Università degli Studi di Padova

Corso di Dottorato in Matematica

Department of Mathematics "Tullio Levi-Civita"

Curriculum: computational mathematics

First and zeroth order optimization methods for data science

Candidate
Damiano Zeffiro

Supervisor
Prof. Francesco Rinaldi

PhD Coordinator
Prof. Giovanni Colombo

ciclo XXXV, 2019-2022

Università degli Studi di Padova, Department of Mathematics "Tullio Levi-Civita".

Acknowledgments

I thank my advisor for his constant help and guidance, prof. Immanuel Bomze for many interesting discussions, and prof. Luis Nunes Vicente for his ideas and his warm hospitality at the University of Lehigh. I also want to thank Dr. Vyacheslav Kungurtsev, prof. Giampaolo Liuzzi and prof. Stefano Lucidi for our collaborations. Last but not least, I want to thank my parents for their love and support.

Contents

Chapter 1

Introduction

Recent data science applications using large datasets often need scalable optimization methods with low per iteration cost and low memory requirements. This has lead to a renewed interest in gradient descent methods, and on tailored variants for problems where gradient descent is unpractical due, e.g., to non smoothness or stochasticity of the optimization objective. Applications include deep neural network training, adversarial attacks in machine learning, sparse signal recovery, cluster detection in networks, etc.

In this thesis, we focus on the theoretical analysis of some of these variants, as well as in the formulation and numerical testing of new variants with better complexity guarantees than existing ones under suitable conditions. The problems we consider have a continuous but sometimes constrained and not necessarily differentiable objective.

All the methods we are concerned with are characterized by the following iterative scheme: at every iteration, a black box oracle is evaluated in the current point to obtain certain local information about the optimization objective. Based on this information, possibly combined with that obtained in previous iterations, the next iterate is chosen. We remark that this is a classic scheme for nonlinear optimization algorithms, used by many previous authors (see, e.g., [191] and references therein). Another feature of the methods we are interested in is that they are all either first or zeroth order methods. The distinction between these two classes is based on the information that can be obtained with the black box oracle. In first order methods, the information consists of the gradient and the value of the objective for the current iterate; in zeroth order methods instead the only information available is the value of the objective for the current iterate.

While each chapter of the thesis can be read independently with some minor overlap in the definitions, broadly speaking our work deals with two specific classes of methods:

- First order projection free methods for the optimization of a smooth objective constrained to a convex set. These are variants of the projected gradient descent method, and are designed to avoid expensive projections on certain classes of problems. The main application of our theoretical analysis is the study of variants of the classic Frank Wolfe (FW) method, characterized by its use of linear minimizations instead of projections and its sparse approximation properties. For these methods, the original contributions of this thesis include proving new support identification properties for a FW variant with quantitative bounds, proposing a technique to provably speed up the convergence of several FW variants for non convex objectives, and an application to a cluster detection problem in networks.

- Direct search methods. These are zeroth order (often also referred to as derivative free) methods that, mimicking the basic idea behind the gradient descent method, try to improve the objective by generating a new iterate moving from the current one along a tentative descent direction with a suitable stepsize. The resulting point is then accepted if some sufficient decrease condition is satisfied. In this thesis, we extend the analysis of some direct search methods to optimization problems with non smooth and stochastic objectives, as well as to optimization problems defined on Riemannian manifolds.

1.1 Outline and main results

We now present an outline of the thesis and give pointers to the main results. Chapter 2 is a survey about the Frank Wolfe method and some of its variants, focusing on applications and recent developments in the theoretical analysis. The method (Algorithm 2) is presented as an instance of a general scheme for first order optimization methods (Algorithm 1), which includes also its main variants, introduced in Section 2.6.3. Some fundamental convergence results are summarized in Table 2.2. In Chapter 3, a unifying framework for the study of projection free methods is described, with a technique to recycle gradient related information in consecutive iterations (Algorithm 3), and linear convergence rate guarantees (Theorem 3.4.13). The main assumptions in this chapter are an angle condition for the descent directions selected by the method, given in Section 3.3 and with examples in

Sections 3.3.1, 3.5.2, and a Kurdyka-Lojasiewicz property for the objective (Assumption 3.1). In Chapter 4, some results are proven about the active set identification property of the away step Frank-Wolfe method (Algorithm 6). In particular, a local identification result for non convex objectives (Theorem 4.3.3) is used to prove qualitative (Theorems 4.4.3 and 4.5.5) and quantitative (Corollary 4.4.5, Theorems 4.5.9 and 4.5.6) active set identification results.

In Chapter 5, a continuous cubic formulation of a cluster detection problem in networks is proposed (problem (P)), together with a Frank-Wolfe variant (Algorithm 8) that provably identifies a local solution of the formulation in finite time (Theorem 5.4.2). Numerical results in Section 5.5 show that this approach is competitive with a state of the art local solver. Chapter 6 consists of a brief survey of direct search methods, focusing on directional direct search approaches. Some popular methods of this kind are described in Section 6.3 as instances of the general scheme 10. In Chapter 7, direct search schemes for smooth (Algorithms 13 and 14) and non smooth (Algorithms 16 and 17) optimization over Riemannian manifolds extending some of the methods discussed in Chapter 6 are presented. Convergence results are given in Theorems 7.3.4, 7.3.6, 7.4.5 and 7.4.6. In Chapter 8, a direct search method for stochastic unconstrained non smooth optimization is analyzed (Algorithm 18), under power law tail bounds on the objective evaluation noise (Assumptions 8.1 and 8.2). Convergence of the method is proved (Theorem 8.3.3) with a number of samples per iteration lower than the one used in other state of the art derivative free methods (see Theorem 8.2.9 and Remark 8.2.10). Chapter 9, some conclusions and potential future developments are discussed.

A detailed introduction can be found at the beginning of each chapter.

1.2 Notation

We denote as $\mathbb{N}_0$ the set of nonnegative integers, and for $a, b \in \mathbb{Z}$ as $[a : b]$ the set of integers between a and b, extremes included. For a set S we denote as $|S|$ and 2^S the cardinality and the set of subsets of S respectively. For a sequence $\{x_k\}_{k \in I}$ we often omit the index set I when it is clear from the context. We denote with e the vector with components all equal to 1, and with e_i the $i-th$ column of the identity matrix, with dimensions depending from the context.

For $p \geq 0$ we denote with $\| \cdot \|_p$ the $p-$th norm: for $x \in \mathbb{R}^n$,

$$\|x\|_p = \sqrt[p]{\sum_{i=0}^{n} |x_i|^p} \,. \tag{1.2.1}$$

For $c \in \mathbb{R}^n$ we denote with $\hat{c}$ the normalized vector $c/\|c\|$ if $c \neq 0$, and 0 otherwise. We define then $\mathrm{supp}(x)$ as the support of x:

$$\mathrm{supp}(x) = \{i \in [1\!:\!n] : x_i \neq 0\}\,, \tag{1.2.2}$$

and use $\|\cdot\|_0$ for the cardinality of the support

$$\|x\|_0 := |\mathrm{supp}(x)|\,. \tag{1.2.3}$$

We say that a function f differentiable in $\Omega \subset \mathbb{R}^n$ has Lipschitz continuous gradient with constant L if for every x, y in Ω

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|\,. \tag{1.2.4}$$

The function f is instead said to be $\mu-$strongly convex in Ω if for every x, y in Ω

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|x - y\|^2\,, \tag{1.2.5}$$

and Ω itself is said to be $\alpha-$strongly convex if, for any $x, y \in \Omega$, $\gamma \in [0, 1]$ and z such that $\|z\| = 1$, it holds that

$$\gamma x + (1 - \gamma)y + \gamma(1 - \gamma)\frac{\alpha}{2}\|x - y\|^2 z \in \Omega\,. \tag{1.2.6}$$

For a compact set $\Omega \subset \mathbb{R}^n$ the linear minimization oracle is defined as the black box oracle $\mathrm{LMO}_\Omega(\cdot)$ that given as input $r \in \mathbb{R}^n$ produces as output a minimizer in Ω of the scalar product with r:

$$\mathrm{LMO}_\Omega(r) \in \operatorname*{arg\,min}_{y \in \Omega} r^\top y\,. \tag{1.2.7}$$

For a bounded polytope $P \subset \mathbb{R}^n$ and $r \in \mathbb{R}^n$, we define as $\mathcal{F}_e(r)$ the face of P exposed by r:

$$\mathcal{F}_e(r) = \arg\max\{r^\top y \mid y \in P\}\,, \tag{1.2.8}$$

where the polytope P will always be clear from the context. Since P is a bounded polytope, hence in particular compact, the function $y \to r^\top y$ constrained to P has always a (finite) maximum, so that $\mathcal{F}_e(r)$ is non empty, and uniquely defined as a subset of P.

