

LIMITED MEMORY KELLEY’S METHOD CONVERGES FOR COMPOSITE CONVEX AND SUBMODULAR OBJECTIVES

SONG ZHOU*, SWATI GUPTA†, AND MADELEINE UDELL‡

Abstract. The original simplicial method (OSM), a variant of the classic Kelley’s cutting plane method, has been shown to converge to the minimizer of composite convex and submodular objectives, though no rate of convergence for this method was known. Moreover, OSM is required to solve subproblems in each iteration whose size grows linearly in the number of iterations. We propose a limited memory version of Kelley’s method (L-KM) that is a novel adaptation of OSM and requires limited memory independent of the iteration (at most $n + 1$ constraints for an n -dimensional problem), while maintaining convergence to the optimal solution. We further show that the dual method of L-KM is a special case of the Fully-Corrective Frank-Wolfe (FCFW) method with approximate correction, thereby deriving a limited memory version of FCFW method and proving a rate of convergence for L-KM. Though we propose L-KM for minimizing composite convex and submodular objectives, our results on limited memory version of FCFW hold for general polytopes, which is of independent interest.

Key words. Kelley’s cutting plane method, submodular functions, Lovász extension, Fully corrective Frank-Wolfe, limited memory simplicial method

AMS subject classifications. 90C25, 90C27, 90C30

1. Introduction. One of the earliest and fundamental methods to minimize non-smooth convex objectives is Kelley’s method, which minimizes the maximum of lower bounds on the convex function given by the supporting hyperplanes to the function at each previously queried point. An approximate solution to the minimization problem is found by minimizing this piecewise linear approximation, and the approximation is then strengthened by adding the supporting hyperplane at the current approximate solution [10, 5]. Many variants of Kelley’s method have been analyzed in the literature [15, 11, 7, for e.g.]. Kelley’s method and its variants are a natural fit for problem involving a piecewise linear function, such as composite convex and submodular objectives. This paper defines a new limited memory version of Kelley’s method adapted to composite convex and submodular objectives, and establishes the first convergence rate for such a method, solving the open problem proposed in [2, 3].

Submodularity is a discrete analogue of convexity and has been used to model combinatorial constraints in a wide variety of machine learning applications, such as MAP inference, document summarization, sensor placement, clustering, and image segmentation [3, and references therein]. Submodular set functions are defined with respect to a ground set of elements V , which we may identify with $\{1, \dots, n\}$ where $|V| = n$. These functions capture the property of diminishing returns: $F : \{0, 1\}^n \rightarrow \mathbb{R}$ is said to be submodular if $F(A \cup \{e\}) - F(A) \geq F(B \cup \{e\}) - F(B)$ for all $A \subseteq B \subseteq V$, $e \notin B$. Lovász gave a convex extension $f : [0, 1]^n \rightarrow \mathbb{R}$ of the submodular set functions which takes the value of the set function on the vertices of the $[0, 1]^n$ hypercube, i.e. $f(\mathbf{1}_S) = F(S)$, where $\mathbf{1}_S$ is the indicator vector of the set $S \subseteq V$ [16]. (See Section 2 for details.)

In this work, we consider the application of a variant of Kelley’s method, the Original Simplicial Method (OSM), to minimize composite convex and submodular objectives

$$(P) \quad \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad g(x) + f(x),$$

where $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is a closed convex function and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the Lovász extension of a given submodular function F . Composite convex and submodular objectives have been used extensively in sparse learning, where the support of the model must satisfy certain combinatorial constraints. OSM was proposed by Bach [3] to minimize such composite objectives. At the i th iteration, it approximates the Lovász extension by a piecewise linear function $f_{(i)}$ whose epigraph is the maximum of the supporting hyperplanes to the function at each previously queried point. Using a minimizer $x^{(i)}$ of $g + f_{(i)}$ and a sub-gradient at $x^{(i)}$, OSM continues with a stronger approximation $f_{(i+1)}$ of the submodular part. It is natural to approximate the submodular part of the objective by a piecewise linear function, since the Lovász extension is piecewise linear

*Tsinghua University, sz557@cornell.edu

† Georgia Institute of Technology, swatig@gatech.edu

‡ Cornell University, ude11@cornell.edu

(with possibly an exponential number of pieces). In OSM, once the algorithm reaches the optimal solution, it terminates, in contrast to sub-gradient methods that might continue to oscillate. Note that the classic Kelley’s Method approximates the full objective function using a piecewise linear function, while OSM only approximates the Lovász extension f and keeps the rest of the composite objective accurate.

In [3], the authors show that OSM converges to the optimum; however, no rate of convergence is given. Moreover, OSM maintains an approximation of the Lovász extension by maintaining a set of linear constraints whose size grows linearly with the number of iterations: OSM does not support deletion of previously obtained cutting planes. Hence the subproblems are harder to solve with each iteration. To address these challenges, we introduce a limited memory version of Kelley’s Method, L-KM, for composite convex and submodular objectives. We show that we can restrict the number of linear constraints used in each approximation $f_{(i)}$ to at most $n + 1$, while preserving convergence.

We next consider DL-KM, the dual method of L-KM, and show that it is a special case of Fully-Corrective Frank-Wolfe (FCFW) with approximate correction [14]. Although the dual method of OSM has been studied in the literature, it is not a special case of FCFW, since it does not support deletion of linear constraints. This connection immediately implies two results: (i) The dual method DL-KM converges linearly while maintaining a set of at most $(n + 1)$ vertices. Hence we obtain a limited memory version of FCFW. (ii) The primal method L-KM has finite convergence. Finally, we remark that the dual method DL-KM applies to general polytopes as well, giving an improved, limited memory, version of FCFW for general polytopes.

1.1. Related Work. The Original Simplicial Method was proposed by Bach (2013) [3]. Before the present work, there was no known rate of convergence of this method, although it had been shown to converge finitely [3]. In 2015, Lacoste-Julien and Jaggi proved global linear convergence of variants of the Frank-Wolfe algorithm, including the Fully Corrective Frank-Wolfe (FCFW) with approximate correction [14]. DL-KM, proposed in this paper, can be shown to be a limited memory special case of the latter (whereas OSM is not), and therefore helps prove convergence of a limited memory version of OSM (and thus, a rate of convergence for the OSM).

Many authors have studied convergence guarantees and reduced memory requirements for variants of Kelley’s method [10, 5]. Unless these variants allow approximation of only part of the objective, they are computationally disadvantaged compared to OSM. Among the earliest work on bounded storage in proximal level bundle methods is a paper by Kiwiel (1995) [11]. This method projects iterates onto successive approximations of the level set of the objective; however, unlike our method, it is sensitive to the choice of parameters (level sets) and oblivious to any simplicial structure: iterates are often not extreme points of the epigraph of the function. Subsequent work on the proximal setup uses trust regions, penalty functions, level sets, and other more complex algorithmic tools than OSM; we refer the reader to [17] for a survey on bundle methods. For the dual problem, a paper by Von Hohenbalken (1977) [21] shares some elements of our proof techniques, however their results only apply to differential problems and do not bound the memory. Another restricted simplicial decomposition method was proposed by Hearn et al (1987) [9], in which the constraint set size is limited by user-defined parameters (e.g., $r = 1$ reduces to the Frank-Wolfe algorithm [8]): it can replace an atom with minimal weight in the current convex combination with a prior iterate of the algorithm, which may be strictly inside the feasible region. In contrast, DL-KM obeys a known upper bound $(n + 1)$ on the number of constraints, and hence requires no tuning of parameters. Moreover, all constraints are vertices of the feasible set.

1.2. Applications. Composite convex and submodular objectives have gained popularity over the last few years in a large number of machine learning applications such as structured regularization or empirical risk minimization [4]: $\min_{w \in \mathbb{R}^n} \sum_i l(y_i, w^\top x_i) + \lambda \Omega(w)$, where w are the model parameters and $\Omega : \mathbb{R}^n \rightarrow \mathbb{R}$ is a regularizer. The Lovász extension can be used to obtain a convex relaxation of a regularizer that penalizes the support of the solution w to achieve structured sparsity, which improves model interpretability or encodes knowledge about the domain. For example, fused regularization uses $\Omega(w) = \sum_i |w_i - w_{i+1}|$, which is the Lovász extension of the generalized cut function, and group regularization uses $\Omega(w) = \sum_g d_g \|w_g\|_\infty$, which is the Lovász extension of the coverage submodular function. (See Appendix A, Table 1 for details on these and other submodular functions.)

Furthermore, minimizing a composite convex and submodular objective is dual to minimizing convex objectives over a submodular polytope (under mild conditions). This duality is central to the present

work. First-order projection based methods like online stochastic mirror descent and its variants require the computation of a Bregman projection $\min_{x \in P} \omega(x) + \nabla \omega(y)^\top (x - y)$ to minimize a strictly convex functions $\omega(\cdot)$ over the set $P \subseteq \mathbb{R}^n$. The computation of this projection becomes a bottleneck in the practical implementation of these methods, though this class of algorithms is known to obtain near optimal convergence guarantees in various settings [19, 1]. The DL-KM used for computing Bregman projections leads to computational speed ups in variants of online mirror descent used for learning over spanning trees to reduce communication delays in networks, [12]), permutations to model scheduling delays [23], and k-sets for principal component analysis [22], to give a few examples of submodular online learning problems. Other example applications of convex minimization over submodular polytopes include computation of densest subgraphs [18], computation of a lower bound for the partition function of log-submodular distributions [6] and distributed routing [13].

