Provable Submodular Minimization via Fujishige-Wolfe's Algorithm*

Deeparnab Chakrabarty† Prateek Jain* Pravesh Kothari‡

Abstract

Owing to several applications in large scale learning and vision problems, fast submodular function minimization (SFM) has become a critical problem. Theoretically, unconstrained SFM can be performed in polynomial time [10, 11]. However, these algorithms are typically not practical. In 1976, Wolfe [22] proposed an algorithm to find the minimum Euclidean norm point in a polytope, and in 1980, Fujishige [4] showed how Wolfe's algorithm can be used for SFM. For general submodular functions, this Fujishige-Wolfe minimum norm algorithm seems to have the best empirical performance.

Despite its good practical performance, very little is known about Wolfe's minimum norm algorithm theoretically. To our knowledge, the only result is an exponential time analysis due to Wolfe [22] himself. In this paper we give the first convergence analysis of Wolfe's algorithm. We prove that in t iterations, Wolfe's algorithm returns an $O(1/t)$-approximate solution to the min-norm point on *any* polytope. We also prove a robust version of Fujishige's theorem which shows that an $O(1/n)$-approximate solution to the min-norm point on the base polytope implies *exact* submodular minimization for integer valued submodular functions. As a corollary, we get the first pseudo-polynomial time guarantee for the Fujishige-Wolfe minimum norm algorithm for unconstrained submodular function minimization.

1 Introduction

An integer-valued[1] function $f : 2^X \to \mathbb{Z}$ defined over subsets of some finite ground set X of n elements is submodular if it satisfies the following *diminishing marginal returns* property: for every $S \subseteq T \subseteq X$ and $i \in X \setminus T$, $f(S \cup \{i\}) - f(S) \geq f(T \cup \{i\}) - f(T)$. Submodularity arises naturally in several applications such as image segmentation [17], sensor placement [18], etc. where minimizing an arbitrary submodular function is an important primitive.

In submodular function minimization (SFM), we assume access to an *evaluation oracle for* f which for any subset $S \subseteq X$ returns the value $f(S)$. We denote the time taken by the oracle to answer a single query as EO. The objective is to find a set $T \subseteq X$ satisfying $f(T) \leq f(S)$ for every $S \subseteq X$. In 1981, Grotschel, Lovasz and Schrijver [8] demonstrated the first polynomial time algorithm for SFM using the ellipsoid algorithm. This algorithm, however, is practically infeasible due to the running time and the numerical issues in implementing the ellipsoid algorithm. In 2001, Schrijver [20] and Iwata et al. [9] independently designed *combinatorial* polynomial time algorithms for SFM. Currently, the best algorithm is by Iwata and Orlin [11] with a running time of $O(n^5\text{EO} + n^6)$.

However, from a practical stand point, none of the provably polynomial time algorithms exhibit good performance on instances of SFM encountered in practice (see §4). This, along with the widespread applicability of SFM in machine learning, has inspired a large body of work on *practically* fast procedures (see [2] for a survey). But most of these procedures focus either on special submodular functions such as decomposable functions [16, 21] or on constrained SFM problems [13, 12, 15, 14].

Fujishige-Wolfe's Algorithm for SFM: For any submodular function f, the *base polytope* $\mathcal{B}_f$ of f is defined as follows:

$$\mathcal{B}_f = \{x \in \mathbb{R}^n : \ x(A) \leq f(A), \ \forall A \subset X, \ \text{ and } \ x(X) = f(X)\}, \tag{1}$$

*Preliminary version appeared in Advances of Neural Information Processing Systems (NIPS), 2014

†Microsoft Research, 9 Lavelle Road, Bangalore 560001.

‡University of Texas at Austin (Part of the work done while interning at Microsoft Research)

[1] One can assume any function is integer valued after suitable scaling though our guarantees will suffer.

where $x(A) := \sum_{i \in A} x_i$ and x_i is the i-th coordinate of $x \in \mathbb{R}^n$. Fujishige [4] showed that if one can obtain the minimum norm point on the base polytope, then one can solve SFM. Finding the minimum norm point, however, is a non-trivial problem; at present, to our knowledge, the only polynomial time algorithm known is via the ellipsoid method. Wolfe [22] described an iterative procedure to find minimum norm points in polytopes as long as linear functions could be (efficiently) minimized over them. Although the base polytope has exponentially many constraints, a simple greedy algorithm can minimize any linear function over it. Therefore using Wolfe's procedure on the base polytope coupled with Fujishige's theorem becomes a natural approach to SFM. This was suggested as early as 1984 in Fujishige [5] and is now called the Fujishige-Wolfe algorithm for SFM.

This approach towards SFM was revitalized in 2006 when Fujishige and Isotani [6, 7] announced encouraging computational results regarding the minimum norm point algorithm. In particular, this algorithm significantly outperformed all known *provably* polynomial time algorithms. Theoretically, however, little is known regarding the convergence of Wolfe's procedure except for the finite, but exponential, running time Wolfe himself proved. Nor is the situation any better for its application on the base polytope. Given the practical success, we believe this is an important, and intriguing, theoretical challenge.