We denote with Δ_{n-1} the standard simplex:

$$\Delta_{n-1} := \{x \in \mathbb{R}^n_+ : e^\top x = 1\} = \mathrm{conv}\{e_i : i \in [1\!:\!n]\}\,.$$

Finally, we use $\operatorname{dist}(\cdot,\cdot)$ for the standard Euclidean distance in $\mathbb{R}^n$ either between points, or between a point and a subset, or between subsets: that is, if A, X are subsets of $\mathbb{R}^n$, then

$$\operatorname{dist}(A, X) = \inf\{\|x - a\| \mid x \in X,\ a \in A\}\,. \tag{1.2.9}$$

An important exception to this notation is made in Chapter 7, where we use dist to denote a Riemannian distance.

Chapter 2

Projection-free optimization methods

Invented some 65 years ago in a seminal paper by Marguerite Straus-Frank and Philip Wolfe, the Frank-Wolfe method recently enjoys a remarkable revival, fuelled by the need of fast and reliable first-order optimization methods in Data Science and other relevant application areas. In this chapter, we explain the success of this approach by illustrating versatility and applicability in a wide range of contexts, combined with an account on recent progress in variants, both improving on the speed and efficiency of this surprisingly simple principle of first-order optimization. We will focus on variants and convergence results most relevant to the contributions in Chapters 3-5. [1]

2.1 A short history

In their seminal work [101], Marguerite Straus-Frank and Philip Wolfe introduced a first-order algorithm for the minimization of convex quadratic objectives over polytopes, now known as Frank-Wolfe method. The main idea of the method is simple: to generate a sequence of feasible iterates by moving at every step towards a minimizer of a linearized objective, the so-called FW vertex. Subsequent works, partly motivated by applications in optimal control theory (see [94] for references), generalized the method to smooth (possibly non-convex) optimization over closed

[1] This chapter is based on the article "Frank-Wolfe and friends: a journey into projection-free first-order optimization methods" in *4OR, vol. 19, iss. 3, pp. 313-345, 2021* [48].

subsets of Banach spaces admitting a linear minimization oracle (see [89, 95]).

Furthermore, while the $O(1/k)$ rate in the original article was proved to be optimal when the solution lies on the boundary of the feasible set [65], improved rates were given in a variety of different settings. In [166] and [89], a linear convergence rate was proved over strongly convex domains assuming a lower bound on the gradient norm, a result then extended in [94] under more general gradient inequalities. In [116], linear convergence of the method was proved for strongly convex objectives with the minimum obtained in the relative interior of the feasible set.

The slow convergence behaviour for objectives with solution on the boundary motivated the introduction of several variants, the most popular being Wolfe's away step [237]. Wolfe's idea was to move away from bad vertices, in case a step of the FW method moving towards good vertices did not lead to sufficient improvement on the objective. This idea was successfully applied in several network equilibrium problems, where linear minimization can be achieved by solving a min-cost flow problem (see [105] and references therein). In [116], some ideas already sketched by Wolfe were formalized to prove linear convergence of the Wolfe's away step method and identification of the face containing the solution in finite time, under some suitable strict complementarity assumptions.

In recent years, the FW method has regained popularity thanks to its ability to handle the structured constraints appearing in machine learning and data science applications efficiently. Examples include LASSO, SVM training, matrix completion, minimum enclosing ball, density mixture estimation, cluster detection, to name just a few (see Section 2.4 for further details).

2.2 Main features of the Frank-Wolfe method

One of the main features of the FW algorithm is its ability to naturally identify sparse and structured (approximate) solutions. For instance, if the optimization domain is the simplex, then after k steps the cardinality of the support of the last iterate generated by the method is at most $k + 1$. Most importantly, in this setting every vertex added to the support at every iteration must be the best possible in some sense, a property that connects the method with many greedy optimization schemes [78]. This makes the FW method pretty efficient on the abovementioned problem class. Indeed, the combination of structured solutions with often noisy data makes the sparse approximations found by the method possibly more desirable than high precision solutions generated by a faster converging approach. In some cases, like in cluster detection (see, e.g., [40]), finding the support of the solution is actually enough to solve the problem independently from the precision achieved.

Another important feature is that the linear minimization used in the method is often cheaper than the projections required by projected-gradient methods. It is important to notice that, even when these two operations have the same complexity, constants defining the related bounds can differ significantly (see [80] for some examples and tests). When dealing with large scale problems, the FW method hence has a much smaller per-iteration cost with respect to projected-gradient methods. For this reason, FW methods fall into the category of *projection-free methods* [160]. Furthermore, the method can be used to approximately solve quadratic subproblems in accelerated schemes, an approach usually referred to as conditional gradient sliding (see, e.g., [66, 161]).

Finally, recent numerical results suggest that in some sparse optimization problems Frank Wolfe variants might be competitive with projected gradient methods even in iteration complexity [32], and thus without taking into account the advantage given by the faster linear minimization oracle.

2.3 Problem and general scheme

We consider the following problem:

$$\min_{x \in \Omega} f(x) \tag{2.3.1}$$

where, unless specified otherwise, Ω is a convex and compact (i.e. bounded and closed) subset of $\mathbb{R}^n$ and f is a differentiable function having Lipschitz continuous gradient with constant $L > 0$. This is a central property required in the analysis of first-order methods. Such a property indeed implies (and for a convex function is equivalent to) the so-called Descent Lemma (see, e.g., [31, Proposition 6.1.2]), which provides a quadratic upper approximation to the function f. Throughout this chapter, we denote by x^* a (global) solution to (2.3.1) and use the symbol $f^* := f(x^*)$ as a shorthand for the corresponding optimal value.

The general scheme of the first-order methods we consider for problem (2.3.1), reported in Algorithm 1, is based upon a set $\mathcal{A}(x, g)$ of directions feasible at x using first-order local information on f around x, in the smooth case $g = -\nabla f(x)$. From this set, a particular $d \in \mathcal{A}(x, g)$ is selected, with the maximal stepsize $\alpha^{\max}$ possibly dependent from auxiliary information available to the method (at iteration k, we thus write $\alpha_k^{\max}$), and not always equal to the maximal feasible stepsize.

Algorithm 1 `First-order method`

```
1: Choose a point $x_0 \in \Omega$
2: for $k = 0, \dots$ do
3:     if $x_k$ satisfies some specific condition then
4:         STOP
5:     end if
6:     Choose $d_k \in \mathcal{A}(x_k, -\nabla f(x_k))$
7:     Set $x_{k+1} = x_k + \alpha_k d_k$, with $\alpha_k \in (0, \alpha_k^{\max}]$ a suitably chosen stepsize
8: end for
```

2.3.1 The classical Frank-Wolfe method

The classical FW method for minimization of a smooth objective f generates a sequence of feasible points $\{x_k\}$ following the scheme of Algorithm 2. At the iteration k it moves toward a vertex i.e., an extreme point, of the feasible set minimizing the scalar product with the current gradient $\nabla f(x_k)$. It therefore makes use of a LMO for the feasible set Ω, defining the descent direction as

$$d_k = d_k^{FW} := s_k - x_k, \quad s_k \in \mathrm{LMO}_{\Omega}(\nabla f(x_k))\,. \tag{2.3.2}$$

In particular, the update at step 6 can be written as

$$x_{k+1} = x_k + \alpha_k (s_k - x_k) = \alpha_k s_k + (1 - \alpha_k) x_k \tag{2.3.3}$$

Since $\alpha_k \in [0, 1]$, by induction x_{k+1} can be written as a convex combination of elements in the set $S_{k+1} := \{x_0\} \cup \{s_i\}_{0 \leq i \leq k}$. When $C = \mathrm{conv}(A)$ for a set A of points with some common property, usually called "elementary atoms", if $x_0 \in A$ then x_k can be written as a convex combination of $k + 1$ elements in A. Note that due to Caratheodory's theorem, we can even limit the number of occurring atoms to $\min\{k, n\} + 1$. In the rest of the paper the primal gap at iteration k is defined as $h_k = f(x_k) - f^*$.