1.3. Summary of contributions. We discuss background and the problem formulations in Section 2. Section 4 describes L-KM, our proposed limited memory version of OSM. We show that L-KM computes approximations to the Lovász extension by using at most $n + 1$ cutting planes for an n -dimensional problem and converges to the optimum. Section 5 shows that the dual of L-KM is a special case of Fully-Corrective Frank-Wolfe (FCFW) and proves sublinear convergence of L-KM and a linear convergence of DL-KM. These results also lead to a novel limited memory version of FCFW which is computationally faster. We present preliminary experiments in Section 6 that highlight the limited memory usage of both L-KM and DL-KM and show that their performance compares favorably with OSM and FCFW respectively.

2. Background. Consider a ground set V of n elements on which the submodular function $F : 2^V \rightarrow \mathbb{R}$ is defined. The function F is said to be submodular if $F(A) + F(B) \geq F(A \cup B) + F(A \cap B)$ for all $A, B \subseteq V$. This is equivalent to the diminishing marginal returns characterization mentioned before. Without loss of generality, we assume $F(\emptyset) = 0$ (otherwise one can set $F'(S) = F(S) - F(\emptyset)$). For $w \in \mathbb{R}^n$, $A \subseteq V$, we define $w(A) = \sum_{k \in A} w(k) = \mathbf{1}_A^\top w$, where $\mathbf{1}_A \in \mathbb{R}^n$ is the indicator vector of A , and let both $w(k)$ and w_k denote the k th element of x .

Given a submodular set function $F : V \rightarrow \mathbb{R}$, the submodular polyhedron and the base polytope are defined as $P(F) = \{w \in \mathbb{R}^n : w(A) \leq F(A), \forall A \subseteq V\}$, and $B(F) = \{w \in \mathbb{R}^n : w(V) = F(V), w \in P(F)\}$, respectively. We use $\text{vert}(B(F))$ to denote the vertex set of $B(F)$. The Lovász extension of F is the piecewise linear function [16]

$$(2.1) \quad f(x) = \max_{w \in B(F)} x^\top w.$$

The Lovász extension can be computed using Edmonds' greedy algorithm for maximizing linear functions over the base polytope (in $O(n \log n + n\gamma)$ time, where γ is the time required to compute the submodular function value). This extension can be defined for any set function, however it is convex if and only if the set function is submodular [16]. We call a permutation π over $[n]$ *consistent* with $x \in \mathbb{R}^n$ if $x_{\pi_i} \geq x_{\pi_j}$ whenever $i \leq j$. Each permutation π corresponds to an extreme point $w_{\pi_k} = F(\{\pi_1, \pi_2, \dots, \pi_k\}) - F(\{\pi_1, \pi_2, \dots, \pi_{k-1}\})$ of the base polytope. Note that the Lovász extension can also be written as

$$f(x) = \sum_k x_{\pi_k} [F(\{\pi_1, \pi_2, \dots, \pi_k\}) - F(\{\pi_1, \pi_2, \dots, \pi_{k-1}\})]$$

where π is a permutation consistent with x and $F(\emptyset) = 0$ by assumption. For $x \in \mathbb{R}^n$, let $\mathcal{V}(x)$ be the set of vertices $B(F)$ that correspond to permutations consistent with x . Note that $\partial f(x) = \text{conv}(\mathcal{V}(x))$, where $\partial f(x)$ is the subdifferential of f at x .

Let F be a non-decreasing submodular function. The symmetric submodular polytope of F is defined as $|P|(F) = \{x \in \mathbb{R}^n : |x| \in P(F)\}$, where $|x|$ stands for the element-wise absolute value of x . Similar to the Lovász extension of F , we have

$$(2.2) \quad f(|x|) = \max_{w \in |P|(F)} x^\top w$$

and $\partial f(|x|) = \text{conv}(\tilde{\mathcal{V}}(x)) := \text{conv}(\{|w| \in \mathcal{V}(x) \text{ such that } w(i)x(i) \geq 0 \text{ for all elements } i\})$.

Given a convex function $g : \mathbb{R}^n \rightarrow \mathbb{R}$, its Fenchel conjugate $g^* : \mathbb{R}^n \rightarrow \mathbb{R}$ is defined as

$$(2.3) \quad g^*(w) \triangleq \max_{x \in \mathbb{R}^n} (w)^\top x - g(x).$$

When g is convex and closed, $g^{**} = g$. Fenchel conjugates are always convex, regardless of the convexity of the original function. Another useful property of Fenchel conjugates is

$$(2.4) \quad w \in \partial g(x) \iff g(x) + g^*(w) = w^\top x,$$

where $\partial g(x)$ is the subdifferential of g at x . We use superscripts $(*)$ to represent Fenchel conjugation and (\star) to represent optimal solutions.

3. Problem Formulation. We are interested in two (dual [3]¹) canonical forms of convex minimization problems involving a given submodular function $F : 2^n \rightarrow \mathbb{R}$. The primal problem is to minimize a composite convex and submodular objective:

$$(3.1) \quad \text{minimize } g(x) + f(x) \quad \text{subject to } x \in \mathbb{R}^n$$

where $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is a closed² convex function, and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the Lovász extension of F . Since indicator functions of convex sets are also convex functions, the primal problem can model constrained minimization problems by setting g to be the corresponding indicator functions. The dual problem is to minimize a convex function over the submodular base polytope $B(F)$:

$$(D) \quad \text{minimize } h(w) \quad \text{subject to } w \in B(F)$$

where $h : \mathbb{R}^n \rightarrow \mathbb{R}$ is a convex function. Solving the primal form (3.1) using general convex optimization methods is challenging due to the non-differentiability of f at any point x that has at least two elements of the same value. On the other hand, in the case of the dual form, the number of facets of $B(F)$ could be exponential in the dimension of the polytope. In fact, there exists a family of matroids³ defined over ground set of size n , such that the extension complexity of the corresponding base polytope is $\Omega(2^{n/2}/n^{5/4}\sqrt{\log(2n)})$ [20], i.e. even after adding polynomial number of variables there is no representation of some submodular base polytopes that use a polynomial number of facets.

As shown in [3], using the definitions of the Lovász extension and the Fenchel conjugates, it is easy to see that weak duality always holds between problem (3.1) and (D).

LEMMA 3.1 (Weak Duality). *Let f be the Lovász extension of a submodular function $F : 2^V \rightarrow \mathbb{R}$ with $|V| = n$ and let $g : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex function. We have $p^* = \text{minimize}_{x \in \mathbb{R}^n} g(x) + f(x) \geq d^*$, where*

$$(3.2) \quad d^* = \text{maximize } -g^*(-w) \quad \text{subject to } w \in B(F)$$

Proof. We first have

$$(3.3) \quad \begin{aligned} & \min_{x \in \mathbb{R}^n} g(x) + f(x) \\ &= \min_{x \in \mathbb{R}^n} g(x) + \max_{w \in B(F)} w^\top x \\ &= \min_{x \in \mathbb{R}^n} \max_{w \in B(F)} g(x) + w^\top x. \end{aligned}$$

¹[3] mentions that strong duality holds between the primal and the dual forms, though the discussion omits the requirement that the solution be attained; hence we include a proof here for completeness.

²A function $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be closed iff $\forall \alpha \in \mathbb{R}, \{x \in \text{dom}(g) : f(x) \leq \alpha\}$ is a closed set.

³Let $r : 2^E \rightarrow \mathbb{Z}$ be a non-negative monotone submodular function that satisfies $r(S) \leq |S|$ for all $S \subseteq E$. Then, the collection of subsets $\mathcal{I} = \{A \subseteq E \mid r(A) = |A|\}$ is called a matroid.

For any given $w^h \in B(F)$, we also have $\min_{x \in \mathbb{R}^n} \max_{w \in B(F)} g(x) + w^\top x \geq \min_{x \in \mathbb{R}^n} g(x) + w^h{}^\top x$. Thus by the definition of g^* , we can see that

$$\begin{aligned}
(3.4) \quad \min_{x \in \mathbb{R}^n} \max_{w \in B(F)} g(x) + w^\top x &\geq \max_{w \in B(F)} \min_{x \in \mathbb{R}^n} g(x) + w^\top x \\
&= \max_{w \in B(F)} - \max_{x \in \mathbb{R}^n} (-w)^\top x - g(x) \\
&= \max_{w \in B(F)} -g^*(-w).
\end{aligned}$$

Combine (3.3) and (3.4), and the theorem follows. \square

Note that problems (D) and (D') have the same form. We refer to both as the dual problem. To argue strong duality, we require the following lemma that characterizes optimality conditions in terms of the intersection of the sub-differentials of f and $-g$. The proof follows from (2.4) and weak duality.

LEMMA 3.2. *Let $X^* = \arg \min_{x \in \mathbb{R}^n} g(x) + f(x)$ and $W^* = \arg \max_{w \in B(F)} -g^*(-w)$ be the set of optimal solutions to the primal and dual problems respectively. When X^* is nonempty, for any $x^* \in X^*$, the negative of $\partial g(x^*)$ has a non-empty intersection with $\partial f(x^*)$: $-\partial g(x^*) \cap \partial f(x^*) \neq \emptyset$. Moreover, any vector in this intersection is a solution to the dual problem: $-\partial g(x^*) \cap \partial f(x^*) \subseteq W^*$.*

Proof. Due to optimality of x^* for (3.1), $0 \in \partial g(x^*) + \partial f(x^*)$. Since both g and f are convex, $\partial g(x^*)$ and $\partial f(x^*)$ are non-empty. This implies $-\partial g(x^*) \cap \partial f(x^*) \neq \emptyset$. Consider $\tilde{w} \in -\partial g(x^*) \cap \partial f(x^*)$. Using (2.4) from preliminaries, we get $-g^*(-\tilde{w}) = g(x^*) + \tilde{w}^\top x^* \stackrel{(a)}{=} g(x^*) + f(x^*)$, where (a) follows from $\tilde{w} \in \partial f(x^*)$. Using weak duality in Lemma 3.1, we get $\tilde{w} \in W^*$. \square

It further follows from Lemma 3.2 that strong duality holds whenever there exists an optimal solution to the primal or the dual problem, as $\emptyset \neq -\partial g(x^*) \cap \partial f(x^*) \subseteq W^*$, and thus W^* is also attained.