In this work, we make some progress towards analyzing the Fujishige-Wolfe method for SFM and, in fact, Wolfe's algorithm in general. In particular, we prove the following two results:

- We prove (in Theorem 4) that for *any* polytope $\mathcal{B}$, Wolfe's algorithm converges to an ε-approximate solution, in $O(1/\varepsilon)$ steps. More precisely, in $O(D^2/\varepsilon)$ iterations, Wolfe's algorithm returns a point $\|x\|_2^2 \leq \min_{z \in \mathcal{B}} z^\top x + \varepsilon$, where $D = \max_{p,q \in \mathcal{B}} \|p - q\|_2$. Note that $\|x\|_2^2 \leq \min_{z \in \mathcal{B}} z^\top x + \varepsilon$ implies $\|x - x_*\|_2^2 \leq 2\varepsilon$.
- We prove (in Theorem 5) a robust version of a theorem by Fujishige [4] relating min-norm points on the base polytope to SFM. In particular, we prove that an approximate min-norm point solution provides an approximate solution to SFM as well. More precisely, if x satisfies $\|x\|_2^2 \leq z^T x + \varepsilon$ for all $z \in \mathcal{B}_f$, then, $f(S_x) \leq \min_S f(S) + 2\sqrt{n\varepsilon}$, where S_x can be constructed efficiently using x. Such a relation was also observed in Bach [1]; our proof is similar and we include it for completeness.

Together, these two results gives us our main result which is a pseudopolynomial bound on the running time of the Fujishige-Wolfe algorithm for submodular function minimization.

Theorem 1. (Main Result.) *Fix a submodular function $f : 2^X \to \mathbb{Z}$. The Fujishige-Wolfe algorithm returns the minimizer of f in $O((n^3\mathrm{EO} + n^4)F^2)$ time where $F := \max_{i=1}^n (|f(\{i\})|, |f([n]) - f([n] \setminus i)|)$.*

Our analysis suggests that the Fujishige-Wolfe's algorithm is dependent on F and has worse dependence on n than the Iwata-Orlin [11] algorithm. To verify this, we conducted empirical study on several standard SFM problems. However, for the considered benchmark functions, running time of Fujishige-Wolfe's algorithm seemed to be independent of F and exhibited better dependence on n than the Iwata-Orlin algorithm. This is described in §4.

2 Preliminaries: Submodular Functions and Wolfe's Algorithm

2.1 Submodular Functions and SFM

Given a ground set X on n elements, without loss of generality we think of it as the first n integers $[n] := \{1, 2, \ldots, n\}$. f be a submodular function. Since submodularity is translation invariant, we assume $f(\emptyset) = 0$. For a submodular function f, we write $\mathcal{B}_f \subseteq \mathbb{R}^n$ for the associated base polyhedron of f defined in (1). Given $x \in \mathbb{R}^n$, one can find the minimum value of $q^\top x$ over $q \in \mathcal{B}_f$ in $O(n \log n + n\mathrm{EO})$ time using the following greedy algorithm: Renumber indices such that $x_1 \leq \cdots \leq x_n$. Set $q_i^* = f([i]) - f([i-1])$. Then, it can be proved that $q^* \in \mathcal{B}_f$ and is the minimizer of the $x^\top q$ for $q \in \mathcal{B}_f$.

The connection between the SFM problem and the base polytope was first established in the following minimax theorem of Edmonds [3].

Theorem 2 (Edmonds [3])**.** *Given any submodular function f with $f(\emptyset) = 0$, we have*

$$\min_{S \subseteq [n]} f(S) = \max_{x \in \mathcal{B}_f} \left(\sum_{i : x_i < 0} x_i \right)$$

The following theorem of Fujishige [4] shows the connection between finding the minimum norm point in the base polytope $\mathcal{B}_f$ of a submodular function f and the problem of SFM on input f. This forms the basis of the Fujishige-Wolfe algorithm. In §3.2, we prove a robust version of this theorem.

Theorem 3 (Fujishige's Theorem [4]). *Let $f : 2^{[n]} \to \mathbb{Z}$ be a submodular function and let $\mathcal{B}_f$ be the associated base polyhedron. Let x^* be the optimal solution to $\min_{x \in \mathcal{B}_f} ||x||$. Define $S = \{i \mid x^*_i < 0\}$. Then, $f(S) \leq f(T)$ for every $T \subseteq [n]$.*

2.2 Wolfe's Algorithm for Minimum Norm Point of a polytope.

We now present Wolfe's algorithm for computing the minimum-norm point in an arbitrary polytope $\mathcal{B} \subseteq \mathbb{R}^n$. We assume a *linear optimization oracle* (LO) which takes input a vector $x \in \mathbb{R}^n$ and outputs a vector $q \in \arg\min_{p \in \mathcal{B}} x^\top p$.