Algorithm 2 `Frank-Wolfe method`

```
Choose a point $x_0 \in \Omega$
for $k = 0, \dots$ do
    if $x_k$ satisfies some specific condition then
        STOP
    end if
    Compute $s_k \in \mathrm{LMO}_\Omega(\nabla f(x_k))$
    $d_k^{FW} = s_k - x_k$
    Set $x_{k+1} = x_k + \alpha_k d_k^{FW}$, with $\alpha_k \in (0, 1]$ a suitably chosen stepsize
end for
```

2.4 Examples

FW methods and variants are a natural choice for constrained optimization on convex sets admitting a linear minimization oracle significantly faster than computing a projection. We present here in particular the traffic assignment problem, submodular optimization, LASSO problem, matrix completion, adversarial attacks, minimum enclosing ball, SVM training, maximal clique search in graphs, sparse optimization.

2.4.1 Traffic assignment

Finding a traffic pattern satisfying the equilibrium conditions in a transportation network is a classic problem in optimization that dates back to Wardrop's paper [235]. Let $\mathcal{G}$ be a network with set of nodes $[1:n]$. Let $\{D(i,j)\}_{i\neq j}$ be demand coefficients, modeling the amount of goods with destination j and origin i. For any i, j with $i \neq j$ let furthermore $f_{ij} : \mathbb{R} \to \mathbb{R}$ be the non-linear (convex) cost functions, and x_{ij}^s be the flow on link (i, j) with destination s. The traffic assignment problem can be modeled as the following non-linear *multicommodity network* problem [105]:

$$\min \left\{ \sum_{i,j} f_{ij} \left(\sum_s x_{ij}^s \right) : \sum_i x_{\ell i}^s - \sum_j x_{j\ell}^s = D(\ell, s)\,, \text{ all } \ell \neq s,\ x_{ij}^s \geq 0 \right\}. \tag{2.4.1}$$

Then the linearized optimization subproblem necessary to compute the FW vertex takes the form

$$\min \left\{ \sum_s \sum_{i,j} c_{ij} x_{ij}^s : \sum_i x_{\ell i}^s - \sum_j x_{j\ell}^s = D(\ell, s),\ \ell \neq s,\ x_{ij}^s \geq 0 \right\} \tag{2.4.2}$$

and can be split in n shortest paths subproblems, each of the form

$$\min\left\{\sum_{i,j} c_{ij}x_{ij}^s : \sum_i x_{\ell i}^s - \sum_j x_{j\ell}^s = D(\ell, s),\ \ell \neq s,\ x_{ij}^s \geq 0\right\} \tag{2.4.3}$$

for a fixed $s \in [1:n]$, with c_{ij} the first-order derivative of f_{ij} (see [105] for further details). A number of FW variants were proposed in the literature for efficiently handling this kind of problems (see, e.g., [31,105,164,185,236] and references therein for further details). Some of those variants represent a good (if not the best) choice when low or medium precision is required in the solution of the problem [202].

In the more recent work [142] a FW variant also solving a shortest path subproblem at each iteration was applied to image and video co-localization.

2.4.2 Submodular optimization

Given a finite set V, a function $r : 2^V \to \mathbb{R}$ is said to be submodular if for every $A, B \subset V$

$$r(A) + r(B) \geq r(A \cup B) + r(A \cap B)\,. \tag{2.4.4}$$

As is common practice in the optimization literature (see e.g. [21, Section 2.1]), here we always assume $s(\emptyset) = 0$. A number of machine learning problems, including image segmentation and sensor placement, can be cast as minimization of a submodular function (see, e.g., [21,69] and references therein for further details):

$$\min_{A \subseteq V} r(A)\,. \tag{2.4.5}$$

Submodular optimization can also be seen as a more general way to relate combinatorial problems to convexity, for example for structured sparsity [21, 136]. By a theorem from [104], problem (2.4.5) can be in turn reduced to an minimum norm point problem over the base polytope

$$B(G) = \{s \in \mathbb{R}^V : \sum_{a \in A} s_a \leq r(A) \text{ for all } A \subseteq V\,, \sum_{a \in V} s_a = r(V)\}\,. \tag{2.4.6}$$

For this polytope, linear optimization can be achieved with a simple greedy algorithm. More precisely, consider the LP

$$\max_{s \in B(F)} w^\intercal s\,.$$

Then if the objective vector w has a negative component, the problem is clearly unbounded. Otherwise, a solution to the LP can be obtained by ordering w in decreasing manner as $w_{j_1} \geq w_{j_2} \geq \ldots \geq w_{j_n}$, and setting

$$s_{j_k} := r(\{j_1, \ldots, j_k\}) - r(\{j_1, \ldots, j_{k-1}\}), \tag{2.4.7}$$

for $k \in [1\!:\!n]$. We thus have a LMO with a $O(n \log n)$ cost. This is the reason why FW variants are widely used in the context of submodular optimization; further details can be found in, e.g., [21, 136].

2.4.3 LASSO problem

The LASSO, proposed by Tibshirani in 1996 [221], is a popular tool for sparse linear regression. Given the training set

$$T = \{(r_i, b_i) \in \mathbb{R}^n \times \mathbb{R} : i \in [1\!:\!m]\},$$

where $r_i^\intercal$ are the rows of an $m \times n$ matrix A, the goal is finding a sparse linear model (i.e., a model with a small number of non-zero parameters) describing the data. This problem is strictly connected with the Basis Pursuit Denoising (BPD) problem in signal analysis (see, e.g., [75]). In this case, given a discrete-time input signal b, and a *dictionary*

$$\{a_j \in \mathbb{R}^m \ : \ j \in [1\!:\!n]\}$$

of elementary discrete-time signals, usually called atoms (here a_j are the columns of a matrix A), the goal is finding a sparse linear combination of the atoms that *approximate* the real signal. From a purely formal point of view, LASSO and BPD problems are equivalent, and both can be formulated as follows:

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & f(x) := \|Ax - b\|_2^2 \\ s.t. & \|x\|_1 \leq \tau, \end{array} \tag{2.4.8}$$

where the parameter τ controls the amount of shrinkage that is applied to the model (related to sparsity, i.e., the number of nonzero components in x). The feasible set is

$$C = \{x \in \mathbb{R}^n : \|x\|_1 \leq \tau\} = \operatorname{conv}\{\pm \tau e_i : \ i \in [1\!:\!n]\}.$$

Thus we have the following LMO in this case:

$$\text{LMO}_C(\nabla f(x_k)) = \operatorname{sign}(-\nabla_{i_k} f(x_k)) \cdot \tau e_{i_k},$$

with $i_k \in \arg\max_i |\nabla_i f(x_k)|$. It is easy to see that the FW per-iteration cost is then $O(n)$. The peculiar structure of the problem makes FW variants well-suited for its solution. This is the reason why LASSO/BPD problems were considered in a number of FW-related papers (see, e.g., [135, 136, 157, 175]).

2.4.4 Matrix completion

Matrix completion is a widely studied problem that comes up in many areas of science and engineering, including collaborative filtering, machine learning, control, remote sensing, and computer vision (just to name a few; see also [64] and references therein). The goal is to retrieve a low rank matrix $X \in \mathbb{R}^{n_1 \times n_2}$ from a sparse set of observed matrix entries $\{U_{ij}\}_{(i,j)\in J}$ with $J \subset [1\!:\!n_1] \times [1\!:\!n_2]$. Thus the problem can be formulated as follows [103]:

$$\begin{array}{cl} \min\limits_{X \in \mathbb{R}^{n_1 \times n_2}} & f(X) := \sum\limits_{(i,j)\in J} (X_{ij} - U_{ij})^2 \\ s.t. & \operatorname{rank}(X) \le \delta, \end{array} \tag{2.4.9}$$

where the function f is given by the squared loss over the observed entries of the matrix and $\delta > 0$ is a parameter representing the assumed belief about the rank of the reconstructed matrix we want to get in the end. In practice, the low rank constraint is relaxed with a nuclear norm ball constraint, where we recall that the nuclear norm $\|X\|_*$ of a matrix X is equal the sum of its singular values. Thus we get the following convex optimization problem:

$$\begin{array}{cl} \min\limits_{X \in \mathbb{R}^{n_1 \times n_2}} & \sum\limits_{(i,j)\in J} (X_{ij} - U_{ij})^2 \\ s.t. & \|X\|_* \le \delta\,. \end{array} \tag{2.4.10}$$

The feasible set is the convex hull of rank-one matrices:

$$\begin{aligned} C &= \{X \in \mathbb{R}^{n_1 \times n_2} : \|X\|_* \le \delta\} \\ &= \operatorname{conv}\{\delta u v^\intercal : u \in \mathbb{R}^{n_1}, v \in \mathbb{R}^{n_2},\ \|u\| = \|v\| = 1\}\,. \end{aligned}$$

If we indicate with A_J the matrix that coincides with A on the indices J and is zero otherwise, then we can write $\nabla f(X) = 2\,(X - U)_J$. Thus we have the following LMO in this case:

$$\mathrm{LMO}_C(\nabla f(X_k)) \in \arg\min\{\operatorname{tr}(\nabla f(X_k)^\intercal X) : \|X\|_* \le \delta\}\,, \tag{2.4.11}$$

which boils down to computing the gradient, and the rank-one matrix $\delta u_1 v_1^\intercal$, with u_1, v_1 right and left singular vectors corresponding to the top singular value of $-\nabla f(X_k)$. Consequently, the FW method at a given iteration approximately reconstructs the target matrix as a sparse combination of rank-1 matrices. Furthermore, as the gradient matrix is sparse (it only has $|J|$ non-zero entries) storage and approximate singular vector computations can be performed much more efficiently than for dense matrices[2]. A number of FW variants has hence been proposed in the literature for solving this problem (see, e.g., [103, 135, 136]).