LEMMA 3.3 (Strong Duality). *Whenever a solution $x^* \in \arg \min_x g(x) + f(x)$ exists, then some $w^* \in \arg \max_{w \in B(F)} -g^*(-w)$ also exists and strong duality holds: $g(x^*) + f(x^*) = -g^*(-w^*)$.*

As a corollary of Lemma 3.2, one can construct a solution to the dual problem given a solution to the primal problem at which g or f is differentiable.

COROLLARY 3.4 (Dual solution). *Whenever there exists an optimal solution x^* to the primal problem (3.1) such that $|\partial g(x^*)| = 1$ or there exists a unique permutation π consistent with x^* i.e. $|\mathcal{V}(x^*)| = 1$, then $w^* = -\nabla g(x^*)$ in the first case, or $w^* = \nabla f(x^*) \in \mathcal{V}(x^*)$ in the second case, is an optimal solution to the dual problem (3.2).*

4. Limited Memory Simplicial Method. We now present our novel limited memory adaptation L-KM of the Original Simplicial Method (OSM). Throughout this section, we assume that a solution to the primal problem (3.1) is attained, and strong duality thus holds. (i.e. $\min_x g(x) + f(x) = \max_{w \in B(F)} -g^*(-w)$). We first briefly review OSM as proposed by Bach [3, Section 7.7] and discuss problems of OSM with respect to memory requirements and the rate of convergence. We then highlight the changes in OSM that enable us to show a bound on the memory requirements while maintaining finite convergence.

4.1. The Original Simplicial Method. To solve the primal problem $\minimize_x g(x) + f(x)$ (3.1), it is natural to approximate the piecewise linear Lovász extension f with cutting planes derived from the values and sub-gradients of the function at previous iterations, which results in piecewise linear approximations of f . This is the basic idea of OSM introduced by Bach in [3]. This approach contrasts with Kelley's Method, which approximates the entire objective function $g + f$. OSM adds a cutting plane to the approximation of f at each iteration, so the number of the linear constraints in its subproblems grows linearly with the number of iterations. Hence it becomes increasingly challenging to solve the subproblem as the number of iterations grows up. Further, in spite of a finite convergence, as mentioned in the introduction there was no known rate of convergence for OSM or its dual method prior to this work. We include a complete description of OSM in Algorithm B.1 in the appendix.

4.2. Limited memory Kelley's Method. To tackle these two problems of memory requirements and lack of convergence rate, we introduce a novel limited memory version L-KM of OSM which ensures that the

Algorithm 4.1 L-KM: The Limited Memory Kelley's Method

1: **Input:** a convex function $g : \mathbb{R}^n \rightarrow \mathbb{R}$ and a submodular function $F : 2^n \rightarrow \mathbb{R}$, suboptimality tolerance parameter $\epsilon > 0$

2: **Output:** an approximate solution x^\sharp s.t. $g(x^\sharp) + f(x^\sharp) - \epsilon \leq \min_{x \in \mathbb{R}^n} g(x) + f(x)$

3: **Initialization:** let $\emptyset \subset \mathcal{V}^{(0)} \subseteq \text{vert}(B(F))$ such that vectors in $\mathcal{V}^{(0)}$ are affinely independent, $p^{(0)} := \infty$, $d^{(0)} := -\infty$, $\Delta^{(0)} := p^{(0)} - d^{(0)}$, $i = 1$

4: **while** $\Delta^{(i-1)} > \epsilon$ **do**

5: let $f_{(i)}(x) := \max_{w \in \text{conv}(\mathcal{V}^{(i-1)})} w^\top x$ ▷ update the approximation of f

6: let $x^{(i)} \in \arg \min_{x \in \mathbb{R}^n} g(x) + f_{(i)}(x)$ ▷ solve the i th subproblem

7: let $p^{(i)} := g(x^{(i)}) + f(x^{(i)})$ ▷ update the upper bound

8: let $d^{(i)} := g(x^{(i)}) + f_{(i)}(x^{(i)})$ ▷ update the lower bound

9: **if** $p^{(i)} \leq p^{(i-1)}$ **then** $z^{(i)} = x^{(i)}$ ▷ update the solution*

10: **else** $z^{(i)} = z^{(i-1)}$ ▷ do not update the solution*

11: let $\Delta^{(i)} := p^{(i)} - d^{(i)}$ ▷ update the optimality gap

12: let $\mathcal{A}^{(i)} := \{w \in \mathcal{V}^{(i-1)} : w^\top x^{(i)} = f_{(i)}(x^{(i)})\}$ ▷ set of tight vectors with respect to $x^{(i)}$ *

13: compute $v^{(i)} \in \mathcal{V}(x^{(i)})$

14: let $\mathcal{V}^{(i)} := \mathcal{A}^{(i)} \cup \{v^{(i)}\}$ ▷ update the set of sub-gradients of f^*

15: $i = i + 1$

16: **return** $z^{(i-1)}$

number of cutting planes maintained by the algorithm at any iteration is bounded by $n + 1$. L-KM only maintains the linear constraints that are tight at the solution $x^{(i)}$ to the current subproblem $\min g(x) + f_{(i)}(x)$. This thrift bounds the size of the subproblems at any iteration, thereby making L-KM cheaper to implement. We describe L-KM in detail in Algorithm 4.1 and mark the differences⁴ compared to OSM with *.

When g is not strongly convex, the first subproblem (Line 6 of Algorithm 4.1) could be unbounded. When g is bounded below, to solve this problem, we can construct $\mathcal{V}^{(0)}$ such that $f_{(1)}$ is bounded, which makes the first subproblem bounded. By Corollary 4.3, this implies that all the subsequent subproblems are bounded as well. We use $z^{(i)}$ and $p^{(i)}$ to keep track of the best iterate and its value so far. L-KM is not a descent method: the objective value at iterate $x^{(i)}$ can increase compared to the objective value at $x^{(i-1)}$ (see e.g. Figure 1 (b), (c)).

The main difference between L-KM and OSM is that, as shown in Lines 12 and 14 of Algorithm 4.1, only vectors $w \in \mathcal{V}^{(i-1)}$ that maximize $w^\top x^{(i)}$ at the current solution $x^{(i)}$ are kept in $\mathcal{V}^{(i)}$ as opposed to OSM that increases $\mathcal{V}^{(i)} \supseteq \mathcal{V}^{(i-1)}$. Since L-KM considers a smaller set of vectors compared to OSM, it is not obvious how the total number of iterations needed for convergence changes, or even whether L-KM converges. Surprisingly, we are able to show that the solution $x^{(i)}$ of the subproblem does not change when we limit the size of this set, as we show in the following lemma. This lemma will be crucial in showing that limiting the set $\mathcal{V}^{(i)}$ in each iteration does not affect the convergence of L-KM. The proof of this lemma follows by observing that $x^{(i)}$ remains locally optimal for $\tilde{P}_{(i)}$ due to preservation of the outer normal cone supported at $w^\sharp \in \mathcal{V}^{(i-1)}$ such that $f_{(i)}(x^{(i)}) = \max_{w \in \text{conv}(\mathcal{V}^{(i-1)})} w^\top x = w^{\sharp T} x^{(i)}$.

LEMMA 4.1. *Let $x^{(i)}$ be a solution to the i th approximation of the composite objective, $P_{(i)}(x) \triangleq \min g(x) + \max_{w \in \text{conv}(\mathcal{V}^{(i-1)})} w^\top x = g(x) + f_{(i)}(x)$ then $x^{(i)}$ is also a solution to the problem $\tilde{P}_{(i)} \triangleq \min_{x \in \mathbb{R}^n} g(x) + \max_{w \in \text{conv}(\mathcal{A}^{(i)})} w^\top x$, where $\mathcal{A}^{(i)} \triangleq \{w \in \mathcal{V}^{(i-1)} : w^\top x^{(i)} = f_{(i)}(x^{(i)})\}$ is the set of vectors in $\mathcal{V}^{(i-1)}$ that attain $f_{(i)}(x^{(i)})$. Moreover, $P_{(i)}$ and $\tilde{P}_{(i)}$ have the same optimal value.*

Proof. There exists at least one $w^\sharp \in \mathcal{V}^{(i-1)}$ such that $f_{(i)}(x^{(i)}) = w^{\sharp T} x^{(i)}$. Therefore, $P_{(i)}(x^{(i)}) = g(x^{(i)}) + w^{\sharp T} x^{(i)} = g(x^{(i)}) + \max_{w \in \text{conv}(\mathcal{A}^{(i)})} w^\top x^{(i)} = \tilde{P}_{(i)}(x^{(i)})$, where the last equality follows from the definition of $\mathcal{A}^{(i)}$. Next, if we can show local optimality of $x^{(i)}$ for $\tilde{P}_{(i)}$, this would imply global optimality

⁴OSM is recovered by deleting steps marked with asterisk and setting $\mathcal{V}^{(i)} \triangleq \mathcal{V}^{(i-1)} \cup \{v^{(i)}\}$ in step 16.

