcal{B}_y))^2$.*

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

$$\frac{1}{R} \frac{1_{A_{r_1 m_1}}^\top 1_{A_{r_2 m_2}}}{\sqrt{|A_{r_1 m_1}||A_{r_2 m_2}|}} = \frac{1}{R} \frac{|A_{r_1 m_1} \cap A_{r_2 m_2}|}{\sqrt{|A_{r_1 m_1}||A_{r_2 m_2}|}}.$$

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 in terms of the symmetric normalized Laplacian of a 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 $|A_{r_1 m_1} \cap A_{r_2 m_2}|$. 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,

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

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.6) by bounding the Cheeger constant h_G of G.

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

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

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

$$c_F(\mathrm{aff}_0(\mathcal{A}_x), \mathrm{aff}_0(\mathcal{B}_y))^2 \leq 1 - \tfrac{R-1}{R} \tfrac{2}{N^2 R^2} \leq 1 - \tfrac{1}{N^2 R^2}.$$

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 $1 - \frac{1}{N^2 R^2}$, i.e.,*

$$\|a_k - a_*\| \leq 2\|a_0 - a_*\|(1 - \tfrac{1}{N^2 R^2})^k \quad \textit{and} \quad \|b_k - b_*\| \leq 2\|b_0 - b_*\|(1 - \tfrac{1}{N^2 R^2})^k.$$

4 A Lower Bound

To probe the tightness of Theorem 12, we construct a "bad" submodular function and decomposition that lead to a slow rate. Appendix B gives the formal details. Our example is an augmented cut function on a cycle: for each $x, y \in V$, define G_{xy} to be the cut function of a single edge (x, y),

$$G_{xy} = \begin{cases} 1 & \text{if } |A \cap \{x, y\}| = 1 \\ 0 & \text{otherwise} . \end{cases}$$

Take N to be even and $R \geq 2$ and define the submodular function $F^{\text{lb}} = F^{\text{lb}}_1 + \cdots + F^{\text{lb}}_R$, where

$$F^{\text{lb}}_1 = G_{12} + G_{34} + \cdots + G_{(N-1)N} \qquad F^{\text{lb}}_2 = G_{23} + G_{45} + \cdots + G_{N1}$$

and $F^{\text{lb}}_r = 0$ for all $r \geq 3$. The optimal solution to the best-approximation problem is the all zeros vector.

Lemma 13. *The cosine of the Friedrichs angle between $\mathcal{A}$ and $\mathrm{aff}(\mathcal{B}^{lb})$ is*

$$c_F(\mathcal{A}, \mathrm{aff}(\mathcal{B}^{lb}))^2 = 1 - \tfrac{1}{R}\left(1 - \cos\left(\tfrac{2\pi}{N}\right)\right).$$

Around the optimal solution 0, the polyhedra $\mathcal{A}$ and $\mathcal{B}^{\text{lb}}$ behave like subspaces, and it is possible to pick initializations $a_0 \in \mathcal{A}$ and $b_0 \in \mathcal{B}^{\text{lb}}$ such that the Friedrichs angle exactly determines the rate of convergence. That means $1 - 1/\kappa_*^2 = c_F(\mathcal{A}, \mathrm{aff}(\mathcal{B}^{\text{lb}}))^2$, and

$$\|a_k\| = (1 - \tfrac{1}{R}(1 - \cos(\tfrac{2\pi}{N})))^k \|a_0\| \quad \text{and} \quad \|b_k\| = (1 - \tfrac{1}{R}(1 - \cos(\tfrac{2\pi}{N})))^k \|b_0\|.$$

Bounding $1 - \cos(x) \leq \frac{1}{2}x^2$ leads to the following lower bound on the rate.

Theorem 14. *There exists a decomposed function F^{lb} and initializations for which the convergence rate of AP is at least $1 - \frac{2\pi^2}{N^2 R}$.*

This theoretical bound can also be observed empirically (Figure 3 in Appendix B).

5 Convergence of the Primal Objective

We have shown that AP generates 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 (P4). 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). The proofs may be found in Appendix C.

Define the matrix $\Gamma = -R^{1/2}S$, where S is the matrix defined in Equation (5). Multiplication by Γ maps a vector $(w_1, \ldots, w_R)$ to $-\sum_r w_r$, 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. [25], 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 $1 - \frac{1}{N^2 R^2}$.*

This linear rate of convergence translates into a linear rate for the original discrete problem.