We start by recalling some definitions. The *affine hull* of a finite set $S \subseteq \mathbb{R}^n$ is $\texttt{aff}(S) = \{y \mid y = \sum_{z \in S} \alpha_z \cdot z, \sum_{z \in S} \alpha_z = 1\}$. The *affine minimizer* of S is defined as $y = \arg\min_{z \in \texttt{aff}(S)} ||z||_2$, and y satisfies the following *affine minimizer property*: for any $v \in \texttt{aff}(S)$, $v^\top y = ||y||^2$. The procedure $\texttt{AffineMinimizer}(S)$ returns (y, α) where y is the affine minimizer and $\alpha = (\alpha_s)_{s \in S}$ is the set of coefficients expressing y as an affine combination of points in S. This procedure can be naively implemented in $O(|S|^3 + n|S|^2)$ as follows. Let B be the $n \times |S|$ matrix where each column in a point in S. Then $\alpha = (B^\top B)^{-1}\mathbf{1}/\mathbf{1}^\top (B^\top B)^{-1}\mathbf{1}$ and $y = B\alpha$. Wolfe [22] showed how on the average all the affine minimization steps in the algorithm below can be done in $O(n^2)$ time *per call*.

Algorithm 1 Wolfe's Algorithm

1. Let q be an arbitrary vertex of $\mathcal{B}$. Initialize $x \leftarrow q$. We always maintain $x = \sum_{i \in S} \lambda_i q_i$ as a convex combination of a subset S of vertices of $\mathcal{B}$. Initialize $S = \{q\}$ and $\lambda_1 = 1$.
2. **WHILE** `(true)`: (MAJOR CYCLE)
   (a) $q := \text{LO}(x)$. *// Linear Optimization: $q \in \arg\min_{p \in \mathcal{B}} x^\top p$.*
   (b) **IF** $||x||^2 \leq x^\top q + \varepsilon^2$ **THEN** `break`. *// Termination Condition. Output x.*
   (c) $S := S \cup \{q\}$.
   (d) **WHILE** `(true)` **:** (MINOR CYCLE)
      i. $(y, \alpha) = \texttt{AffineMinimizer}(S)$. *//$y = \arg\min_{z \in \texttt{aff}(S)} ||z||$.*
      ii. **IF** $\alpha_i \geq 0$ for all i **THEN** `break`. *//If $y \in \texttt{conv}(S)$, then end minor loop.*
      iii. **ELSE**
         // If $y \notin \texttt{conv}(S)$, then update x to the intersection of the boundary of $\texttt{conv}(S)$ and the segment joining y and previous x. Delete points from S which are not required to describe the new x as a convex combination.
         $\theta := \min_{i:\alpha_i < 0} \lambda_i/(\lambda_i - \alpha_i)$ *// Recall, $x = \sum_i \lambda_i q_i$.*
         Update $x \leftarrow \theta y + (1-\theta)x$. *// By definition of θ, the new x lies in $\texttt{conv}(S)$.*
         Update $\lambda_i \leftarrow \theta\alpha_i + (1-\theta)\lambda_i$. *//This sets the coefficients of the new x*
         $S = \{i : \lambda_i > 0\}$. *// Delete points which have $\lambda_i = 0$. This deletes at least one point.*
   (e) Update $x \leftarrow y$. *// After the minor loop terminates, x is updated to be the affine minimizer of the current set S.*
3. **RETURN** x.

When $\varepsilon = 0$, the algorithm on termination (if it terminates) returns the minimum norm point in $\mathcal{B}$ since $||x||^2 \leq x^\top x_* \leq ||x|| \cdot ||x_*||$. For completeness, we sketch Wolfe's argument in [22] of finite termination. Note that $|S| \leq n$ always; otherwise the affine minimizer is 0 which either terminates the program or starts a minor cycle which decrements $|S|$. Thus, the number of minor cycles in a major cycle $\leq n$, and it suffices to bound the number of major cycles. Each major cycle is associated with a set S whose affine minimizer, which is the current x, lies in the convex hull of S. Wolfe calls such sets *corrals*. Next, we show that $||x||$ strictly decreases across iterations (major or minor

cycle) of the algorithm, which proves that no corral repeats, thus bounding the number of major cycles by the number of corrals. The latter is at most $\binom{N}{n}$, where N is the number of vertices of $\mathcal{B}$.

Consider iteration j which starts with x_j and ends with x_{j+1}. Let S_j be the set S at the beginning of iteration j. If the iteration is a major cycle, then x_{j+1} is the affine minimizer of $S_j \cup \{q_j\}$ where $q_j = \text{LO}(x_j)$. Since $x_j^\top q_j < ||x_j||^2$ (the algorithm doesn't terminate in iteration j) and $x_{j+1}^\top q_j = ||x_{j+1}||^2$ (affine minimizer property), we get $x_j \neq x_{j+1}$, and so $||x_{j+1}|| < ||x_j||$ (since the affine minimizer is unique). If the iteration is a minor cycle, then $x_{j+1} = \theta x_j + (1-\theta) y_j$, where y_j is the affine minimizer of S_j and $\theta < 1$. Since $||y_j|| < ||x_j||$ ($y_j \neq x_j$ since $y_j \notin \texttt{conv}(S_j)$), we get $||x_{j+1}|| < ||x_j||$.