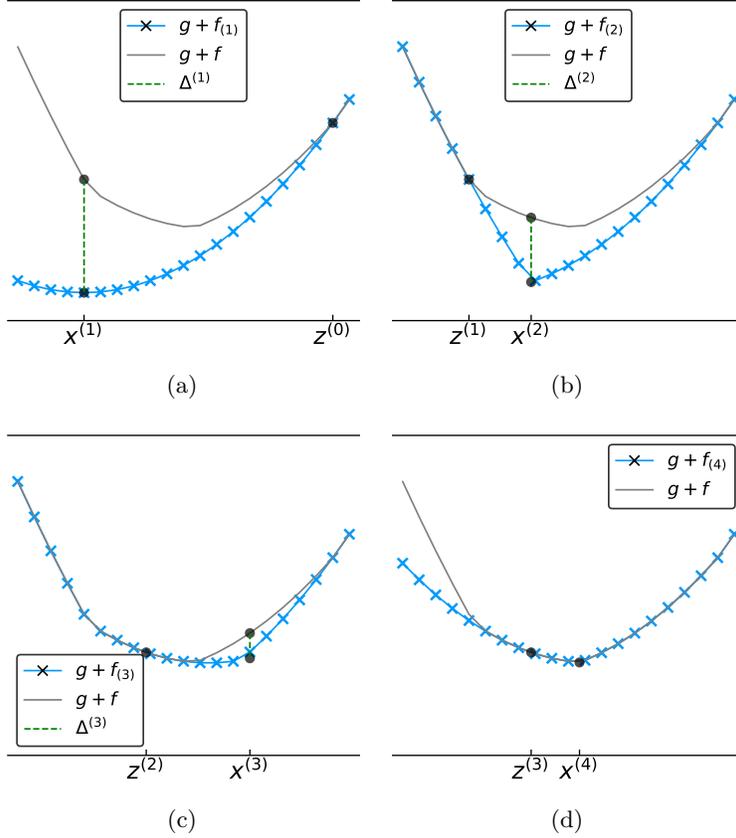


Fig. 1: An illustration of L-KM: blue curve denotes the i th function approximation $g + f_{(i)}$. In (d), note that L-KM approximation $g + f_{(4)}$ (marked with \times) is obtained by dropping the leftmost constraint in $g + f_{(3)}$ (in (c)), unlike OSM.

of $x^{(i)}$ for $\tilde{P}_{(i)}$ due to convexity of $\tilde{P}_{(i)}$. Suppose for a contradiction that for any given $\epsilon > 0$, there exists a ball $B(x^{(i)}, \epsilon)$ that contains y such that $\tilde{P}_{(i)}(y) < \tilde{P}_{(i)}(x^{(i)})$. That is, $g(y) + \max_{w \in \text{conv}(\mathcal{A}^{(i)})} w^\top y < g(x^{(i)}) + \max_{w \in \text{conv}(\mathcal{A}^{(i)})} w^\top x^{(i)} \leq g(y) + \max_{w \in \text{conv}(\mathcal{V}^{(i-1)})} w^\top y$, i.e. $\exists y \in B(x^{(i)}, \epsilon)$ such that $\max_{w \in \text{conv}(\mathcal{A}^{(i)})} w^\top y < \max_{w \in \text{conv}(\mathcal{V}^{(i-1)})} w^\top y$. However, $x^{(i)}$ must be strictly contained in the outer normal cone of $\text{conv}(\mathcal{V}^{(i-1)})$ supported at w^\dagger and in a small ball around $x^{(i)}$, $\max_{w \in \text{conv}(\mathcal{A}^{(i)})} w^\top y = \max_{w \in \text{conv}(\mathcal{V}^{(i-1)})} w^\top y$. \square

We next show that the sequence of lower bounds obtained by L-KM is non-decreasing. This is a direct result of Lemma 4.1 and the inclusion relation $\mathcal{A}^{(i)} \subseteq \mathcal{V}^{(i)} \subseteq \mathcal{V} = \text{vert}(B(F))$.

COROLLARY 4.2. *The sequence of lower bounds $\{d^{(i)}\}$ constructed by L-KM are non-decreasing, i.e. $d^{(i-1)} \leq d^{(i)}$ for all $i \geq 1$.*

Proof. We have

$$\begin{aligned}
(4.1) \quad d^{(i)} &= g(x^{(i)}) + f_{(i)}(x^{(i)}) && \triangleright \text{definition of } d^{(i)} \\
&= g(x^{(i)}) + \max_{w \in \text{conv}(\mathcal{V}^{(i-1)})} w^\top x^{(i)} && \triangleright \text{definition of } f_{(i)} \\
&\leq g(x^*) + \max_{w \in \text{conv}(\mathcal{V}^{(i-1)})} w^\top x^* && \triangleright \text{optimality of } x^{(i)} \\
&\leq g(x^*) + \max_{w \in B(F)} w^\top x^* && \triangleright \mathcal{V}^{(i-1)} \subseteq B(F) \\
&= g(x^*) + f(x^*), && \triangleright \text{from (2.1)}
\end{aligned}$$

for any optimal solution x^* . The non-decreasing property of $\{d^{(i)}\}$ holds obviously when $i = 1$ since $d^{(0)} = -\infty$. When $i \geq 2$, using Lemma 4.1, from the inclusion relation $\mathcal{A}^{(i-1)} \subseteq \mathcal{V}^{(i-1)}$, we have

$$\begin{aligned}
(4.2) \quad d^{(i-1)} &= g(x^{(i-1)}) + \max_{w \in \text{conv}(\mathcal{A}^{(i-1)})} w^\top x^{(i-1)} \leq g(x^{(i)}) + \max_{w \in \text{conv}(\mathcal{A}^{(i-1)})} w^\top x^{(i)} \\
&\leq g(x^{(i)}) + \max_{w \in \text{conv}(\mathcal{V}^{(i-1)})} w^\top x^{(i)} = d^{(i)}.
\end{aligned}$$

Therefore, $d^{(i)}$'s are non-decreasing lower bounds of the optimal value. \square

COROLLARY 4.3. *If the subproblem in line 6 of Algorithm 4.1 is bounded below in the i th iteration, then the subproblems remain bounded below in the subsequent iterations.*

Note that since $B(F)$ has finitely many vertices, we can see OSM converges finitely: in the worst case, the method adds each vertex until the surrogate function $g + f_{(i)}$ minimized at the i th iteration matches the original objective function $g + f$ everywhere. (Recall OSM never deletes a vertex.) We next show that our proposed method L-KM does not stall at suboptimal iterates:

LEMMA 4.4. *For all $i \geq 0$, when $x^{(i)}$ is suboptimal, $x^{(i+1)} \neq x^{(i)}$.*

Proof. We first prove $\mathcal{V}^{(i-1)} \cap \partial f(x^{(i)}) = \emptyset$ when $x^{(i)}$ is suboptimal. If $x^{(i)}$ is suboptimal and there exists a $w^\natural \in \text{conv}(\mathcal{V}^{(i-1)}) \cap \partial f(x^{(i)})$, then

$$(4.3) \quad g(x^{(i)}) + f(x^{(i)}) \stackrel{(a)}{=} g(x^{(i)}) + w^\natural \top x^{(i)} \stackrel{(b)}{\leq} g(x^{(i)}) + f_{(i)}(x^{(i)}),$$

where (a) follows from $w^\natural \in \partial f(x^{(i)})$ and (b) follows from $w^\natural \in \mathcal{V}^{(i-1)}$. Thus $x^{(i)}$ is a minimizer of $g + f$ as $g(x) + f_{(i)}(x) \leq g(x) + f(x), \forall x \in \mathbb{R}^n$, and we have a contradiction.

Now suppose that $x^{(i)}$ is suboptimal and $x^{(i+1)} = x^{(i)}$. We have

$$(4.4) \quad d^{(i+1)} = g(x^{(i)}) + f_{(i+1)}(x^{(i)}) \stackrel{(a)}{\geq} g(x^{(i)}) + v^{(i) \top} x^{(i)} \stackrel{(b)}{=} g(x^{(i)}) + f(x^{(i)}) \geq p^*,$$

where (a) follows from $v^{(i)} \in \mathcal{V}^{(i)}$ (since $f(x) \geq v^\top x$ for all $v \in \mathcal{V}^{(i)}$) and (b) follows from $v^{(i)} \in \partial f(x^{(i)})$ respectively. This argument shows $x^{(i)}$ is optimal, and we have a contradiction. \square

We show next that the size of the subproblems stays bounded by $n + 1$. This follows from the affine independence of the sets $\mathcal{V}^{(i)}$, as we show in the following lemma:

LEMMA 4.5. *For all $i \geq 0$, vectors in $\mathcal{V}^{(i)}$ are affinely independent. Moreover, $|\mathcal{V}^{(i)}| \leq n + 1$.*

Proof. We prove this claim by induction. The claim is true for $i = 0$ by the definition of $\mathcal{V}^{(0)}$. Suppose that the claim is true for every $i < i_0$. When the duality gap $\Delta^{(i_0)} > \epsilon$, we have $v^{(i_0) \top} x^{(i_0)} = f(x^{(i_0)}) > f_{(i_0)}(x^{(i_0)})$. From $\mathcal{A}^{(i_0)} \subseteq \{w \in \mathbb{R}^n : w^\top x^{(i_0)} = f_{(i_0)}(x^{(i_0)})\}$, we have $v^{(i_0)} \notin \text{affine}(\mathcal{A}^{(i_0)})$. Hence $\mathcal{V}^{(i_0)} = \mathcal{A}^{(i_0)} \cup v^{(i_0)}$ is a set of affinely independent vectors.

Otherwise, when $\Delta^{(i_0)} \leq \epsilon$, the algorithm terminates in the i_0 th iteration.

Since any set of more than $n + 1$ vectors in \mathbb{R}^n must be affinely dependent, we have $|\mathcal{V}^{(i)}| \leq n + 1$. \square

Thus, we have shown that L-KM solves a series of limited memory convex subproblems whose number of linear constraints do not exceed $n + 1$. L-KM produces non-decreasing lower bounds and does not stall at suboptimal solutions. In Section 5, we will show a rate of convergence of the method.