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 $1 - \frac{1}{2N^2 R^2}$.*

6 Discussion

In this work, we analyze projection methods for parallel 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 ϵ is logarithmic in $1/\epsilon$, not linear as in previous work [35]. Our rate is uniform over all submodular functions. Moreover, our proof highlights how the number R of components and the facial structure of $\mathcal{B}$ 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 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, $F(V) = F(S) + F(V\backslash S)$ for some nonempty $S \subsetneq V$. Then the subsets $A_{rm} \subseteq V$ defining the relevant faces of $\mathcal{B}$ satisfy either $A_{rm} \subseteq S$ or $A_{rm} \subseteq S^c$ [2]. 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. For example, Jegelka et al. [25] also considered the related Douglas–Rachford (DR) algorithm. DR between subspaces converges linearly with rate c_F [6], 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 cyclic updates (instead of parallel ones), multiple polyhedra, and stochastic algorithms.

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 work is supported in part by the Office of Naval Research under grant number N00014-11-1-0688, the US ARL and the US ARO under grant number W911NF-11-1-0391, and the NSF under grant number DGE-1106400.

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. B. Cruz, T. T. Nghia, 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. *Journal of Approximation Theory*, 185:63–79, 2014.

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

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

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

[20] 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.

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

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

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

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

[25] 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.

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

[27] 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.

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

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

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

[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

$$\begin{aligned}
\kappa_*^{-2} d(p_k, E)^2 &\leq d(p_k, Q')^2 \\
&\leq \|p_k - \Pi_Q p_k + v\|^2 \\
&\leq \|(p_k - \Pi_Q p_k) - (e - \Pi_Q e)\|^2 \\
&\leq \|p_k - e\|^2 - \|\Pi_Q p_k - \Pi_Q e\|^2 \\
&\leq d(p_k, E)^2 - d(q_k, H)^2.
\end{aligned}$$

It follows that $d(q_k, H) \leq (1 - \kappa_*^{-2})^{1/2} d(p_k, E)$. Similarly, we have $d(p_{k+1}, E) \leq (1 - \kappa_*^{-2})^{1/2} d(q_k, H)$. When combining these, induction shows that

$$\begin{aligned}
d(p_k, E) &\leq (1 - \kappa_*^{-2})^k d(p_0, E) \\
d(q_k, H) &\leq (1 - \kappa_*^{-2})^k d(q_0, H).
\end{aligned}$$

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

$$\begin{aligned}
\|p_k - p_*\| &\leq 2\|p_0 - p_*\|(1 - \kappa_*^{-2})^k \\
\|q_k - q_*\| &\leq 2\|q_0 - q_*\|(1 - \kappa_*^{-2})^k.
\end{aligned}$$

A.2 Connection Between κ and c_F in the Subspace Case

In this section, we introduce a simple lemma connecting κ 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. □

A.3 Proof of Lemma 3

It suffices to prove the statement for $p \in P \backslash E$. For $p \in P \backslash E$, 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 $d(\Pi_P q, E) > 0$. We have

$$\kappa(p) = \frac{d(p, E)}{d(p, Q')} \leq \frac{\|p - e\|}{\|p - q\|} \leq \frac{\|q - e\|}{\|q - p''\|} \leq \frac{d(q, E)}{d(q, P)} = \kappa(q). \tag{8}$$

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 θ 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.

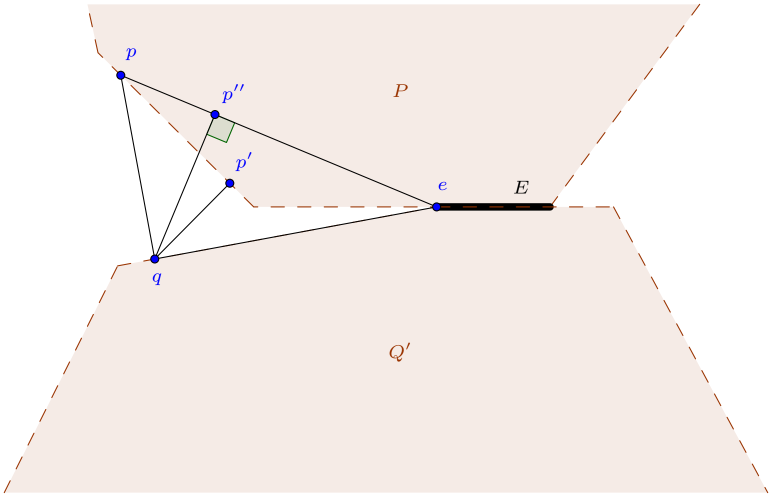

Figure 2: Illustration of the proof of Lemma 3.