2.4.5 Adversarial attacks in machine learning

Adversarial examples are maliciously perturbed inputs designed to mislead a properly trained learning machine at test time. An *adversarial attack* hence consists in taking a correctly classified data point x_0 and slightly modifying it to create a new data point that leads the considered model to misclassification (see, e.g., [67, 73, 112] for further details). A possible formulation of the problem (see, e.g., [72, 112]) is given by the so called *maximum allowable ℓ_p-norm attack* that is,

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \ & f(x_0 + x) \\ s.t. \quad & \|x\|_p \le \varepsilon\,, \end{aligned} \tag{2.4.12}$$

where f is a suitably chosen attack loss function, x_0 is a correctly classified data point, x represents the additive noise/perturbation, $\varepsilon > 0$ denotes the magnitude of the attack, and $p \ge 1$. It is easy to see that the LMO has a cost $O(n)$. If x_0 is a feature vector of a dog image correctly classified by our learning machine, our adversarial attack hence suitably perturbs the feature vector (using the noise vector x), thus getting a new feature vector $x_0 + x$ classified, e.g., as a cat. In case a target adversarial class is specified by the attacker, we have a *targeted attack*. In some scenarios, the goal may not be to push x_0 to a specific target class, but rather push it away from its original class. In this case we have a so called *untargeted attack*. The attack function f will hence be chosen depending on the kind of attack we aim to perform over the considered model. Due to its specific structure, problem (2.4.12) can be nicely handled by means of tailored FW variants. Some FW frameworks for adversarial attacks were recently described in, e.g., [72, 147, 213].

[2] Details related to the LMO cost can be found in, e.g., [136].

2.4.6 Minimum enclosing ball

Given a set of points $P = \{p_1, \dots, p_n\} \subset \mathbb{R}^d$, the minimum enclosing ball problem (MEB, see, e.g., [78, 246]) consists in finding the smallest ball containing P. Such a problem models numerous important applications in clustering, nearest neighbor search, data classification, machine learning, facility location, collision detection, and computer graphics, to name just a few. We refer the reader to [155] and the references therein for further details. Denoting by $c \in \mathbb{R}^d$ the center and by $\sqrt{\gamma}$ (with $\gamma \geq 0$) the radius of the ball, a convex quadratic formulation for this problem is

$$\min_{(c,\gamma)\in\mathbb{R}^d\times\mathbb{R}} \gamma \tag{2.4.13}$$

$$s.t. \quad \|p_i - c\|^2 \leq \gamma\,, \quad \text{all } i \in [1\!:\!n]\,. \tag{2.4.14}$$

This problem can be formulated via Lagrangian duality as a convex *Standard Quadratic Optimization Problem* (StQP, see, e.g. [44])

$$\min\{x^\intercal A^\intercal Ax - b^\intercal x : x \in \Delta_{n-1}\} \tag{2.4.15}$$

with $A = [p_1, ..., p_n]$ and $b^\intercal = [p_1^\intercal p_1, \dots, p_n^\intercal p_n]$. The feasible set is the standard simplex Δ_{n-1}, and the LMO is defined as follows:

$$\text{LMO}_{\Delta_{n-1}}(\nabla f(x_k)) = e_{i_k},$$

with $i_k \in \arg\min_i \nabla_i f(x_k)$. It is easy to see that cost per iteration is $O(n)$. When applied to (2.4.15), the FW method can find an ε-cluster in $O(\frac{1}{\varepsilon})$, where an ε-cluster is a subset P' of P such that the MEB of P' dilated by $1 + \varepsilon$ contains P [78]. The set P' is given by the atoms in P selected by the LMO in the first $O(\frac{1}{\varepsilon})$ iterations. Further details related to the connections between FW methods and MEB problems can be found in, e.g., [5, 6, 78] and references therein.

2.4.7 Training linear Support Vector Machines

Support Vector Machines (SVMs) represent a very important class of machine learning tools (see, e.g., [226] for further details). Given a labeled set of data points, usually called *training set*:

$$TS = \{(p_i, y_i),\ p_i \in \mathbb{R}^d,\ y_i \in \{-1, 1\},\ i = 1, \dots, n\},$$

the linear SVM training problem consists in finding a linear classifier $w \in \mathbb{R}^d$ such that the label y_i can be deduced with the "highest possible confidence" from $w^\intercal p_i$. A convex quadratic formulation for this problem is the following [78]:

$$\begin{array}{ll} \min\limits_{w\in\mathbb{R}^d,\rho\in\mathbb{R}} & \rho + \frac{\|w\|^2}{2} \\ s.t. & \rho + y_i\, w^\intercal p_i \geq 0\,, \quad \text{all } i \in [1\!:\!n]\,, \end{array} \tag{2.4.16}$$

where the slack variable ρ stands for the negative margin and we can have $\rho < 0$ if and only if there exists an exact linear classifier, i.e. w such that $w^\intercal p_i = \text{sign}(y_i)$. The dual of (2.4.16) is again an StQP:

$$\min\{x^\intercal A^\intercal Ax : x \in \Delta_{n-1}\} \tag{2.4.17}$$

with $A = [y_1p_1, ..., y_np_n]$. Notice that problem (2.4.17) is equivalent to an MNP problem on $\text{conv}\{y_ip_i : i \in [1:n]\}$, see Section 2.8.2 below. Some FW variants (like, e.g., the Pairwise Frank-Wolfe) are closely related to classical working set algorithms, such as the SMO algorithm used to train SVMs [157]. Further details on FW methods for SVM training problems can be found in, e.g., [78, 135].

2.4.8 Finding maximal cliques in graphs

In the context of network analysis the clique model refers to subsets with every two elements in a direct relationship. Let $G = (V, E)$ be a simple undirected graph with V and E set of vertices and edges, respectively. A clique in G is a subset $C \subseteq V$ such that $(i, j) \in E$ for each $(i, j) \in C$, with $i \neq j$. The goal in finding a clique C such that $|C|$ is maximal (i.e., it is not contained in any strictly larger clique). This corresponds to find a local minimum for the following equivalent (this time non-convex) StQP (see, e.g., [40, 43, 133] for further details):

$$\max\left\{x^\intercal A_G x + \frac{1}{2}\|x\|^2 : x \in \Delta_{n-1}\right\} \tag{2.4.18}$$

where A_G is the adjacency matrix of G. Due to the peculiar structure of the problem, FW methods can be fruitfully used to find maximal cliques, (see, e.g., [133]). In Chapter 5, the application of a FW variant to a generalization of (2.4.18) will be discussed.

2.4.9 Finding sparse points in a set

Given a non-empty polyhedron $P \subset \mathbb{R}^n$, the goal is finding a sparse point $x \in P$ (i.e., a point with as many zero components as possible). This sparse optimization problem can be used to model a number of real-world applications in fields like, e.g., machine learning, pattern recognition and signal processing (see [207] and references therein). Ideally, what we would like to get is an optimal solution for the following problem:

$$\min \{\|x\|_0 : x \in P\}\,. \tag{2.4.19}$$

Since the zero norm is non-smooth, a standard procedure is to replace the original formulation (2.4.19) with an equivalent concave optimization problem of the form:

$$\min \left\{ \sum_{i=1}^{n} \phi(y_i) : x \in P,\ -y \le x \le y \right\}, \tag{2.4.20}$$

where $\phi : [0, +\infty[\ \to \mathbb{R}$ is a suitably chosen smooth concave univariate function bounded from below, like, e.g.,

$$\phi(t) = (1 - e^{-\alpha t})\,,$$

with α a large enough positive parameter (see, e.g., [181, 207] for further details). The LMO in this case gives a vertex solution for the linear programming problem:

$$\min \left\{ c_k^\intercal y : x \in P,\ -y \le x \le y \right\},$$

with $(c_k)_i$ the first-order derivative of ϕ calculated in $(y_k)_i$. Variants of the unit-stepsize FW method have been proposed in the literature (see, e.g., [181, 207]) to tackle the smooth equivalent formulation (2.4.20).