4.3. Lovász Extension Over Absolute Values. A simple variant of L-KM can be used to minimize $g(x) + f(|x|)$: simply substitute $B(F)$, $f(x)$ and $\mathcal{V}(x^{(i)})$ with $|P|(F)$, $f(|x|)$ and $\bar{\mathcal{V}}(x^{(i)})$ respectively in Algorithm 4.1. All the results discussed earlier in this chapter hold for this variant as well. To initialize the algorithm, when g is bounded below, we can set $\mathcal{V}^{(0)} = \{v, -v\}$ for some vertex v of $|P|(F)$. This implies that $f_{(1)} \geq 0$, and $g(x) + f_{(1)}(x)$ is thus bounded below over \mathbb{R}^n .

5. DL-KM, Connections to FCFW and Convergence Analysis. In this section we present DL-KM, the dual of L-KM, as a special case of the Fully-Corrective Frank-Wolfe algorithm that uses limited memory. This will help us also provide a direct proof of finite convergence of L-KM. The results in this section will be applicable to general polytopes, not just the submodular base polytopes. We assume that g is closed, which implies $g^{**} = g$.

The dual of OSM can be seen as a first-order method which exploits the fact that it is easy to do linear optimization over the base polytope $B(F)$ [3]. We use this duality to re-write L-KM in the dual form and derive the dual method from scratch for L-KM in Appendix C. DL-KM is thus limited memory like L-KM. For ease of readability and to better demonstrate the connections between DL-KM and FCFW method over general polytopes, we re-write DL-KM for the minimization of the convex function $h(w) = g^*(-w)$ over the base polytope $B(F)$ (see Algorithm 5.1). For general polytopes P , $B(F)$ can be replaced verbatim with P in Algorithm 5.1.

Algorithm 5.1 DL-KM: Dual of Limited memory Kelley’s Method

- 1: **Input:** a convex function $h : \mathbb{R}^n \rightarrow \mathbb{R}$, polytope $B(F) \subseteq \mathbb{R}^n$, suboptimality parameter $\epsilon > 0$
 - 2: **Output:** an approximate solution $w^\sharp \in B(F)$ s.t. $h(w^\sharp) - \epsilon \leq \min_{w \in B(F)} h(w)$
 - 3: **Initialization:** set $\bar{\Delta}^{(0)} = \infty$, $i = 1$, and let $\mathcal{V}^{(0)} \subseteq \text{vert}(B(F))$ be a nonempty set of affinely independent vectors from $\text{vert}(B(F))$ such that $\inf_{x \in \text{conv}(\mathcal{V}^{(0)})} g(x)$ is finite
 - 4: **while** $\bar{\Delta}^{(i-1)} > \epsilon$ **do**
 - 5: let $w^{(i)} \in \arg \min_{w \in \text{conv}(\mathcal{V}^{(i-1)})} h(w)$
 - 6: compute $x^{(i)} \in -\partial h(w^{(i)})$
 - 7: compute $v^{(i)} \in \arg \max_{v \in \text{vert}(B(F))} v^\top x^{(i)}$
 - 8: let $\bar{\Delta}^{(i)} := (v^{(i)} - w^{(i)})^\top x^{(i)}$ ▷ update the optimality gap
 - 9: let $\mathcal{A}^{(i)} := \{v \in \mathcal{V}^{(i-1)} : v^\top x^{(i)} = w^{(i)\top} x^{(i)}\}$ ▷ set of tight vertices w.r.t sub-gradient $x^{(i)}$
 - 10: let $\mathcal{V}^{(i)} := \mathcal{A}^{(i)} \cup \{v^{(i)}\}$ ▷ update $\mathcal{V}^{(i)}$
 - 11: $i = i + 1$
 - 12: **return** $w^{(i-1)}$
-

In the i th iteration, DL-KM minimizes $h(\cdot)$ over $\text{conv}(\mathcal{V}^{(i-1)})$ which is a subset of the polytope $B(F)$ (or a general polytope P), to get an approximate solution $w^{(i)}$. When the optimality gap is not desirable, we generate a new vertex $v^{(i)}$ by maximizing the sub-gradient $x^{(i)}$ of $-h(w^{(i)})$, and define $\mathcal{V}^{(i)}$ by expanding $\text{conv}(\mathcal{A}^{(i)})$ (the tight set of vertices with respect to $x^{(i)}$) to include $v^{(i)}$. Since the current approximate solution $w^{(i)}$ is contained in the next feasible set $\text{conv}(\mathcal{V}^{(i)})$, the next subproblem is also feasible. Through duality, we have $\bar{d}^* = -p^*$ and $h(w^{(i)}) = -d^{(i)}$, i.e. the lower bounds in L-KM.

To analyze the rate of convergence for L-KM, we present the Fully-Corrective Frank-Wolfe (FCFW) Method with approximate correction [14] in Algorithm 5.2 and show that DL-KM is a limited memory special case of FCFW in Theorem 5.2.

Before we prove the main theorem in this section, we need the following lemma:

LEMMA 5.1. *Consider a representation of $w^{(i)}$ as a convex combination of vertices $v \in \mathcal{V}^{(i-1)}$,*

$$w^{(i)} = \sum_{v \in \mathcal{V}^{(i-1)}} \lambda_v^{(i)} v,$$

where $0 \leq \lambda_v^{(i)} \leq 1$ and $\sum_{v \in \mathcal{V}^{(i)}} \lambda_v^{(i)} = 1$. In this decomposition, only vertices $v \in \mathcal{A}^{(i)}$ can have positive weights $\lambda_v^{(i)}$. Other vertices $v \in \mathcal{V}^{(i-1)} \setminus \mathcal{A}^{(i)}$ have weights $\lambda_v^{(i)} = 0$.

Algorithm 5.2 FCFW: Fully-Corrective Frank-Wolfe Method with approximate correction [14]

- 1: **Input:** convex function $h : \mathbb{R}^n \rightarrow \mathbb{R}$, polytope $P \subseteq \mathbb{R}^n$, suboptimality parameter $\epsilon > 0$
 - 2: **Output:** an approximate solution $w^\sharp \in P$ s.t. $h(w^\sharp) - \epsilon \leq \min_{w \in P} h(w)$
 - 3: **Initialization:** let $w^{(0)} \in \text{vert}(P)$, $x^{(0)} \in -\partial h(w^{(0)})$, $v^{(0)} \in \arg \max_{v \in \text{vert}(P)} v^\top x^{(0)}$, $\mathcal{V}^{(0)} := \{w^{(0)}\}$, $v^{(0)}, \bar{\Delta}^{(0)} = (v^{(0)} - w^{(0)})^\top x^{(0)}$, $i = 1$
 - 4: **while** $\bar{\Delta}^{(i-1)} > \epsilon$ **do**
 - 5: compute $w^{(i)}, \mathcal{V}^{(i)}$ such that
 - (a). $w^{(i)} \in \text{conv}(\mathcal{V}^{(i)})$
 - (b). $h(w^{(i)}) \leq \min_{\lambda \in [0,1]} h(w^{(i-1)} + \lambda(v^{(i-1)} - w^{(i-1)}))$ \triangleright make some progress with FW step
 - (c). $\max_{v \in \mathcal{V}^{(i)}} (w^{(i)} - v)^\top x^{(i)} \leq \epsilon$ \triangleright make the away step gap small enough
 - 6: compute $v^{(i)} \in \arg \max_{v \in \text{vert}(P)} v^\top x^{(i)}$
 - 7: let $\bar{\Delta}^{(i)} := (v^{(i)} - w^{(i)})^\top x^{(i)}$ \triangleright update the optimality gap
 - 8: $i = i + 1$
 - 9: **return** $w^{(i-1)}$
-

Proof. Using the identity between the iterates $x^{(i)}$ in the primal and dual algorithms, detailed in (C.1), we have

$$\begin{aligned}
 (5.1) \quad \text{conv}(\mathcal{A}^{(i)}) &= \text{conv}(\{v \in \mathcal{V}^{(i-1)} : v^\top x^{(i)} = w^{(i)\top} x^{(i)}\}) \\
 &= \text{conv}(\{v \in \mathcal{V}^{(i-1)} : v^\top x^{(i)} = \max_{w \in \text{conv}(\mathcal{V}^{(i-1)})} w^\top x^{(i)}\}) \\
 &= \arg \max_{w \in \text{conv}(\mathcal{V}^{(i-1)})} w^\top x^{(i)}.
 \end{aligned}$$

Then for any representation $w^{(i)} = \sum_{v \in \mathcal{V}^{(i-1)}} \lambda_v^{(i)} v$ of $w^{(i)}$ as a convex combination of $v \in \mathcal{V}^{(i)}$, we have

$$\begin{aligned}
 (5.2) \quad 0 &= \left(w^{(i)} - \sum_{v \in \mathcal{V}^{(i-1)}} \lambda_v^{(i)} v \right)^\top x^{(i)} \\
 &= \sum_{v \in \mathcal{V}^{(i-1)}} \lambda_v^{(i)} [(w^{(i)})^\top x^{(i)} - v^\top x^{(i)}], \\
 &= \sum_{v \in \mathcal{V}^{(i-1)} \setminus \mathcal{A}^i} \lambda_v^{(i)} [(w^{(i)})^\top x^{(i)} - v^\top x^{(i)}],
 \end{aligned}$$

since $w^{(i)\top} x^{(i)} = v^\top x^{(i)}$ for every $v \in \mathcal{A}^{(i)}$. Using (5.1), we have $v^\top x^{(i)} < w^{(i)\top} x^{(i)}$ for any $v \in \mathcal{V}^{(i-1)} \setminus \mathcal{A}^{(i)}$. Thus $\lambda_v^{(i)} = 0$ for any $v \in \mathcal{V}^{(i-1)} \setminus \mathcal{A}^{(i)}$. \square