A.4 Proof of Proposition 4

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

$$\begin{array}{ccccccc} P & \supseteq & P_1 & \supseteq & \cdots & \supseteq & P_J \\ Q' & \supseteq & Q'_1 & \supseteq & \cdots & \supseteq & Q'_J \end{array}.$$

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 $x \in \mathbb{R}^D \backslash E_j$ (this is just the function κ 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 $p_j \in \text{relint}(P_j) \backslash E_j$, $q_j \in \text{relint}(Q'_j) \backslash E_j$, and $q_j = \Pi_{Q'_j} p_j$ for each j. Furthermore, we will have

$$\kappa(p) \leq \kappa_1(p_1) \leq \cdots \leq \kappa_J(p_J). \tag{9}$$

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 $\epsilon > 0$ sufficiently small so that every face of either P or Q' that intersects $B_\epsilon(e)$, the ball of radius ϵ centered on e, necessarily contains e. Now choose $0 \leq t_0 < 1$ sufficiently close to 1 so that $\|p^{t_0} - e\| < \epsilon$. It follows that e is contained in the face of P whose relative interior contains p^{t_0}. It further follows that e is contained in the face of Q' whose relative interior contains $\Pi_{Q'} p^{t_0}$ because

$$\|\Pi_{Q'} p^{t_0} - e\| = \|\Pi_{Q'} p^{t_0} - \Pi_{Q'} e\| \leq \|p^{t_0} - e\| < \epsilon.$$

To initialize the construction, set

$$p_1 = p^{t_0}$$
$$q_1 = \Pi_{Q'} p^{t_0},$$

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 $x_k \in \operatorname{relint}(P_j)$ and $y_k \in \operatorname{relint}(Q'_j)$ for every k, or there is some k for which either $x_k \notin \operatorname{relint}(P_j)$ or $y_k \notin \operatorname{relint}(Q'_j)$. Note that P_j and Q'_j intersect and that AP between them will not terminate after a finite number of steps.

Suppose that $x_k \in \operatorname{relint}(P_j)$ and $y_k \in \operatorname{relint}(Q'_j)$ for every k. Then set $J = j$ and terminate the procedure. Otherwise, choose k' such that either $x_{k'} \notin \operatorname{relint}(P_j)$ or $y_{k'} \notin \operatorname{relint}(Q'_j)$. 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 $p_{j+1} \in \operatorname{relint}(P_{j+1})$ and $q_{j+1} \in \operatorname{relint}(Q'_{j+1})$. 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 $\dim(\operatorname{aff}(P_{j+1})) < \dim(\operatorname{aff}(P_j))$ and if $Q'_{j+1} \neq Q'_j$ then $\dim(\operatorname{aff}(Q'_{j+1})) < \dim(\operatorname{aff}(Q'_j))$. 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 $\operatorname{aff}(P_J)$ and $\operatorname{aff}(Q'_J)$. Using this fact, along with [12, Theorem 9.3], we see that $\Pi_{\operatorname{aff}(P_J)\cap\operatorname{aff}(Q'_J)} p_J \in E_J$. Using this, along with Lemma 17(b), we see that

$$\kappa_J(p_J) \leq (1 - c_F(\operatorname{aff}_0(P_J), \operatorname{aff}_0(Q'_J))^2)^{-1/2}. \tag{10}$$

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

$$\begin{aligned} c_F(\operatorname{null}(S), \operatorname{null}(T)) &= c_F(\operatorname{range}(S^\top)^\perp, \operatorname{range}(T^\top)^\perp) \\ &= c_F(\operatorname{range}(S^\top), \operatorname{range}(T^\top)), \end{aligned}$$

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

$$\sigma_j = \max_{u,v} \{u^\top S T^\top v \,|\, u \perp \operatorname{span}(u_1, \ldots, u_{j-1}), v \perp \operatorname{span}(v_1, \ldots, v_{j-1}), \|u\| = 1, \|v\| = 1\}.$$