2.5 Stepsizes

Popular rules for determining the stepsize are:

- unit stepsize:

$$\alpha_k = 1,$$

  mainly used when the problem has a concave objective function. Finite convergence can be proved, under suitable assumptions, both for the unit-stepsize FW and some of its variants described in the literature (see, e.g., [207] for further details).

- *diminishing stepsize*:

$$\alpha_k = \frac{2}{k+2}\,, \tag{2.5.1}$$

  mainly used for the classic FW (see, e.g., [102, 136]).

- *exact line search*:

$$\alpha_k = \min \underset{\alpha \in [0, \alpha_k^{\max}]}{\arg\min}\ \varphi(\alpha) \quad \text{with } \varphi(\alpha) := f(x_k + \alpha\, d_k)\,, \tag{2.5.2}$$

  where we pick the smallest minimizer of the function φ for the sake of being well-defined even in rare cases of ties (see, e.g., [47, 157]).

- *Armijo line search:* the method iteratively shrinks the step size in order to guarantee a sufficient reduction of the objective function. It represents a good way to replace exact line search in cases when it gets too costly. In practice, we fix parameters $\delta \in (0, 1)$ and $\gamma \in (0, \frac{1}{2})$, then try steps $\alpha = \delta^m \alpha_k^{\max}$ with $m \in \{0, 1, 2, \dots\}$ until the sufficient decrease inequality

$$f(x_k + \alpha\, d_k) \le f(x_k) + \gamma\alpha\, \nabla f(x_k)^\intercal d_k \tag{2.5.3}$$

  holds, and set $\alpha_k = \alpha$ (see, e.g., [46] and references therein).

- *Lipschitz constant dependent step size:*

$$\alpha_k = \alpha_k(L) := \min \left\{ -\frac{\nabla f(x_k)^\intercal d_k}{L\|d_k\|^2}, \alpha_k^{max} \right\}\,, \tag{2.5.4}$$

  with L the Lipschitz constant of ∇f (see, e.g., [47, 201]).

The Lipschitz constant dependent step size can be seen as the minimizer of the quadratic model $m_k(\cdot; L)$ overestimating f along the line $x_k + \alpha\, d_k$:

$$m_k(\alpha; L) = f(x_k) + \alpha\, \nabla f(x_k)^\intercal d_k + \frac{L\alpha^2}{2}\, \|d_k\|^2 \ge f(x_k + \alpha\, d_k)\,, \tag{2.5.5}$$

where the inequality follows by the standard Descent Lemma.

In case L is unknown, it is even possible to approximate L using a backtracking line search (see, e.g., [150, 201]).

We now report a lower bound for the improvement on the objective obtained with the stepsize (2.5.4), often used in the convergence analysis.

Lemma 2.5.1. *If* α_k *is given by* (2.5.4) *and* $\alpha_k < \alpha_k^{\max}$ *then*

$$f(x_{k+1}) \le f(x_k) - \frac{1}{2L}(\nabla f(x_k)^\intercal \widehat{d}_k)^2\,. \tag{2.5.6}$$

Proof. We have

$$\begin{aligned} f(x_k + \alpha_k\, d_k) &\le f(x_k) + \alpha_k \nabla f(x_k)^\intercal d_k + \tfrac{L\alpha_k^2}{2}\, \|d_k\|^2 \\ &= f(x_k) - \tfrac{(\nabla f(x_k)^\intercal d_k)^2}{2L\|d_k\|^2} = f(x_k) - \tfrac{1}{2L}(\nabla f(x_k)^\intercal \widehat{d_k})^2\,, \end{aligned} \tag{2.5.7}$$

where we used the standard Descent Lemma in the inequality. □

2.6 Properties of the FW method and its variants

2.6.1 The FW gap

A key parameter often used as a measure of convergence is the FW gap

$$G(x) = \max_{s \in \Omega} -\nabla f(x)^\intercal (s - x) \,, \tag{2.6.1}$$

which is always nonnegative and equal to 0 only in first order stationary points. This gap is, by definition, readily available during the algorithm. If f is convex, using that $\nabla f(x)$ is a subgradient we obtain

$$G(x) \geq -\nabla f(x)^\intercal (x^* - x) \geq f(x) - f^* \,, \tag{2.6.2}$$

so that $G(x)$ is an upper bound on the optimality gap at x. Furthermore, $G(x)$ is a special case of the Fenchel duality gap [158].

If $\Omega = \Delta_{n-1}$ is the simplex, then G is related to the Wolfe dual as defined in [78]. Indeed, this variant of Wolfe's dual reads

$$\begin{aligned} \max \;& f(x) + \lambda(e^\intercal x - 1) - u^\intercal x \\ \text{s.t.} \;& \nabla_i f(x) - u_i + \lambda = 0 \,, \quad i \in [1 \colon n] \,, \\ & (x, u, \lambda) \in \mathbb{R}^n \times \mathbb{R}^n_+ \times \mathbb{R} \end{aligned} \tag{2.6.3}$$

and for a fixed $x \in \mathbb{R}^n$, the optimal values of (u, λ) are

$$\lambda_x = -\min_j \nabla_j f(x) \,, \quad u_i(x) := \nabla_i f(x) - \min_j \nabla_j f(x) \geq 0 \,.$$

Performing maximization in problem (2.6.3) iteratively, first for (u, λ) and then for x, this implies that (2.6.3) is equivalent to

$$\begin{aligned} & \max\nolimits_{x \in \mathbb{R}^n} \left[f(x) + \lambda_x(e^\intercal x - 1) - u(x)^\intercal x \right] \\ = \; & \max\nolimits_{x \in \mathbb{R}^n} \left[f(x) - \max\nolimits_j (e_j - x)^\intercal \nabla f(x) \right] = \max\nolimits_{x \in \mathbb{R}^n} \left[f(x) - G(x) \right] \,. \end{aligned} \tag{2.6.4}$$

Furthermore, since Slater's condition is satisfied, strong duality holds by Slater's theorem [57], resulting in $G(x^*) = 0$ for every solution x^* of the primal problem.

The FW gap is related to several other measures of convergence (see e.g. [160, Section 7.5.1]). First, consider the projected gradient

$$\widetilde{g}_k := \pi_\Omega(x_k - \nabla f(x_k)) - x_k \,. \tag{2.6.5}$$

with π_B the projection on a convex and closed subset $B \subseteq \mathbb{R}^n$. We have $\|\widetilde{g}_k\| = 0$ if and only if x_k is stationary, with

$$\begin{aligned} \|\widetilde{g}_k\|^2 &= \widetilde{g}_k^\intercal \widetilde{g}_k \le \widetilde{g}_k^\intercal [(x_k - \nabla f(x_k)) - \pi_\Omega(x_k - \nabla f(x_k))] + \widetilde{g}_k^\intercal \widetilde{g}_k \\ &= -\widetilde{g}_k^\intercal \nabla f(x_k) = -(\pi_\Omega(x_k - \nabla f(x_k)) - x_k)^\intercal \nabla f(x_k) \\ &\le \max_{y \in \Omega} -(y - x_k)^\intercal \nabla f(x_k) = G(x_k) , \end{aligned} \tag{2.6.6}$$

where we used $[y - \pi_\Omega(x)]^\intercal [x - \pi_\Omega(x)] \le 0$ in the first inequality, with $x = x_k - \nabla f(x_k)$ and $y = x_k$.

Let now $N_\Omega(x)$ denote the normal cone to Ω at a point $x \in \Omega$:

$$N_\Omega(x) := \{r \in \mathbb{R}^n : r^\intercal (y - x) \le 0 \text{ for all } y \in \Omega\} . \tag{2.6.7}$$

First-order stationarity conditions are equivalent to $-\nabla f(x) \in N_\Omega(x)$, or

$$\text{dist}(N_\Omega(x), -\nabla f(x)) = \| - \nabla f(x) - \pi_{N_\Omega(x)}(-\nabla f(x))\| = 0 .$$

The FW gap provides a lower bound on the distance from the normal cone $\text{dist}(N_\Omega(x), -\nabla f(x))$, inflated by the diameter $D > 0$ of Ω, as follows:

$$\begin{aligned} G(x_k) &= -(s_k - x_k)^\intercal \nabla f(x_k) \\ &= (s_k - x_k)^\intercal [\pi_{N_\Omega(x_k)}(-\nabla f(x_k)) - (\pi_{N_\Omega(x_k)}(-\nabla f(x_k)) + \nabla f(x_k))] \\ &\le \|s_k - x_k\| \, \|\pi_{N_\Omega(x_k)}(-\nabla f(x_k)) + \nabla f(x_k)\| \\ &\le D \, \text{dist}(N_\Omega(x_k), -\nabla f(x_k)) , \end{aligned} \tag{2.6.8}$$

where in the first inequality we used $(s_k - x_k)^\intercal [\pi_{N_\Omega(x_k)}(-\nabla f(x_k))] \le 0$ together with the Cauchy-Schwarz inequality, and $\|s_k - x_k\| \le D$ in the second.