We are now ready to prove the main theorem:

THEOREM 5.2. *The dual method DL-KM is a special case of FCFW with approximate correction as the approximate solution $w^{(i)}$ and the subset of vertices $\mathcal{V}^{(i)}$ constructed in DL-KM satisfy the conditions in Line 5(a)-(c) of Algorithm 5.2 in each iteration $i \geq 1$.*

Proof. Since DL-KM computes $w^{(i)} \in \arg \min_{w \in \text{conv}(\mathcal{V}^{(i-1)})} h(w)$, by first-order optimality conditions we get $w^{(i)\top} x^{(i)} \geq v^\top x^{(i)}$ for all $w \in \text{conv}(\mathcal{V}^{(i-1)})$ (recall $x^{(i)} \in -\partial h(w^{(i)})$). Since $v \in \mathcal{A}^{(i)}$ for all $v \in \mathcal{V}^{(i-1)}$ such that $v^\top x^{(i)} = w^{(i)\top} x^{(i)}$, only vertices in $\mathcal{A}^{(i)}$ can have non-zero convex multipliers in the decomposition of $w^{(i)}$ using Lemma 5.1. Therefore, $w^{(i)} \in \text{conv}(\mathcal{A}^{(i)}) \subseteq \text{conv}(\mathcal{V}^{(i)})$ and (a) holds. Condition (b) holds because $\text{conv}(\{w^{(i-1)}, v^{(i-1)}\}) \subseteq \text{conv}(\mathcal{V}^{(i-1)})$, over which $w^{(i)}$ is optimal. Lastly (c) holds because by construction of $\mathcal{A}^{(i)}$ and $v^{(i)}$, $(w^{(i)} - v)^\top x^{(i)} = 0$ for $v \in \mathcal{A}^{(i)}$ and $(w^{(i)} - v)^\top x^{(i)} < 0$ when $v = v^{(i)}$. \square

Therefore, known convergence rates for FCFW [14] are applicable to DL-KM:

THEOREM 5.3. *Suppose g is closed, $\frac{1}{\mu}$ -strongly smooth and $\frac{1}{L}$ -strongly convex. Let M be the diameter*

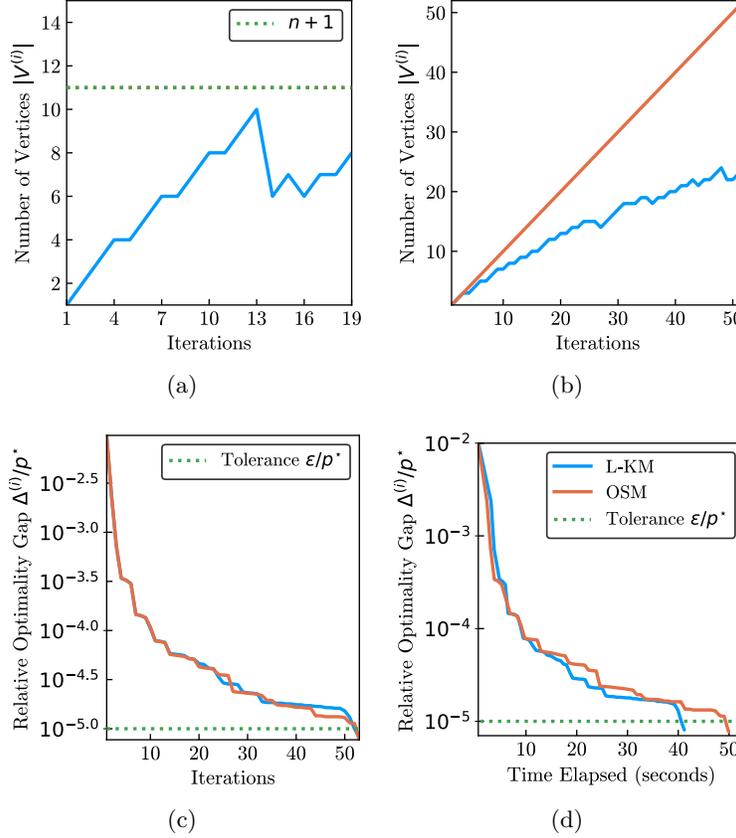


Fig. 2: Dimension $n = 10$ in (a), $n = 100$ in (b), (c) and (d). The methods converged in (a), (b), (c) and (d).

and δ be the pyramidal width⁵ of P , then the lower bounds $d^{(i)}$ in Algorithm 4.1 converges linearly at the rate of $1 - \rho$, i.e. $p^* - d^{(i+1)} \leq (1 - \rho)(p^* - d^{(i)})$, where $\rho \triangleq \frac{\mu}{4L} \left(\frac{\delta}{M}\right)^2$.

THEOREM 5.4. *Suppose g is closed, $\frac{1}{L}$ -strongly convex and let M be the diameter of P , the duality gap $\Delta^{(i)}$ in Algorithm 4.1 converges sub-linearly: $\Delta^{(i+1)} \leq (p^* - d^{(i)}) + LM^2/2$ when $(p^* - d^{(i)}) \geq LM^2/2$ and $\Delta^{(i)} \leq M\sqrt{2(p^* - d^{(i)})L}$ otherwise.*

Note that the same analysis would not apply to the dual of OSM since OSM does not support deletion of constraints. Therefore, it is not possible to bound the away step gap in Line 5(c).

6. Experiments and Conclusion. We present computational results in this section. The experiments were implemented in Julia, using Convex.jl and Submodular.jl. Convex solvers MOSEK and Gurobi were called to solve the subproblems.

6.1. Experiment 1. In the first experiment, we explore the behavior of L-KM on the problem of minimizing the non-separable composite function $g + f$, where $g(x) = x^\top(A + n\mathbf{I}_n)x + b^\top x$ for $x \in \mathbb{R}^n$ and f is the Lovász extension of the submodular function $F(A) = \frac{|A|(2n-|A|+1)}{2}$ for $A \subseteq [n]$. The entries of $A \in M_n$ and $b \in \mathbb{R}^n$ were randomly sampled from the uniform distributions on $[-1, 1]$, and $[0, n]$, respectively. In this experiment, the subproblems were solved to a relative tolerance of 10^{-7} using the interior point solver MOSEK.

Primal convergence: We first solve the problem with $n = 10$ to verify that the number of constraints

⁵See Appendix C for definitions of the diameter and pyramidal width.

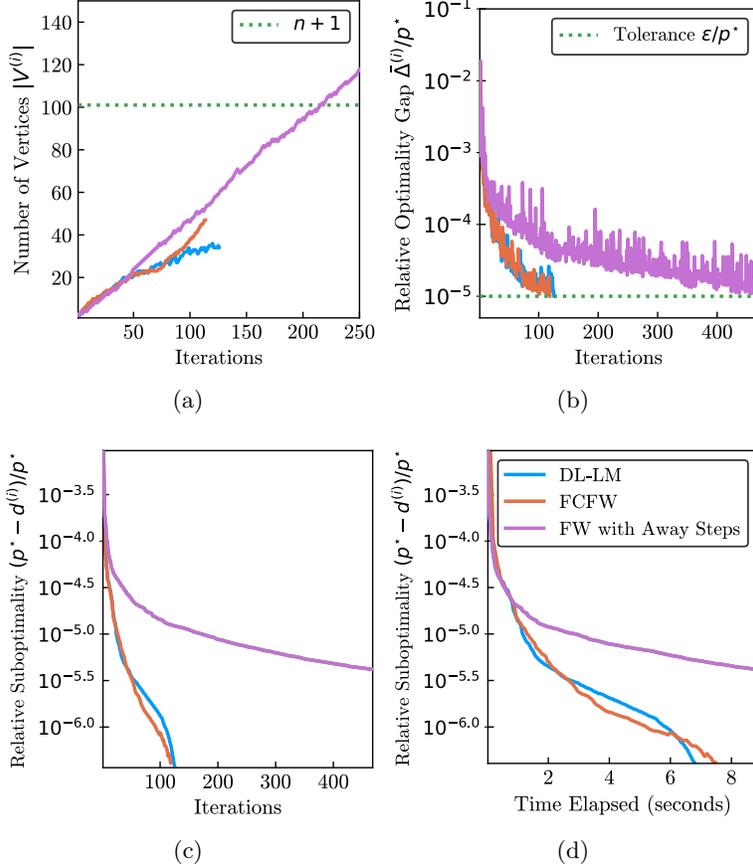


Fig. 3: DL-KM and FCFW converged in all plots, FW with away steps has converged in (b), (c) and (d).

does not exceed $n + 1$. Note that the number of constraints can sometimes decrease even before it reaches $n + 1$ (Figure 2(a)). We next compare the memory used in each iteration (Figure 2(b)), the optimality gap per iteration (Figure 2(c)), and the running time (Figure 2(d)) of L-KM and OSM by solving the problem for $n = 100$ up to a relative accuracy of 10^{-5} . Note that L-KM uses much less memory compared to OSM, converges in almost the same number of iterations, and completes later iterations much faster than OSM.

Dual convergence: We compare the convergence of DL-KM, FCFW and Frank-Wolfe with away steps for the dual problem $\max_{w \in B(F)} -(-w - b)^\top A^{-1}(-w - b)$ for $n = 100$ up to relative accuracy of 10^{-5} . DL-KM maintains smaller sized subproblems (Figure 3(a)), and it converges faster than FCFW as the number of iterations increases (Figure 3(d)). The duality gap in FCFW (provably) converges linearly. Moreover, as shown in Figures 3(b) and (c), DL-KM and FCFW converge much faster than Frank-Wolfe with away steps.