Since the $\{u_j\}_j$ are orthonormal, so are the $\{x_j\}_j$. Similarly, since the $\{v_j\}_j$ are orthonormal, so are the $\{y_j\}_j$. Suppose that all of the singular values of $ST^\top$ 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

$$X \cap Y = \operatorname{span}(x_1, \ldots, x_\ell) = \operatorname{span}(y_1, \ldots, y_\ell),$$

and so

$$\begin{aligned} \sigma_{\ell+1} &= \sup_{u,v} \{u^\top S T^\top v \,|\, u \in \operatorname{span}(u_1, \ldots, u_\ell)^\perp, v \in \operatorname{span}(v_1, \ldots, v_\ell)^\perp, \|u\| = 1, \|v\| = 1\} \\ &= \sup_{x,y} \{x^\top y \,|\, x \in X \cap (X \cap Y)^\perp, y \in Y \cap (X \cap Y)^\perp, \|x\| = 1, \|y\| = 1\} \\ &= c_F(X, Y). \end{aligned}$$

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, $\operatorname{vol}(U) = \sum_{i \in U} \delta_i$, 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

$$h_G = \min_{\emptyset \neq U \subsetneq V_G} \frac{|E(U, U^c)|}{\min(\operatorname{vol}(U), \operatorname{vol}(U^c))}.$$

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

$$L_{ij} = \left\{ \begin{array}{ll} -w_{ij} & i \neq j \\ \delta_i & \text{otherwise} \end{array} \right. .$$

Let D be the $|V_G| \times |V_G|$ diagonal matrix defined by $D_{ii} = \delta_i$. Then $\mathcal{L} = D^{-1/2} L D^{-1/2}$ 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

$$\begin{aligned} \min(\operatorname{vol}(U), \operatorname{vol}(U^c)) &\leq \frac{1}{2} \operatorname{vol}(V_G) \\ &= \frac{1}{2} \sum_{(r,m)} \left(\sum_{(r',m') \neq (r,m)} |A_{rm} \cap A_{r'm'}| \right) \\ &= \frac{1}{2} \sum_{(r,m)} (R-1)|A_{rm}| \\ &= \frac{1}{2} NR(R-1). \end{aligned}$$

Since G is connected, for any nonempty set $U \subsetneq V_G$, 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 $v \in A_{r_1 m_1} \cap A_{r_2 m_2}$ 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

$$h_G \geq \frac{R-1}{\frac{1}{2}NR(R-1)} = \frac{2}{NR}.$$

□

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

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*

$$\begin{aligned}\|u_k\| &= c_F(U,V)^{2k}\|u_0\| \\ \|v_k\| &= c_F(U,V)^{2k}\|v_0\|.\end{aligned}$$

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 $u_*^\top v_* = c_F(U,V)$, which we can do by compactness. By Lemma 17(a),

$$c_F(U,V) = v_*^\top u_* = v_*^\top \Pi_V u_* \leq \|\Pi_V u_*\| \leq c_F(U,V).$$

Set $u_0 = u_*$ and generate the sequences $\{u_k\}_{k\geq 0}$ and $\{v_k\}_{k\geq 0}$ via AP. Since $\|\Pi_V u_0\| = c_F(U,V)$, Lemma 17(c) implies that $\kappa(u_0) = (1 - c_F(U,V)^2)^{-1/2}$. Since κ attains its maximum at u_0, Lemma 3 implies that κ attains the same value at every element of the sequences $\{u_k\}_{k\geq 0}$ and $\{v_k\}_{k\geq 0}$. Therefore, Lemma 17(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 $U' \cap B_\epsilon(0) = U \cap B_\epsilon(0)$ and $V' \cap B_\epsilon(0) = V \cap B_\epsilon(0)$ 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 AP between U' and V' initialized at u_0' satisfy*

$$\begin{aligned}\|u_k'\| &= c_F(U,V)^{2k}\|u_0'\| \\ \|v_k'\| &= c_F(U,V)^{2k}\|v_0'\|.\end{aligned}$$

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

B.2 Proof of Lemma 13

Observe that we can write

$$\mathrm{aff}(\mathcal{B}^{\mathrm{lb}}) = \{(s_1, -s_1, \ldots, s_{\frac{N}{2}}, -s_{\frac{N}{2}}, -t_{\frac{N}{2}}, t_1, -t_1, \ldots, t_{\frac{N}{2}}, 0, \ldots, 0, \ldots, 0, \ldots, 0) \mid s_i, t_j \in \mathbb{R}\}.$$

We can write $\mathrm{aff}(\mathcal{B}^{\mathrm{lb}})$ as the nullspace of the matrix