3 Analysis

Our refined analysis of Wolfe's algorithm is encapsulated in the following theorem.

Theorem 4. *Let $\mathcal{B}$ be an* arbitrary *polytope of diameter D and which is contained in a ball of radius R. After $O(D^2/\varepsilon + \log(R/\varepsilon))$ iterations, Wolfe's algorithm returns a point $x \in \mathcal{B}$ which satisfies $||x||^2 \leq \min_{q \in \mathcal{B}} x^\top q + \varepsilon$, for all points $q \in \mathcal{B}$. In particular, this implies $||x - x_*||^2 \leq 2\varepsilon$.*

The above theorem shows that Wolfe's algorithm converges to the minimum norm point at an $1/t$-rate. We stress that the above is for *any* polytope. To apply this to SFM, we prove the following robust version of Fujishige's theorem connecting the minimum norm point in the base polytope and the set minimizing the submodular function value.

Theorem 5. *Fix a submodular function f with base polytope $\mathcal{B}_f$. Let $x \in \mathcal{B}_f$ be such that $||x||^2 \leq x^\top q + \varepsilon$ for all $q \in \mathcal{B}_f$. Renumber indices such that $x_1 \leq \cdots \leq x_n$. Then there exists a $1 \leq k \leq n$ such that $f(\{1, 2, \ldots, k\}) \leq \min_S f(S) + 2\sqrt{n\varepsilon}$. In particular, if $\varepsilon < \frac{1}{4n}$ and f is integer-valued, then S is a minimizer.*

We remark here that in a previous version of the paper we had a weaker result – n instead of $\sqrt{n}$. We thank Francis Bach for pointing us that we can get $\sqrt{n}$. Theorem 4 and Theorem 5 implies our main theorem.

Theorem 1. (Main Result.) *Fix a submodular function $f : 2^X \to \mathbb{Z}$. The Fujishige-Wolfe algorithm returns the minimizer of f in $O((n^3 \text{EO} + n^4)F^2)$ time where $F := \max_{i=1}^n (|f(\{i\})|, |f([n]) - f([n] \setminus i)|)$.*

Proof. The vertices of $\mathcal{B}_f$ are well understood: for every permutation σ of $[n]$, we have a vertex with $x_{\sigma(i)} = f(\{\sigma(1), \ldots, \sigma(i)\}) - f(\{\sigma(1), \ldots, \sigma(i-1)\})$. This implies that the diameter of $\mathcal{B}_f$ is $\leq \sqrt{n}F$. Choose $\varepsilon = 1/4n$. From Theorem 4 we know that if we run $O(n^2F^2)$ iterations of Wolfe, we will get a point $x \in \mathcal{B}_f$ such that $||x||^2 \leq x^\top q + \varepsilon$ for all $q \in \mathcal{B}_f$. Theorem 5 implies this solves the SFM problem. The running time for each iteration is dominated by the time for the subroutine to compute the affine minimizer of S which is at most $O(n^2)$ per iteration, and the linear optimization oracle. For $\mathcal{B}_f$, $\text{LO}(x)$ can be implemented in $O(n \log n + n\text{EO})$ time. This proves the theorem. □

We prove Theorem 4 and Theorem 5 in §3.1 and §3.2, respectively.

3.1 Analysis of Wolfe's Min-norm Point Algorithm

The stumbling block in the analysis of Wolfe's algorithm is the interspersing of major and minor cycles which oscillates the size of S preventing it from being a good measure of progress. Instead, in our analysis, we use the norm of x as the measure of progress. Already we have seen that $||x||$ strictly decreases. It would be nice to quantify how much the decrease is, say, across one major cycle. This, at present, is out of our reach even for major cycles which contain two or more minor cycles in them. However, we can prove significant drop in norm in major cycles which have at most one minor cycle in them. We call such major cycles *good*. The next easy, but very useful, observation is the following: one cannot have too many bad major cycles without having too many good major cycles. In our previous version, we had a weaker result of T/n. Lacoste-Julien and Jaggi observed that indeed $T/2$ suffices. We thank them for allowing us to include this result in this version of our paper.

Lemma 1. *[19] For any T, the first T iterations contain at least $T/2$ iterations with at least one good major cycle.*

Proof. This basically follow since two minor cycles decreases the size of S. But to decrease the size of S, there must be correspondingly two consecutive major cycles as well. □

Before proceeding, we introduce some notation.

Definition 1. *Given a point $x \in \mathcal{B}$, let us denote $\texttt{err}(x) := ||x||^2 - ||x_*||^2$. Given a point x and q, let $\Delta(x, q) := ||x||^2 - x^\top q$ and let $\Delta(x) := \max_{q \in \mathcal{B}} \Delta(x, q) = ||x||^2 - \min_{q \in \mathcal{B}} x^\top q$. Observe that $\Delta(x) \geq \texttt{err}(x)/2$ since $\Delta(x) \geq ||x||^2 - x^\top x_* \geq (||x||^2 - ||x_*||^2)/2$.*

We now use t to index all good major cycles. Let x_t be the point x at the beginning of the t-th good major cycle. The next theorem shows that the norm significantly drops across good major cycles. Recall D is the diameter of the polytope $\mathcal{B}$ which is contained in a ball of radius R.