2.6.2 $O(1/k)$ rate for convex objectives

If f is non-convex, it is possible to prove a $O(1/\sqrt{k})$ rate for $\min_{i \in [1:k]} G(x_i)$ (see, e.g., [156]). On the other hand, if f is convex, we have an $O(1/k)$ rate on the optimality gap (see, e.g., [101, 166]) for all the stepsizes discussed in Section 2.5. Here we include a proof for the Lipschitz constant dependent stepsize α_k given by (2.5.4).

Theorem 2.6.1. *If f is convex and the stepsize is given by* (2.5.4)*, then for every $k \ge 1$*

$$f(x_k) - f^* \le \frac{2LD^2}{k+2} . \tag{2.6.9}$$

Before proving the theorem we prove a lemma concerning the decrease of the objective in the case of a full FW step, that is a step with $d_k = d_k^{FW}$ and with α_k equal to 1, the maximal feasible stepsize.

Lemma 2.6.2. *If* $\alpha_k = 1$ *and* $d_k = d_k^{FW}$ *then*

$$f(x_{k+1}) - f^* \leq \frac{1}{2} \min \left\{ L\|d_k\|^2, f(x_k) - f^* \right\} . \tag{2.6.10}$$

Proof. If $\alpha_k = 1 = \alpha_k^{\max}$ then by Definitions (2.3.2) and (2.6.1)

$$G(x_k) = -\nabla f(x_k)^\intercal d_k \geq L\|d_k\|^2 , \tag{2.6.11}$$

the last inequality following by Definition (2.5.4) and the assumption that $\alpha_k = 1$. By the standard Descent Lemma it also follows

$$f(x_{k+1}) - f^* = f(x_k + d_k) - f^* \leq f(x_k) - f^* + \nabla f(x_k)^\intercal d_k + \frac{L}{2} \|d_k\|^2 . \tag{2.6.12}$$

Considering the definition of d_k and convexity of f, we get

$$f(x_k) - f^* + \nabla f(x_k)^\intercal d_k \leq f(x_k) - f^* + \nabla f(x_k)^\intercal (x^* - x_k) \leq 0 ,$$

so that (2.6.12) entails $f(x_{k+1}) - f^* \leq \frac{L}{2} \|d_k\|^2$. To conclude, it suffices to apply to the RHS of (2.6.12) the inequality

$$f(x_k) - f^* + \nabla f(x_k)^\intercal d_k + \tfrac{L}{2} \|d_k\|^2 \leq f(x_k) - f^* - \tfrac{1}{2} G(x_k) \leq \tfrac{f(x_k) - f^*}{2} \tag{2.6.13}$$

where we used (2.6.11) in the first inequality and $G(x_k) \geq f(x_k) - f^*$ in the second. □

We can now proceed with the proof of the main result.

Theorem 2.6.1. For $k = 0$ and $\alpha_0 = 1$ then by Lemma 2.6.2

$$f(x_1) - f^* \leq \frac{L\|d_0\|^2}{2} \leq \frac{LD^2}{2} . \tag{2.6.14}$$

If $\alpha_0 < 1$ then

$$f(x_0) - f^* \leq G(x_0) < L\|d_0\|^2 \leq LD^2 . \tag{2.6.15}$$

Therefore in both cases (2.5.6) holds for $k = 0$.
Reasoning by induction, if (2.6.9) holds for k with $\alpha_k = 1$, then the claim is clear

by (2.6.10). On the other hand, if $\alpha_k < \alpha_k^{\max} = 1$ then by Lemma 2.5.1, we have

$$\begin{aligned} f(x_{k+1}) - f^* &\leq f(x_k) - f^* - \tfrac{1}{2L}(\nabla f(x_k)^\intercal \widehat{d}_k)^2 \\ &\leq f(x_k) - f^* - \tfrac{(\nabla f(x_k)^\intercal d_k)^2}{2LD^2} \\ &\leq f(x_k) - f^* - \tfrac{(f(x_k)-f^*)^2}{2LD^2} \\ &= (f(x_k) - f^*)(1 - \tfrac{f(x_k)-f^*}{2LD^2}) \leq \tfrac{2LD^2}{k+3}, \end{aligned} \tag{2.6.16}$$

where we used $\|d_k\| \leq D$ in the second inequality, $\nabla f(x_k)^\intercal d_k = G(x_k) \geq f(x_k) - f^*$ in the third inequality; and the last inequality follows by induction hypothesis. □

As can be easily seen from above argument, the convergence rate of $O(1/k)$ is true also in more abstract normed spaces than $\mathbb{R}^n$, e.g. when Ω is a convex and weakly compact subset of a Banach space (see, e.g., [89, 95]). A generalization for some unbounded sets is given in [100]. The bound is tight due to a zigzagging behaviour of the method near solutions on the boundary, leading to a rate of $\Omega(1/k^{1+\delta})$ for every $\delta > 0$ (see [65] for further details), when the objective is a strictly convex quadratic function and the domain is a polytope.
Also the minimum FW gap $\min_{i\in[0:k]} G(x_i)$ converges at a rate of $O(1/k)$ (see [102, 136]). In [102], a broad class of stepsizes is examined, including $\alpha_k = \frac{1}{k+1}$ and $\alpha_k = \bar{\alpha}$ constant. For these stepsizes a convergence rate of $O\left(\frac{\ln(k)}{k}\right)$ is proved.

2.6.3 Variants

We present here some active set FW variants. Such variants mostly aim to improve over the $O(1/k)$ rate and also ensure support identification in finite time. They generate a sequence of active sets $\{A_k\}$, such that $x_k \in \text{conv}(A_k)$, and define alternative directions making use of these active sets (see Figure 2.1).

For the *pairwise FW (PFW)* and the *AFW* (see [78, 157]) we have that A_k must always be a subset of S_k, with x_k a convex combination of the elements in A_k. The away vertex v_k is then defined by

$$v_k \in \operatorname*{arg\,max}_{y\in A_k} \nabla f(x_k)^\intercal y\,. \tag{2.6.17}$$

The AFW direction, introduced in [237], is hence given by

$$\begin{aligned} d_k^{AS} &= x_k - v_k \\ d_k &\in \arg\max\{-\nabla f(x_k)^\intercal d : d \in \{d_k^{AS}, d_k^{FW}\}\}\,, \end{aligned} \tag{2.6.18}$$

while the PFW direction, as defined in [157] and inspired by the early work [184], is

$$d_k^{PFW} = d_k^{FW} + d_k^{AS} = s_k - v_k \,, \tag{2.6.19}$$

with s_k defined in (2.3.2).

The *FW method with in-face directions (FDFW)* (see [103, 116]), also known as Decomposition invariant Conditional Gradient (DiCG) when applied to polytopes [24], is defined exactly as the AFW, but with the minimal face $\mathcal{F}(x_k)$ of Ω containing x_k as the active set. The *extended FW (EFW)* was introduced in [126] and is also known as simplicial decomposition [231]. At every iteration the method minimizes the objective in the current active set A_{k+1}

$$x_{k+1} \in \underset{y \in \operatorname{conv}(A_{k+1})}{\arg\min} \; f(y) \,, \tag{2.6.20}$$

where $A_{k+1} \subseteq A_k \cup \{s_k\}$ (see, e.g., [78], Algorithm 4.2). A more general version of the EFW, only approximately minimizing on the current active set, was introduced in [157] under the name of fully corrective FW. In Table 2.1, we report the main features of the classic FW and of the variants under analysis.