6.2. Experiment 2. In the second experiment, we select the dimension of the problem to be $n = 200$. We minimize the non-separable composite function $g(x) + f(|x|)$, where $g(x) = \|Ax - b\|_1$ for $x \in \mathbb{R}^n$ and f is the Lovász extension of the submodular function $F(A) = \frac{|A|(2n - |A| + 1)}{2}$ for $A \subseteq [n]$. We choose $A = P^\top \Lambda P \in S_n$, where Λ is a diagonal matrix half of whose eigenvalues are 25 and the other half are 10^{-1} , and P is a random orthonormal matrix. The entries of $b \in \mathbb{R}^n$ were chosen uniformly at random from $[0, n]$ as in the first experiment. In this experiment, the subproblems are linear, and are solved using the simplex solver Gurobi.

From Figure 4(b) and (c), we can see that although L-KM requires more iterations to converge, it takes less time in total to converge. As the number of iteration increases, L-KM takes less time to complete each

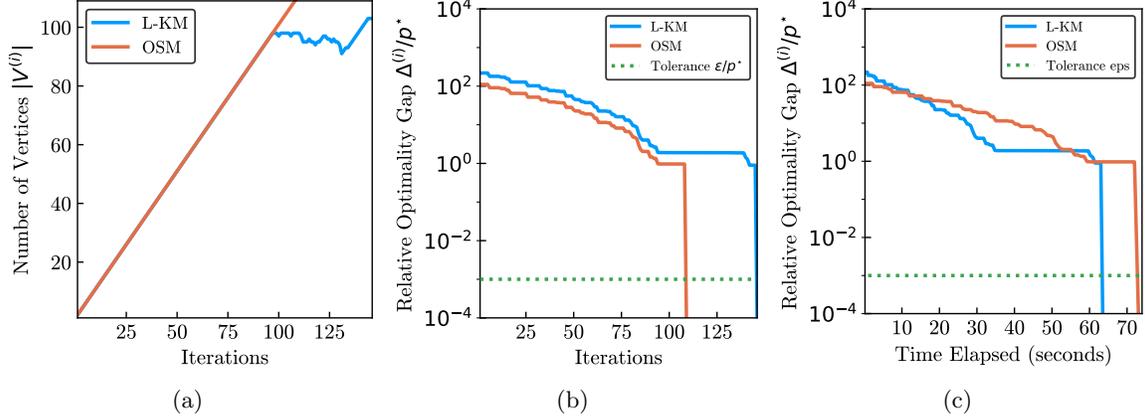


Fig. 4: Experimental results for an ill-conditioned problem of dimension 200. L-KM and OSM converged in all plots.

iteration compared to OSM.

6.3. Identifying Active Constraints. In our computational experiments, we use an equivalent definition for the active sets $\mathcal{A}^{(i)} := \{v \in \mathcal{V}^{(i-1)} : v^\top x^{(i)} = w^{(i)\top} x^{(i)}\}$ that is more stable numerically. We note that the active vertices correspond to active constraints, and use the dual values to decide which vertices are active. Concretely, we re-write the subproblem in Line 6 of Algorithm 4.1 as

$$(6.1) \quad \begin{aligned} \min_{x \in \mathbb{R}^n, t \in \mathbb{R}} \quad & g(x) + t \\ \text{s.t.} \quad & w^\top x \leq t, \forall w \in \mathcal{V}^{(i-1)}. \end{aligned}$$

Using the KKT conditions, we can see that the inequality constraints whose dual multipliers have positive values correspond to active vertices. Using primal-dual solvers for the subproblems in L-KM, we determine $\mathcal{A}^{(i)}$ by examining the dual multipliers of the inequality constraints in 6.1.

7. Conclusion. This paper defines a new limited memory version of Kelley’s method adapted to composite convex and submodular objectives, and establishes the first convergence rate for such a method, solving the open problem proposed in [2, 3]. We give guarantees on the memory requirements and convergence rate, and demonstrate compelling performance in practice.

Appendix A. Additional Background. We list some examples of popular submodular functions in Table 1.

Appendix B. The Original Simplicial Method.

In the i th iteration, OSM generates a lower linear approximation of f :

$$(B.1) \quad f_{(i)}(x) \triangleq \max_{w \in \text{conv}(\mathcal{V}^{(i-1)})} w^\top x,$$

where $\mathcal{V}^{(i-1)}$ is the set of all the sub-gradients of f (vertices of $B(F)$) collected in the previous iterations. A subproblem

$$(B.2) \quad \min_{x \in \mathbb{R}^n} g(x) + f_{(i)}(x)$$

is solved to get the solution $x^{(i)}$ and an updated lower bound $d^{(i)} \triangleq g(x^{(i)}) + f_{(i)}(x^{(i)})$ of the optimal value p^* . Afterwards, we calculate the optimality gap $\Delta^{(i)} \triangleq f(x^{(i)}) - f_{(i)}(x^{(i)})$. The method stops if the optimality

Problem	Submodular function, $S \subseteq E$ (unless specified)
n experts (simplex), $E = \{1, \dots, n\}$	$f(S) = 1$
k out of n experts (k-simplex), $E = \{1, \dots, n\}$	$f(S) = \min\{ S , k\}$
Permutations over $E = \{1, \dots, n\}$	$f(S) = \sum_{s=1}^{ S } (n+1-s)$
k-truncated permutations over $E = \{1, \dots, n\}$	$f(S) = (n-k) S $ for $ S \leq k$, $f(S) = k(n-k) + \sum_{s=k+1}^{ S } (n+1-s)$ if $ S \geq k$
Spanning trees on $G = (V, E)$	$f(S) = V(S) - \kappa(S)$, $\kappa(S)$ is the number of connected components of S
Matroids over ground set E : $M = (E, (\cdot), (\cdot)) \subseteq 2^E$	$f(S) = r_M(S)$, the rank function of the matroid
Coverage of T: given $T_1, \dots, T_n \subseteq T$	$f(S) = \bigcup_{i \in S} T_i $, $E = \{1, \dots, n\}$
Cut functions on a directed graph $D = (V, E)$, $c: E \rightarrow \mathbb{R}_+$	$f(S) = c(\delta^{out}(S))$, $S \subseteq V$
Flows into a sink vertex t , given a directed graph $D = (V, E)$ and costs $c: E \rightarrow \mathbb{R}_+$	$f(S) = \max$ flow from $S \subseteq V \setminus \{t\}$ into t
Maximal elements in E , $h: E \rightarrow \mathbb{R}$	$f(S) = \max_{e \in S} h(e)$, $f(\emptyset) = \min_{e \in E} h(e)$
Entropy H of random variables X_1, \dots, X_n	$f(S) = H(\bigcup_{i \in S} X_i)$, $E = \{1, \dots, n\}$

Table 1: Problems and the submodular functions (on ground set of elements E) that give rise to them.

Algorithm B.1 The Original Simplicial Method

- 1: **Input:** a convex function $g: \mathbb{R}^n \rightarrow \mathbb{R}$ and a submodular function $F: 2^n \rightarrow \mathbb{R}$, suboptimality tolerance $\epsilon > 0$
 - 2: **Output:** an approximate solution $x^\#$ s.t. $g(x^\#) + f(x^\#) - \epsilon \leq \min_{x \in \mathbb{R}^n} g(x) + f(x)$
 - 3: **Initialization:** let $x^{(0)} \in \mathbb{R}^n$ that has mutually distinct elements, $p^{(0)} := g(x^{(0)}) + f(x^{(0)})$, $d^{(0)} := -\infty$, $\Delta^{(0)} := p^{(0)} - d^{(0)}$, $\mathcal{V}^{(0)} := \mathcal{V}(x^{(0)})$, $i = 1$
 - ▷ $x^{(0)}$ could be set at other values
 - 4: **while** $\Delta^{(i-1)} > \epsilon$ **do**
 - 5: let $f_{(i)}(x) := \max_{w \in \text{conv}(\mathcal{V}^{(i-1)})} w^\top x$ ▷ update the approximation of f
 - 6: let $x^{(i)} \in \arg \min_{x \in \mathbb{R}^n} g(x) + f_{(i)}(x)$ ▷ solve the i th subproblem
 - 7: let $p^{(i)} := g(x^{(i)}) + f(x^{(i)})$ ▷ update the upper bound
 - 8: let $d^{(i)} := g(x^{(i)}) + f_{(i)}(x^{(i)})$ ▷ update the lower bound
 - 9: let $\Delta^{(i)} := p^{(i)} - d^{(i)}$ ▷ update the optimality gap
 - 10: compute $v^{(i)} \in \mathcal{V}(x^{(i)})$
 - 11: let $\mathcal{V}^{(i)} := \mathcal{V}^{(i-1)} \cup \{v^{(i)}\}$ ▷ update the set of sub-gradients of f
 - 12: $i = i + 1$
 - 13: **return** $x^{(i-1)}$
-

gap is small enough, otherwise, we update the set of sub-gradients of f by computing $\mathcal{V}^{(i)} \triangleq \mathcal{V}^{(i-1)} \cup \{v^{(i)}\}$ where $v^{(i)} \in \mathcal{V}(x^{(i)})$, and go on to the $(i+1)$ th iteration.

From the inclusion relation $\text{conv}(\mathcal{V}^{(i)}) \subseteq \text{conv}(\mathcal{V}^{(i+1)}) \subseteq B(F)$, we have $f_{(i)}(x) \leq f_{(i+1)}(x) \leq f(x)$, $\forall x \in \mathbb{R}^n$, $i \in \mathbb{N}$. Thus $f_{(i+1)}$ is a better approximation of f than $f_{(i)}$, and $d^{(i)}$'s are non-decreasing lower bounds of p^* .