Theorem 6. *Fix a good major cycle t, and define $\Delta = \min_x \Delta(x)$ where the minimum is taken over x_t and the point obtained after a minor cycle, if any. Then,*

$$\texttt{err}(x_t) - \texttt{err}(x_{t+1}) \geq \Delta^2/3D^2.$$

We now complete the proof of Theorem 4 using Theorem 6.

Proof of Theorem 4. Using Theorem 6, we get that $\texttt{err}(x_t) - \texttt{err}(x_{t+1}) \geq \texttt{err}(x_{t+1})^2/12D^2$ since $\Delta(x) \geq \texttt{err}(x)/2$ for all x and the error monotonically decreases. We claim that in $t^* \leq \frac{24D^2}{\varepsilon} + 8\log_2\left(\frac{R}{\varepsilon}\right)$ good major cycles, we reach x_t with $\texttt{err}(x_{t^*}) \leq \varepsilon$.

For simplicity, let $C := 12D^2$. Let $\delta := \texttt{err}(x_t) - \texttt{err}(x_{t+1})$. We get $\delta \geq (\texttt{err}(x_t) - \delta)^2/C$ which in turn implies $\delta(1 + \frac{2\texttt{err}(x_t)}{C}) \geq \frac{\texttt{err}(x_t)^2}{C}$. Dividing both sides by the paranthesized term of the LHS gives $\delta \geq \frac{\texttt{err}(x_t)^2}{C + 2\texttt{err}(x_t)}$. In summary, we get

$$\texttt{err}(x_{t+1}) \leq \texttt{err}(x_t)\left(1 - \frac{\texttt{err}(x_t)}{C + 2\texttt{err}(x_t)}\right)$$

Now let $e_0 := \texttt{err}(x_0)$. Note that $e_0 \leq R^2$. Define $t_0, t_1, \ldots$ such that for all $k \geq 1$ we have $\texttt{err}(x_t) > e_0/2^k$ for $t \in [t_{k-1}, t_k)$. That is, t_k is the first time t at which $\texttt{err}(x_t) \leq e_0/2^k$. Note that for $t \in [t_{k-1}, t_k)$, we have $\texttt{err}(x_{t+1}) \leq \texttt{err}(x_t)\left(1 - \frac{e_0}{2^k C + 4e_0}\right)$. This implies in $(2^k C/e_0 + 4)$ good major cycles after t_{k-1}, we will have $\texttt{err}(x_t) \leq \texttt{err}(x_{t_{k-1}})/2$; we have used the fact that $(1-\delta)^{1/\delta} < 1/2$ when $\delta < 1$. That is, $t_k \leq t_{k-1} + 2^k C/e_0 + 4$. We are interested in $t^* = t_K$ where $2^K = e_0/\varepsilon$. We get $t^* \leq \frac{C}{e_0}\left(1 + 2 + \cdots + 2^K\right) + 4K \leq \frac{2C}{\varepsilon} + 4\log_2(e_0/\varepsilon)$. Substituting $C = 12D^2$ and noting $e_0 \leq R^2$ gives the required statement.

Next, we claim that in $t^{**} < t^* + t'$ good major cycles, where $t' = 3D^2/\varepsilon$, we obtain an $x_{t^{**}}$ with $\Delta(x_{t^{**}}) \leq \varepsilon$. This is because, if not, then, using Theorem 6, in each of the good major cycles $t^* + 1, t^* + 2, \ldots t^* + t'$, $\texttt{err}(x)$ falls additively by $> \varepsilon^2/3D^2$ and thus $\texttt{err}(x_{t^*+t'}) < \texttt{err}(x_{t^*}) - \varepsilon \leq 0$, which is a contradiction. Therefore, in $O(D^2/\varepsilon + \log(R/\varepsilon))$ good major cycles, the algorithm obtains an $x = x_{t^{**}}$ with $\Delta(x) \leq \varepsilon$, proving Theorem 4. □

The rest of this subsection is dedicated to proving Theorem 6.

Proof of Theorem 6: We start off with a simple geometric lemma.

Lemma 2. *Let S be a subset of $\mathbb{R}^n$ and suppose y is the minimum norm point of $\texttt{aff}(S)$. Let x and q be* arbitrary *points in $\texttt{aff}(S)$. Then,*

$$||x - y||^2 = ||x||^2 - ||y||^2 \geq \frac{\Delta(x, q)^2}{||q - x||^2} \tag{2}$$

Proof. Since y is the minimum norm point in $\texttt{aff}(S)$, we have $x^\top y = q^\top y = ||y||^2$. In particular, $||x - y||^2 = ||x||^2 - ||y||^2$. Therefore,

$$\Delta(x, q) = \|x\|^2 - x^T q = \|x\|^2 - x^\top y + y^\top q - x^T q = (y - x)^T (q - x) \leq \|y - x\| \cdot \|q - x\|$$

by Cauchy-Schwartz. □

The above lemma takes case of major cycles with no minor cycles in them.