Variant	Direction	Active set
FW	$d_k = d_k^{FW} = s_k - x_k, \quad s_k \in \arg\max\{\nabla f(x_k)^\intercal x : x \in \Omega\}$	-
AFW	$d_k \in \arg\max\{-\nabla f(x_k)^\intercal d : d \in \{x_k - v_k, d_k^{FW}\},\ v_k \in A_k\}$	$A_{k+1} \subseteq A_k \cup \{s_k\}$
PFW	$d_k = s_k - v_k, \quad v_k \in \arg\max\{\nabla f(x_k)^\intercal v_k : v_k \in A_k\}$	$A_{k+1} \subseteq A_k \cup \{s_k\}$
EFW	$d_k = y_k - x_k, \quad y_k \in \arg\min\{f(y) : y \in \operatorname{conv}(A_k)\}$	$A_{k+1} \subseteq A_k \cup \{s_k\}$
FDFW	$d_k \in \arg\max\{-\nabla f(x_k)^\intercal d : d \in \{x_k - v_k, d_k^{FW}\},\ v_k \in A_k\}$	$A_k = \mathcal{F}(x_k)$

Table 2.1: FW method and variants covered in this chapter.

2.6.4 Sparse approximation properties

As discussed in the previous section, for the classic FW method and the AFW, PFW, EFW variants x_k can always be written as a convex combination of elements in $A_k \subset S_k$, with $|A_k| \leq k+1$. Even for the FDFW we still have the weaker property that x_k must be an affine combination of elements in $A_k \subset A$ with $|A_k| \leq k+1$. It turns out that the convergence rate of methods with this property is $\Omega(\frac{1}{k})$ in high dimension. More precisely, if $\Omega = \mathrm{conv}(A)$ with A compact, the $O(1/k)$ rate of the classic FW method is worst case optimal given the sparsity constraint

$$x_k \in \mathrm{aff}(A_k) \text{ with } A_k \subset A,\ |A_k| \le k+1\,. \tag{2.6.21}$$

An example where the $O(1/k)$ rate is tight was presented in [136]. Let $\Omega = \Delta_{n-1}$ and $f(x) = \|x - \frac{1}{n}e\|^2$. Clearly, $f^* = 0$ with $x^* = \frac{1}{n}e$. Then it is easy to see that $\min\{f(x) - f^* : \|x\|_0 \le k+1\} \ge \frac{1}{k+1} - \frac{1}{n}$ for every $k \in \mathbb{N}$, so that in particular under (2.6.21) with $A_k = \{e_i : i \in [1\!:\!n]\}$, the rate of any FW variant must be $\Omega(\frac{1}{k})$.

2.6.5 Affine invariance

The FW method and the AFW, PFW, EFW are affine invariant [136]. More precisely, let $\mathbb{P}$ be a linear transformation, $\hat{f}$ be such that $\hat{f}(\mathbb{P}x) = f(x)$ and $\hat{\Omega} = \mathbb{P}(\Omega)$. Then for every sequence $\{x_k\}$ generated by the methods applied to (f, Ω), the sequence $\{y_k\} := \{\mathbb{P}x_k\}$ can be generated by the FW method with the same stepsizes applied to $(\hat{f}, \hat{\Omega})$. As a corollary, considering the special case where $\mathbb{P}$ is the matrix collecting the elements of A as columns, one can prove results on $\Omega = \Delta_{|A|-1}$ and generalize them to $\hat{\Omega} := \mathrm{conv}(A)$ by affine invariance.

An affine invariant convergence rate bound for convex objectives can be given using the curvature constant

$$\kappa_{f,\Omega} := \sup\left\{2\tfrac{f(\alpha y+(1-\alpha)x)-f(x)-\alpha\nabla f(x)^\intercal(y-x)}{\alpha^2} : \{x,y\} \subset \Omega,\ \alpha \in (0,1]\right\}\,. \tag{2.6.22}$$

It is easy to prove that $\kappa_{f,\Omega} \le LD^2$ if D is the diameter of Ω. In the special case where $\Omega = \Delta_{n-1}$ and $f(x) = x^\intercal \tilde{A}^\intercal \tilde{A}x + b^\intercal x$, then $\kappa_{f,\Omega} \le \mathrm{diam}(A\Delta_{n-1})^2$ for $A^\intercal = [\tilde{A}^\intercal, b]$; see [78].

When the method uses the stepsize sequence (2.5.1), it is possible to give the following affine invariant convergence rate bounds (see [102]):

$$\begin{aligned} f(x_k) - f^* &\le \frac{2\kappa_{f,\Omega}}{k+4}\,, \\ \min_{i\in]0:k]} G(x_i) &\le \frac{9\kappa_{f,\Omega}}{2k}\,, \end{aligned} \tag{2.6.23}$$

thus in particular slightly improving the rate we gave in Theorem 2.6.1 since we have that $\kappa_{f,\Omega} \le LD^2$.

2.6.6 Inexact linear oracle

In many real-world applications, linear subproblems can only be solved approximately. This is the reason why the convergence of FW variants is often analyzed under some error term for the linear minimization oracle (see, e.g., [58,59,102,136,154]). A common assumption, relaxing the FW vertex exact minimization property, is to have access to a point (usually a vertex) $\tilde{s}_k$ such that

$$\nabla f(x_k)^\intercal(\tilde{s}_k - x_k) \leq \min_{s\in\Omega} \nabla f(x_k)^\intercal(\{1, ..., -\}x_k) + \delta_k\,, \tag{2.6.24}$$

for a sequence $\{\delta_k\}$ of non negative approximation errors.
If the sequence $\{\delta_k\}$ is constant and equal to some $\delta > 0$, then trivially the lowest possible approximation error achieved by the FW method is δ. At the same time, [102, Theorem 5.1] implies a rate of $O(\frac{1}{k} + \delta)$ if the stepsize $\alpha_k = \frac{2}{k+2}$ is used.

The $O(1/k)$ rate can be instead retrieved by assuming that $\{\delta_k\}$ converges to 0 quickly enough, and in particular if

$$\delta_k = \frac{\delta \kappa_{f,C}}{k+2} \tag{2.6.25}$$

for a constant $\delta > 0$. Under (2.6.25), in [136] a convergence rate of

$$f(x_k) - f^* \leq \frac{2\kappa_{f,\Omega}}{k+2}(1+\delta) \tag{2.6.26}$$

was proved for the FW method with α_k given by exact line search or equal to $\frac{2}{k+2}$, as well as for the EFW.

A linearly convergent variant making use of an approximated linear oracle recycling previous solutions to the linear minimization subproblem is studied in [58]. In [102,125], the analysis of the classic FW method is extended to the case of inexact gradient information. In particular in [102], assuming the availability of the (δ, L) oracle introduced in [90], a convergence rate of $O(1/k + \delta k)$ is proved.

2.7 Improved rates for strongly convex objectives

2.7.1 Linear convergence for FW variants

In the rest of this section we assume that f is μ-strongly convex (1.2.5). We also assume that the stepsize is given by exact line search or by (2.5.4).

Under this assumption, an asymptotic linear convergence rate for the FDFW on polytopes was given in the early work [116]. Furthermore, in [109] a linearly

Method	Objective	Domain	Assumptions	Rate	Article
FW	NC	Generic	-	$\mathcal{O}(1/\sqrt{k})$	[156]
FW	C	Generic	-	$\mathcal{O}(1/k)$	[101]
FW	SC	Generic	$x^* \in \mathrm{ri}(\Omega)$	Linear	[116]
Variants	SC	Polytope	-	Linear	[157]
FW	SC	Strongly convex	-	$\mathcal{O}(1/k^2)$	[108]
FW	SC	Strongly convex	$\min \lVert\nabla f(x)\rVert > 0$	Linear	[89]

Table 2.2: Known convergence rates for the FW method and the variants covered in this chapter. NC, C and SC stand for non-convex, convex and strongly convex respectively.

convergent variant was proposed, making use however of an additional local linear minimization oracle.

Recent works obtain linear convergence rates by proving the condition

$$-\nabla f(x_k)^\intercal \widehat{d_k} \geq \frac{\tau}{\|x_k - x^*\|} \nabla f(x_k)^\intercal (x_k - x^*) \tag{2.7.1}$$

for some $\tau > 0$ and some $x^* \in \arg\min_{x \in C} f(x)$. As we shall see in the next lemma, under (2.7.1) it is not difficult to prove linear convergence rates in the number of *good steps*. These are FW steps with $\alpha_k = 1$ and steps in any descent direction with $\alpha_k < 1$.