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

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

$$T_{\mathrm{lb},1} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 & & & & & \\ & & 1 & 1 & & & \\ & & & & \ddots & & \\ & & & & & 1 & 1 \end{pmatrix} \qquad T_{\mathrm{lb},2} = \frac{1}{\sqrt{2}}\begin{pmatrix} & 1 & 1 & & & & \\ & & & 1 & 1 & & \\ & & & & & \ddots & \\ 1 & & & & & & 1 \end{pmatrix}.$$

Recall that we can write $\mathcal{A}$ as the nullspace of the matrix S defined in Equation (5). It follows from Lemma 6 that $c_F(\mathcal{A}, \mathrm{aff}(\mathcal{B}^{\mathrm{lb}}))$ equals the largest singular value of $ST_{\mathrm{lb}}^\top$ that is less than one. We have

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

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

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

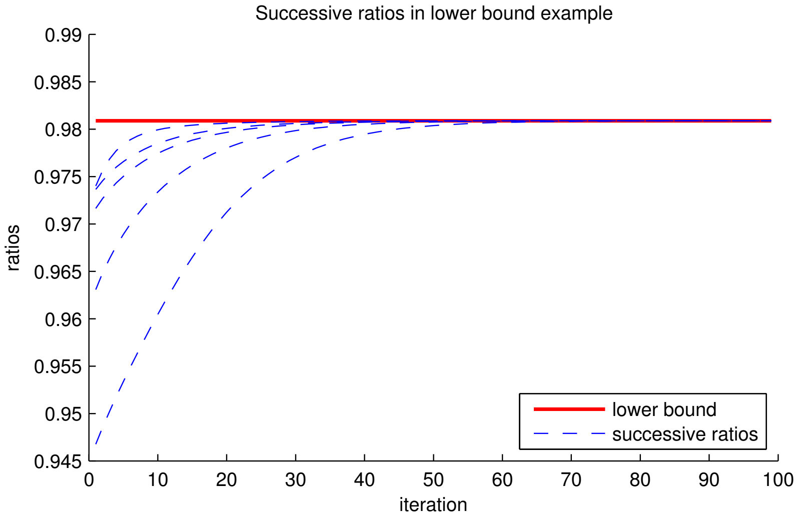

Figure 3: We run five trials of AP between $\mathcal{A}$ and $\mathcal{B}^{\text{lb}}$ with random initializations, where $N = 10$ and $R = 10$. For each trial, we plot the ratios $d(a_{k+1}, E)/d(a_k, E)$, where $E = \mathcal{A} \cap \mathcal{B}^{\text{lb}}$ is the optimal set. The red line shows the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

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

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

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

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

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

$$T_{\text{lb},0}^\top T_{\text{lb},0} = \frac{1}{2} \begin{pmatrix} 2 & 1 & & & 1 \\ 1 & 2 & & & \\ & & \ddots & & \\ & & & 2 & 1 \\ 1 & & & 1 & 2 \end{pmatrix}.$$

The eigenvalues of $T_{\text{lb},0}^\top T_{\text{lb},0}$ are given by $\lambda_j = 1 + \cos\left(\frac{2\pi j}{N}\right)$ for $0 \leq j \leq N-1$ (see Gray [18, Section 3.1] for a derivation). Therefore,

$$c_F(\mathcal{A}, \text{aff}(\mathcal{B}^{\text{lb}}))^2 = 1 - \tfrac{1}{R}(1 - \cos(\tfrac{2\pi}{N})).$$

B.3 Lower Bound Illustration

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

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

where $E = \mathcal{A} \cap \mathcal{B}^{\text{lb}}$ 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 $\mathcal{A}$ and $\mathcal{B}^{\text{lb}}$ 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 $\mathcal{A}$ and $\mathcal{B}^{\text{lb}}$ 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

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

Recall that we defined $F_{\max} = \max_A |F(A)|$. 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 $1 - \frac{1}{N^2R^2}$. 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 $\|x_k - x_*\| \leq \|\Gamma\|\|b_k - b_*\| \leq \sqrt{R}\|b_k - b_*\|$. 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

$$\begin{aligned}
\tfrac{1}{2}\|x_k\|^2 - \tfrac{1}{2}\|x_*\|^2 &\leq x_k^\top (x_k - x_*) \\
&\leq \|x_k\|\|x_k - x_*\| \\
&\leq 3F_{\max}\sqrt{R}\|b_k - b_*\| \\
&\leq 6F_{\max}\sqrt{R}\|b_0 - b_*\|(1 - \tfrac{1}{N^2R^2})^k.
\end{aligned} \tag{12}$$