Lemma 3 (Progress in Major Cycle with no Minor Cycles)**.** *Let t be the index of a good major cycle with no minor cycles. Then* $\texttt{err}(x_t) - \texttt{err}(x_{t+1}) \geq \Delta^2(x_t)/D^2$.

Proof. Let S_t be the set S at start of the tth good major cycle, and let q_t be the point minimizing $x_t^\top q$. Let $S = S_t \cup q_t$ and let y be the minimum norm point in $\texttt{aff}(S)$. Since there are no minor cycles, $y \in \texttt{conv}(S)$. Abuse notation and let $x_{t+1} = y$ be the iterate at the call of the next major cycle (and not the next good major cycle). Since the norm monotonically decreases, it suffices to prove the lemma statement for this x_{t+1}. Now apply Lemma 2 with $x = x_t$ and $q = q_t$ and $S = S_t \cup q_t$. We have that $\texttt{err}(x_t) - \texttt{err}(x_{t+1}) = ||x_t||^2 - ||y||^2 \geq \Delta(x_t, q_t)^2/D^2 = \Delta(x_t)^2/D^2$. □

Now we have to argue about major cycles with exactly one minor cycle. The next observation is a useful structural result.

Lemma 4 (New Vertex Survives a Minor Cycle.)**.** *Consider any (not necessarily good) major cycle. Let x_t, S_t, q_t be the parameters at the beginning of this cycle, and let $x_{t+1}, S_{t+1}, q_{t+1}$ be the parameters at the beginning of the next major cycle. Then, $q_t \in S_{t+1}$.*

Proof. Clearly $S_{t+1} \subseteq S_t \cup q_t$ since q_t is added and then maybe minor cycles remove some points from S. Suppose $q_t \notin S_{t+1}$. Well, then $S_{t+1} \subseteq S_t$. But x_{t+1} is the affine minimizer of S_{t+1} and x_t is the affine minimizer of S_t. Since S_t is the larger set, we get $||x_t|| \leq ||x_{t+1}||$. This contradicts the strict decrease in the norm. □

Lemma 5 (Progress in an iteration with exactly one minor cyvle)**.** *Suppose the tth good major cycle has exactly one minor cycle. Let z_t be the current iterate after the minor cycle. Then,* $\texttt{err}(x_t) - \texttt{err}(x_{t+1}) \geq \min(\Delta(x_t), \Delta(z_t))^2/16D^2$.

Proof. Let x_t, S_t, q_t be the parameters at the beginning of the tth good major cycle. Let y_t be the affine minimizer of $S_t \cup q_t$. Since there is one minor cycle, $y_t \notin \texttt{conv}(S_t \cup q_t)$. $z_t = \theta x_t + (1-\theta)y$ is the intermediate x after the minor cycle, that is, the point in the line segment $[x_t, y]$ which lies in $\texttt{conv}(S_t \cup q_t)$. Let S' be the set after the single minor cycle is run. Since there is just one minor cycle, we get x_{t+1} (abusing notation once again since the next major cycle may not be good) is the affine minimizer of S'.

From Lemma 2 to S_t, x_t, q_t, and using the fact that q_t is the minimizer of $x_t^\top q$ over all q, we have:

$$||x_t - y_t|| \geq \Delta(x_t)/D \tag{3}$$

Applying Lemma 2 to S', z_t, q_t (we can do this since Lemma 4 implies $q_t \in S'$), we get

$$||z_t||^2 - ||x_{t+1}||^2 \geq \Delta(z_t, q_t)^2/D^2 \tag{4}$$

Next, we try to lower bound $\Delta(z_t, q_t)$ in terms of $\Delta(z)$. Let $q_z := \arg\min_{p \in \mathcal{B}} p^\top z$. Then

$$\begin{aligned}
\Delta(z_t, q_t) &= ||z_t||^2 - z_t^\top q_t \\
&= ||z_t||^2 - z_t^\top q_z + z_t^\top (q_z - q_t) \\
&\geq \Delta(z_t) + (z_t - x_t)^\top (q_z - q_t) &(5)\\
&\geq \Delta(z_t) - ||z_t - x_t|| \cdot ||q_z - q_t|| &(6)
\end{aligned}$$

Inequality (5) follows from definition of q_t, that is, $q_t = \arg\min_{p \in \mathcal{B}} p^\top x_t$, and so $x_t^\top q_z \geq x_t^\top q_t$. Inequality (6) is Cauchy-Schwartz.