Similar to Lemma 4.1, OSM does not stall at suboptimal iterations, thus $\mathcal{V}^{(i)}$ strictly expands in each iteration. Since $B(F)$ has finite number of vertices, OSM terminates within finite steps.

Appendix C. The Dual of the L-KM.

C.1. The Duality Between L-KM and DL-KM. To establish the duality between Algorithm 4.1 and 5.1, we prove the equivalence of $x^{(i)}$, $\mathcal{A}^{(i)}$ and $\mathcal{V}^{(i)}$ in Algorithm 4.1 and 5.1, which imply the equivalence of $v^{(i)}$:

LEMMA C.1. *When g is closed, the vector $x^{(i)}$ defined in line 6 of Algorithm 5.1 is also an optimal solution to the i th subproblem of Algorithm 4.1.*

Proof. We first prove

$$(C.1) \quad w^{(i)\top} x^{(i)} = f_{(i)}(x^{(i)}) = \max_{w \in \text{conv}(\mathcal{V}^{(i-1)})} w^\top x^{(i)}.$$

If there exists a $w^\natural \in \text{conv}(\mathcal{V}^{(i-1)})$ such that $(w^\natural - w^{(i)})^\top x^{(i)} < 0$, $w^\natural - w^{(i)}$ is thus a feasible descent direction in $\text{conv}(\mathcal{V}^{(i-1)})$ since the line segment between w^\natural and $w^{(i)}$ lies in $\text{conv}(\mathcal{V}^{(i-1)})$. This contradicts the optimality of $w^{(i)}$ in $\text{conv}(\mathcal{V}^{(i-1)})$.

Since g is closed and $x^{(i)} \in -\partial h(w^{(i)})$, we have

$$(C.2) \quad g(x^{(i)}) + g^*(-w^{(i)}) = -w^{(i)\top} x^{(i)} = -f_{(i)}(x^{(i)}),$$

and $x^{(i)}$ is an optimal solution to the i th subproblem of Algorithm 4.1. \square

LEMMA C.2. *When g is closed, the definition of $\mathcal{A}^{(i)}$ and $\mathcal{V}^{(i)}$ in Algorithm 4.1 and 5.1 are equivalent.*

Proof. We prove this by induction. The claim is obviously true for $i = 0$ since $\mathcal{V}(0)$ is also the set of the extreme points of $B(F)$. Suppose that the claim is true for $i < i_0$. Lemma C.1 has established the equivalence of $x^{(i_0)}$ in the two algorithms. So from (C.1),

$$(C.3) \quad \{v \in \mathcal{V}^{(i-1)} : v^\top x^{(i)} = w^{(i)\top} x^{(i)}\} = \{v \in \mathcal{V}^{(i-1)} : v^\top x^{(i)} = f_{(i)}(x^{(i)})\},$$

and the $\mathcal{V}^{(i)}$ defined in the two algorithms are equivalent. \square

C.2. Definition of Diameter and Pyramid Width. The diameter of a set $\mathcal{P} \subseteq \mathbb{R}^n$ is defined as

$$(C.4) \quad \text{Diam}(\mathcal{P}) \triangleq \max_{v, w \in \mathcal{P}} \|v - w\|_2.$$

Given a direction $x \in \mathbb{R}^n$, the directional width of a set $\mathcal{P} \subseteq \mathbb{R}^n$ with respect to x is defined as

$$(C.5) \quad \text{dirW}(\mathcal{P}, x) \triangleq \max_{v, w \in \mathcal{P}} (v - w)^\top \frac{x}{\|x\|_2}.$$

Pyramid directional width and pyramid width are defined by Lacoste-Julien and Jaggi in [14] for a finite sets of vectors $\mathcal{V} \subseteq \mathbb{R}^n$. Similarly, the pyramidal width of a polytope $\mathcal{P} = \text{conv}(V)$ can be defined as:

Pyramid Directional Width. Let $\mathcal{V} \subseteq \mathbb{R}^n$ be a finite set of vectors in \mathbb{R}^n . The pyramid directional width of \mathcal{V} with respect to a direction x and a base point $w \in \text{conv}(\mathcal{V})$ is defined as

$$(C.6) \quad \text{PdirW}(\mathcal{V}, x, w) \triangleq \min_{A \in \mathcal{A}(w)} \text{dirW}(A \cup \{v(\mathcal{V}, x)\}, x),$$

where $\mathcal{A}(w) \triangleq \{A \subseteq \mathcal{V} : \text{the convex multipliers are non-zero for all } v \in A \text{ in the decomposition of } w\}$ and $v(\mathcal{V}, x)$ is a vector in $\arg \max_{v \in \mathcal{V}} v^\top x$. The pyramid directional width got its name because the set $A \cup \{v(\mathcal{V}, x)\}$ has the shape of a pyramid with A being the base and $v(\mathcal{V}, x)$ being the summit.

Pyramid Width. The pyramid width of \mathcal{P} is defined as

$$(C.7) \quad \text{PWidth}(\mathcal{P}) \triangleq \min_{\mathcal{K} \in \text{face}(\mathcal{P})} \min_{x \in \text{cone}(\mathcal{K} - w) \setminus \{0\}, w \in \mathcal{K}} \text{PdirW}(\mathcal{K} \cap \text{vert}(\mathcal{P}), x, w),$$

where $\text{face}(\mathcal{P})$ stands for the faces of \mathcal{P} and $\text{cone}(\mathcal{K} - w)$ is equivalent to the set of vectors pointing inwards \mathcal{K} .

Acknowledgments. Parts of this work were completed while the first author was at the Operations Research Center at the Department of Mathematical Science at Tsinghua University, and while the second author was a research fellow at the Simons Institute, UC Berkeley. The third author was supported in part by DARPA Award FA8750-17-2-0101.

REFERENCES

- [1] J. AUDIBERT, S. BUBECK, AND G. LUGOSI, *Regret in online combinatorial optimization*, Mathematics of Operations Research, 39 (2013), pp. 31–45.
- [2] F. BACH, *Duality between subgradient and conditional gradient methods*, SIAM Journal on Optimization, 25 (2015), pp. 115–129.
- [3] F. BACH ET AL., *Learning with submodular functions: A convex optimization perspective*, Foundations and Trends® in Machine Learning, 6 (2013), pp. 145–373.
- [4] F. R. BACH, *Structured sparsity-inducing norms through submodular functions*, in Advances in Neural Information Processing Systems, 2010, pp. 118–126.
- [5] E. W. CHENEY AND A. A. GOLDSTEIN, *Newton’s method for convex programming and Tchebycheff approximation*, Numerische Mathematik, 1 (1959), pp. 253–268.
- [6] J. DJOLONGA AND A. KRAUSE, *From MAP to marginals: Variational inference in bayesian submodular models*, in Advances in Neural Information Processing Systems, 2014, pp. 244–252.
- [7] Y. DRORI AND M. TEBoulLE, *An optimal variant of Kelleys cutting-plane method*, Mathematical Programming, 160 (2016), pp. 321–351.
- [8] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Research Logistics Quarterly, 3 (1956), pp. 95–110.
- [9] D. HEARN, S. LAWPHONGPANICH, AND J. VENTURA, *Restricted simplicial decomposition: Computation and extensions*, Mathematical Programming Study, 31 (1987), pp. 99–118.
- [10] J. E. KELLEY, JR, *The cutting-plane method for solving convex programs*, Journal of the Society for Industrial and Applied Mathematics, 8 (1960), pp. 703–712.
- [11] K. C. KIWIEL, *Proximal level bundle methods for convex nondifferentiable optimization, saddle-point problems and variational inequalities*, Mathematical Programming, 69 (1995), pp. 89–109.
- [12] W. M. KOOLEN, M. K. WARMUTH, AND J. KIVINEN, *Hedging structured concepts*, COLT, (2010).
- [13] W. KRICHENE, S. KRICHENE, AND A. BAYEN, *Convergence of mirror descent dynamics in the routing game*, in European Control Conference (ECC), IEEE, 2015, pp. 569–574.
- [14] S. LACOSTE-JULIEN AND M. JAGGI, *On the global linear convergence of Frank-Wolfe optimization variants*, in Advances in Neural Information Processing Systems, 2015, pp. 496–504.
- [15] C. LEMARÉCHAL, A. NEMIROVSKII, AND Y. NESTEROV, *New variants of bundle methods*, Mathematical programming, 69 (1995), pp. 111–147.
- [16] L. LOVÁSZ, *Submodular functions and convexity*, Mathematical Programming: The State of the Art, (1983).
- [17] M. MÄKELÄ, *Survey of bundle methods for nonsmooth optimization*, Optimization methods and software, 17 (2002), pp. 1–29.
- [18] K. NAGANO, Y. KAWAHARA, AND K. AIHARA, *Size-constrained submodular minimization through minimum norm base*, in Proceedings of the 28th International Conference on Machine Learning (ICML), 2011, pp. 977–984.
- [19] A. S. NEMIROVSKI AND D. B. YUDIN, *Problem complexity and method efficiency in optimization*, Wiley-Interscience, New York, (1983).
- [20] T. ROTHVOSS, *Some 0/1 polytopes need exponential size extended formulations*, Mathematical Programming, 142 (2013), pp. 255–268.
- [21] B. VON HOHENBALKEN, *Simplicial decomposition in nonlinear programming algorithms*, Mathematical Programming, 13 (1977), pp. 49–68.
- [22] M. K. WARMUTH AND D. KUZMIN, *Randomized PCA algorithms with regret bounds that are logarithmic in the dimension*, in Advances in Neural Information Processing Systems, 2006, pp. 1481–1488.
- [23] S. YASUTAKE, K. HATANO, S. KIJIMA, E. TAKIMOTO, AND M. TAKEDA, *Online linear optimization over permutations*, in Algorithms and Computation, Springer, 2011, pp. 534–543.