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

$$\begin{aligned}
f(x_k) - f(x_*) &\leq s^\top (x_k - x_*) \\
&\leq \|s\|\|x_k - x_*\| \\
&\leq 3F_{\max}\sqrt{R}\|b_k - b_*\| \\
&\leq 6F_{\max}\sqrt{R}\|b_0 - b_*\|(1 - \tfrac{1}{N^2R^2})^k.
\end{aligned} \tag{13}$$

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

$$(f(x_k) + \tfrac{1}{2}\|x_k\|^2) - (f(x_*) + \tfrac{1}{2}\|x_*\|^2) \leq 12F_{\max}\sqrt{R}\|b_0 - b_*\|(1 - \tfrac{1}{N^2R^2})^k. \tag{14}$$

C.2 Proof of Theorem 16

We will make use of the following result, shown in [2, Proposition 10.5] and stated below for convenience.

Proposition 21. *Let $(w, s) \in \mathbb{R}^N \times B(F)$ be a pair of primal-dual candidates for the minimization of $\frac{1}{2}\|w\|^2 + f(w)$, with duality gap $\epsilon = \frac{1}{2}\|w\|^2 + f(w) + \frac{1}{2}\|s\|^2$. Then if A is the suplevel set of w with smallest value of F, then*

$$F(A) - s_-(V) \leq \sqrt{N\epsilon/2}.$$

Using this result in our setting, recall that 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

$$\begin{aligned} F(A_k) - F(A_*) &\leq F(A_k) - (s_*)_-(V) \\ &\leq \sqrt{\tfrac{N}{2}((f(x_k) + \tfrac{1}{2}\|x_k\|^2) - (f(x_*) + \tfrac{1}{2}\|x_*\|^2))} \\ &\leq \sqrt{6F_{\max} N R^{1/2} \|b_0 - b_*\|}\,(1 - \tfrac{1}{N^2R^2})^{k/2} \\ &\leq \sqrt{6F_{\max} N R^{1/2} \|b_0 - b_*\|}(1 - \tfrac{1}{2N^2R^2})^{k}, \end{aligned}$$

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 ϵ 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 $\frac{1}{k}$, 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.

C.3 Running times

Theorem 16 implies that the number of iterations required for an accuracy of ϵ is at most

$$2N^2R^2 \log\left(\frac{\sqrt{6F_{\max} N R^{1/2} \|b_0 - b_*\|}}{\epsilon}\right). \tag{15}$$

Each iteration involves minimizing each of the F_r separately. For comparison, the number of iterations required in Stobbe and Krause [35] is

$$24\sqrt{N}R\frac{F_{\max}}{\epsilon}.$$

The dependence of this algorithm on N and R is better, but its dependence on the constants $F_{\max}/\epsilon$ is worse. For example, to obtain the exact discrete solution, we need $\epsilon < \min_{S,T} |F(S) - F(T)|$. This is one for integer-valued functions (in which case the lower rate may be desirable), but can otherwise become very small. The constant $F_{\max}$ can be of order $O(N)$ in general (or even larger if the function becomes very negative). For empirical comparisons, we refer the reader to [25].

The running times of the combinatorial algorithm by Kolmogorov [29] apply to *integer-valued* functions (as opposed to the generic ones above) and range from $O((N + R)^2 \log F_{\max})$ for cuts to $O((N + Q^2R)(N + Q^2R + QR\tau_2) \log F_{\max})$, where $Q \leq N$ is the maximal cardinality of the support of any F_r, and τ_2 is the time required to minimize a simple function. This is better than (15) if Q is a small constant, and worse as Q gets closer to N.

For comparison, if not exploiting decomposition, one may use combinatorial algorithms, the Frank-Wolfe algorithm (conditional gradient descent), or a subgradient method. The combinatorial algorithm by Orlin [34] has a running time of $O(N^5\tau_1 + N^6)$, and the algorithm by Iwata [23] (for integer-valued functions) a time of $O((N^4\tau_1 + N^5) \log F_{\max})$, where τ_1 is the time required to evaluate F. For an accuracy of ϵ in the discrete objective, Frank-Wolfe will take $64N\frac{F_{\max}}{\epsilon^2}$ iterations, each taking time $O(N \log N)$. The subgradient method behaves similarly.