Now we are armed to prove the lemma. To start with note that we may assume $||x_t||^2 - ||z_t|^2 \leq \Delta(x_t)\Delta(z_t)/3D^2$ for otherwise the theorem holds vacuously. Now we use the fact that $z_t = \theta x_t + (1-\theta)y_t$ for some $\theta \in [0, 1]$. Using the fact that $||x_t - y_t||^2 = ||x_t||^2 - ||y_t||^2$, we get that

$$(1-\theta^2)||x_t - y_t||^2 \leq \frac{\Delta(x_t)\Delta(z_t)}{3D^2} \tag{7}$$

Now we can upper bound $||x_t - z_t||$ as follows.

$$\begin{aligned} ||x_t - z_t|| &= (1-\theta)||x_t - y_t|| \\ &\leq (1-\theta^2)||x_t - y_t|| \\ &\leq \frac{\Delta(x_t)\Delta(z_t)}{3D^2} \cdot \frac{1}{||x_t - y_t||} \qquad \text{Due to (7)} \\ &\leq \frac{\Delta(z_t)}{3D} \qquad \text{Due to (3)} \end{aligned}$$

Substituting in (6) gives $\Delta(z_t, q_t) \geq 2\Delta(z_t)/3$ which in turn substituted in (4) gives $||z_t||^2 - ||x_{t+1}||^2 \geq \Delta(z_t)^2/3D^2$. This completes the proof. □

Lemma 3 and Lemma 5 complete the proof of Theorem 6.

3.2 A Robust version of Fujishige's Theorem

In this section we prove Theorem 5 which we restate below.

Theorem 5. *Fix a submodular function f with base polytope $\mathcal{B}_f$. Let $x \in \mathcal{B}_f$ be such that $||x||^2 \leq x^\top q + \varepsilon$ for all $q \in \mathcal{B}_f$. Renumber indices such that $x_1 \leq \cdots \leq x_n$. Then there exists a $1 \leq k \leq n$ such that $f(\{1, 2, \ldots, k\}) \leq \min_S f(S) + 2\sqrt{n\varepsilon}$. In particular, if $\varepsilon < \frac{1}{4n}$ and f is integer-valued, then S is a minimizer.*

Before proving the theorem, note that setting $\varepsilon = 0$ gives Fujishige's theorem Theorem 3.

Proof. We claim that the following inequality holds. Below, $[i] := \{1, \ldots, i\}$.

$$\sum_{i=1}^{n-1} (x_{i+1} - x_i) \cdot (f([i]) - x([i])) \leq \varepsilon \tag{8}$$

We prove the above shortly; for now we assume this and move on. Let i_z be the largest i such that $x_i \leq 0$. Let k be the smallest index $\geq i_z$ such that $f([k]) - x([k]) \leq \sqrt{n\varepsilon}$; k exists since $f([n]) - x([n]) = 0$. Note that $f([k]) \leq \sum_{i:x_i \leq 0} x_i + \sum_{i_z < i \leq k} x_i + \sqrt{n\varepsilon}$. We now bound the second summand. (8) implies $(x_k - x_{i_z}) \cdot \sqrt{n\varepsilon} \leq \varepsilon$, that is, $x_k \leq \sqrt{\frac{\varepsilon}{n}}$ since $x_{i_z} \leq 0$. In particular, $\sum_{i:0<x_i \leq x_k} x_i \leq n x_k \leq \sqrt{n\varepsilon}$. Together, we get $f([k]) \leq \sum_{i:x_i \leq 0} x_i + 2\sqrt{n\varepsilon}$. which implies the theorem due to Theorem 2.

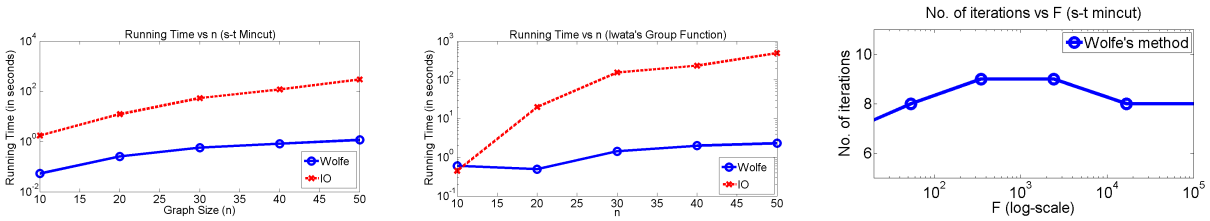
Now we prove (8). Let $z \in \mathcal{B}_f$ be the point which minimizes $z^\top x$. By the Greedy algorithm described in Section 2.1, we know that $z_i = f([i]) - f([i-1])$. Next, we write x in a different basis as follows: $x = \sum_{i=1}^{n-1} (x_i - x_{i+1}) \mathbf{1}_{[i]} + x_n \mathbf{1}_{[n]}$. Here $\mathbf{1}_{[i]}$ is used as the shorthand for the vector which has 1's in the first i coordinates and 0s everywhere else. Taking dot product with $(x - z)$, we get

$$||x||^2 - x^\top z = (x-z)^\top x = \sum_{i=1}^{n-1} (x_i - x_{i+1}) \left(x^\top \mathbf{1}_{[i]} - z^\top \mathbf{1}_{[i]}\right) + x_n \left(x^\top \mathbf{1}_{[n]} - z^\top \mathbf{1}_{[n]}\right) \tag{9}$$

Since $z_i = f([i]) - f([i-1])$, we get $x^\top \mathbf{1}_{[i]} - z^\top \mathbf{1}_{[i]}$ is $x([i]) - f([i])$. Therefore the RHS of (9) is the LHS of (8). The LHS of (9), by the assumption of the theorem, is at most ε implying (8). □

4 Discussion and Conclusions

We have shown that the Fujishige-Wolfe algorithm solves SFM in $O((n^3 \mathrm{EO} + n^4)F^2)$ time, where F is the maximum change in the value of the function on addition or deletion of an element. Although this is the first pseudopolynomial time analysis of the algorithm, we believe there is room for improvement and hope our work triggers more interest.



(a) (b) (c)

Figure 1: Running time comparision of Iwata-Orlin's (IO) method [11] vs Wolfe's method. (a): s-t mincut function, (b) Iwata's 3 groups function [16]. (c): Total number of iterations required by Wolfe's method for solving s-t mincut with increasing F

Note that our anlaysis of the Fujishige-Wolfe algorithm is weaker than the best known method in terms of time complexity (IO method by [11]) on the dependence on F. In contrast, we found this algorithm significantly outperforming the IO algorithm empirically – we show two plots here. In Figure 1 (a), we run both on Erdos-Renyi graphs with $p = 0.8$ and randomly chosen s, t nodes. In Figure 1 (b), we run both on the Iwata group functions [16] with 3 groups. Perhaps more interestingly, in Figure 1 (c), we ran the Fujishige-Wolfe algorithm on the simple path graph where s, t were the end points, and changed the capacities on the edges of the graph which changed the parameter F. As can be seen, the number of iterations of the algorithm remains constant even for exponentially increasing F.

References

[1] F. Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2–3):145–373, 2013. 2

[2] Francis Bach. Convex analysis and optimization with submodular functions: a tutorial. *CoRR*, abs/1010.4207, 2010. 1

[3] Jack Edmonds. Matroids, submodular functions and certain polyhedra. *Combinatorial Structures and Their Applications*, pages 69–87, 1970. 2

[4] Satoru Fujishige. Lexicographieally optimal base of a polymatroid with respect to a weight vector. *Math. Oper. Res.*, 5:186–196, 1980. 1, 2, 3

[5] Satoru Fujishige. Submodular systems and related topics. *Math. Programming Study*, 1984. 2

[6] Satoru Fujishige, Takumi Hayashi, and Shigueo Isotani. The minimum-norm-point algorithm applied to submodular function minimization and linear programming. 2006. 2

[7] Satoru Fujishige and Shigueo Isotani. A submodular function minimization algorithm based on the minimum-norm base. *Pacific Journal of Optimization*, 7:3, 2011. 2

[8] Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981. 1

[9] Satoru Iwata, Lisa Fleischer, and Satoru Fujishige. A combinatorial, strongly polynomial-time algorithm for minimizing submodular functions. In *STOC*, pages 97–106, 2000. 1

[10] Satoru Iwata, Lisa Fleischer, and Satoru Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *J. ACM*, 48(4):761–777, 2001. 1

[11] Satoru Iwata and James B. Orlin. A simple combinatorial algorithm for submodular function minimization. In *SODA*, pages 1230–1237, 2009. 1, 2, 8

[12] Rishabh Iyer, Stefanie Jegelka, and Jeff Bilmes. Curvature and optimal algorithms for learning and minimizing submodular functions. *CoRR*, abs/1311.2110, 2013. 1

[13] Rishabh Iyer, Stefanie Jegelka, and Jeff Bilmes. Fast semidifferential-based submodular function optimization. In *ICML (3)*, pages 855–863, 2013. 1

[14] Rishabh K. Iyer and Jeff A. Bilmes. Submodular optimization with submodular cover and submodular knapsack constraints. In *NIPS*, pages 2436–2444, 2013. 1

[15] Stefanie Jegelka, Francis Bach, and Suvrit Sra. Reflection methods for user-friendly submodular optimization. In *NIPS*, pages 1313–1321, 2013. 1

[16] Stefanie Jegelka, Hui Lin, and Jeff A. Bilmes. On fast approximate submodular minimization. In *NIPS*, pages 460–468, 2011. 1, 8

[17] Pushmeet Kohli and Philip H. S. Torr. Dynamic graph cuts and their applications in computer vision. In *Computer Vision: Detection, Recognition and Reconstruction*, pages 51–108. 2010. 1

[18] Andreas Krause, Ajit Paul Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008. 1

[19] S. Lacoste-Julien and M. Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *Adv. in Neu. Inf. Proc. Sys. (NIPS)*, 2015. 4

[20] Alexander Schrijver. A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *J. Comb. Theory, Ser. B*, 80(2):346–355, 2000. 1

[21] Peter Stobbe and Andreas Krause. Efficient minimization of decomposable submodular functions. In *NIPS*, pages 2208–2216, 2010. 1

[22] Phillip Wolfe. Finding the nearest point in a polytope. *Math. Programming*, 11:128 – 149, 1976. 1, 2, 3