Lemma 2.7.1. *If the step k is a good step and* (2.7.1) *holds, then*

$$h_{k+1} \leq \max\left\{\tfrac{1}{2}, 1 - \tfrac{\tau^2\mu}{L}\right\} h_k \,. \tag{2.7.2}$$

Proof. If the step k is a full FW step then Lemma 2.6.2 entails $h_{k+1} \leq \frac{1}{2} h_k$. In the remaining case, first observe that by strong convexity

$$\begin{aligned} f^* &= f(x^*) \geq f(x_k) + \nabla f(x_k)^\intercal (x^* - x_k) + \tfrac{\mu}{2}\|x_k - x^*\|^2 \\ &\geq \min_{\alpha \in \mathbb{R}} \left[f(x_k) + \alpha \nabla f(x_k)^\intercal (x^* - x_k) + \tfrac{\alpha^2\mu}{2}\|x_k - x^*\|^2 \right] \\ &= f(x_k) - \tfrac{1}{2\mu\|x_k - x^*\|^2} \left[\nabla f(x_k)^\intercal (x_k - x^*)\right]^2 \,, \end{aligned} \tag{2.7.3}$$

which means

$$h_k \leq \frac{1}{2\mu\|x_k - x^*\|^2} \left[\nabla f(x_k)^\intercal (x_k - x^*)\right]^2 \,. \tag{2.7.4}$$

We can then proceed using the bound (2.5.6) from Lemma 2.5.1 in the following way:

$$\begin{array}{rcl} h_{k+1} & = & f(x_{k+1}) - f^* \leq f(x_k) - f^* - \frac{1}{2L}\left[\nabla f(x_k)^\intercal \widehat{d_k}\right]^2 \\ & \leq & h_k - \frac{\tau^2}{2L\|x_k - x^*\|^2}\left[\nabla f(x_k)^\intercal (x_k - x^*)\right]^2 \\ & \leq & h_k\left(1 - \frac{\tau^2\mu}{L}\right), \end{array} \tag{2.7.5}$$

where we used (2.7.1) in the second inequality and (2.7.4) in the third one. □

As a corollary, under (2.7.1) we have the rate

$$f(x_k) - f^* = h_k \leq \max\left\{\frac{1}{2}, 1 - \frac{\tau^2\mu}{L}\right\}^{\gamma(k)} h_0 \tag{2.7.6}$$

for any method with non increasing $\{f(x_k)\}$ and following Algorithm 1, with $\gamma(k) \leq k$ an integer denoting the number of good steps until step k. It turns out that for all the variants we introduced in this chapter we have $\gamma(k) \geq Tk$ for some constant $T > 0$. When x^* is in the relative interior of Ω, the FW method satisfies (2.7.1) and we have the following result (see [116, 157]):

Theorem 2.7.2. *If* $x^* \in \mathrm{ri}(\Omega)$*, then*

$$f(x_k) - f^* \leq \left[1 - \frac{\mu}{L}\left(\frac{\mathrm{dist}(x^*, \partial\Omega)}{D}\right)^2\right]^k (f(x_0) - f^*)\,. \tag{2.7.7}$$

Proof. We can assume for simplicity $\mathrm{int}(\Omega) \neq \emptyset$, since otherwise we can restrict ourselves to the affine hull of Ω. Let $\delta = \mathrm{dist}(x^*, \partial\Omega)$ and $g = -\nabla f(x_k)$. First, by assumption we have $x^* + \delta\widehat{g} \in \Omega$. Therefore

$$g^\intercal d_k^{FW} \geq g^\intercal((x^* + \delta\widehat{g}) - x) = \delta g^\intercal \widehat{g} + g^\intercal(x^* - x) \geq \delta\|g\| + f(x) - f^* \geq \delta\|g\|\,, \tag{2.7.8}$$

where we used $x^* + \delta\widehat{g} \in \Omega$ in the first inequality and convexity in the second. We can conclude

$$g^\intercal \frac{d_k^{FW}}{\|d_k^{FW}\|} \geq g^\intercal \frac{d_k^{FW}}{D} \geq \frac{\delta}{D}\|g\| \geq \frac{\delta}{D} g^\intercal\left(\frac{x_k - x^*}{\|x_k - x^*\|}\right). \tag{2.7.9}$$

The thesis follows by Lemma 2.7.1, noticing that for $\tau = \frac{\mathrm{dist}(x^*, \partial\Omega)}{D} \leq \frac{1}{2}$ we have $1 - \tau^2\frac{\mu}{L} > \frac{1}{2}$. □

In [157], the authors proved that directions generated by the AFW and the PFW on polytopes satisfy condition (2.7.1), with $\tau = \mathrm{PWidth}(A)/D$ and $\mathrm{PWidth}(A)$, pyramidal width of A. While $\mathrm{PWidth}(A)$ was originally defined with a rather complex minmax expression, in [200] it was then proved

$$\mathrm{PWidth}(A) = \min_{F \in \mathrm{faces}(C)} \mathrm{dist}(F, \mathrm{conv}(A \setminus F)) \,. \tag{2.7.10}$$

This quantity can be explicitly computed in a few special cases. For $A = \{0,1\}^n$ we have $\mathrm{PWidth}(A) = 1/\sqrt{n}$, while for $A = \{e_i\}_{i \in [1:n]}$ (so that Ω is the $n-1$ dimensional simplex)

$$\mathrm{PWidth}(A) = \begin{cases} \frac{2}{\sqrt{n}} & \text{if } n \text{ is even} \\ \frac{2}{\sqrt{n-1/n}} & \text{if } n \text{ is odd.} \end{cases} \tag{2.7.11}$$

Conditions like (2.7.1) with τ dependent on the number of vertices used to represent x_k as a convex combination were given in [24] and [27] for the FDFW and the PFW respectively. In particular, in [27] a geometric constant Ω_Ω called vertex-facet distance was defined as

$$\Omega_\Omega = \min\{\mathrm{dist}(v, H) : v \in V(\Omega)\,, H \in \mathcal{H}(\Omega),\ v \notin H\}\,, \tag{2.7.12}$$

with $V(\Omega)$ the set of vertices of Ω, and $\mathcal{H}(\Omega)$ the set of supporting hyperplanes of Ω (containing a facet of Ω). Then condition (2.7.1) is satisfied for $\tau = \Omega_\Omega/s$, with d_k the PFW direction and s the number of points used in the active set A_k.
In [24], a geometric constant H_s was defined depending on the minimum number s of vertices needed to represent the current point x_k, as well as on the proper[3] inequalities $q_i^\intercal x \leq b_i$, $i \in [1:m]$, appearing in a description of Ω. For each of these inequalities the *second gap* g_i was defined as

$$g_i = \max_{v \in V(\Omega)} q_i^\intercal v - \operatorname*{secondmax}_{v \in V(\Omega)} q_i^\intercal v\,, \quad i \in [1:m]\,, \tag{2.7.13}$$

with the secondmax function giving the second largest value achieved by the argument. Then H_s is defined as

$$H_s := \max\left\{ \sum_{j=1}^{n} \left(\sum_{i \in S} \frac{a_{ij}}{g_i} \right)^2 : S \in \binom{[1:m]}{s} \right\}. \tag{2.7.14}$$

The arguments used in the paper imply that (2.7.1) holds with $\tau = \frac{1}{2D\sqrt{H_s}}$ if d_k is a FDFW direction and x_k the convex combination of at most s vertices. We refer the reader to [200] and [206] for additional results on these and related constants.

[3] i.e., those inequalities strictly satisfied for some $x \in \Omega$.

The linear convergence results for strongly convex objectives are extended to compositions of strongly convex objectives with affine transformations in [27], [157], [200]. In [117], the linear convergence results for the AFW and the FW method with minimum in the interior are extended with respect to a generalized condition number $L_{f,\Omega,D}/\mu_{f,\Omega,D}$, with D a distance function on Ω.

For the AFW, the PFW and the FDFW, linear rates with no bad steps ($\gamma(k) = k$) are given in [209] (see Chapter 3) for non-convex objectives satisfying a Kurdyka-Łojasiewicz inequality. In [208], condition (2.7.1) was proved for the FW direction and orthographic retractions on some convex sets with smooth boundary. The work [79] introduces a new FW variant using a subroutine to align the descent direction with the projection on the tangent cone of the negative gradient, thus implicitly maximizing τ in (2.7.1).

2.7.2 Strongly convex domains

When Ω is strongly convex we have a $O(1/k^2)$ rate (see, e.g., [108, 149]) for the classic FW method. Furthermore, when Ω is β_Ω-strongly convex and $\|\nabla f(x)\| \geq c > 0$, then we have the linear convergence rate (see [89, 94, 150, 166])

$$h_{k+1} \leq \max\left\{\tfrac{1}{2}, 1 - \tfrac{L}{2c\beta_\Omega}\right\} h_k \,. \tag{2.7.15}$$

Finally, it is possible to interpolate between the $O(1/k^2)$ rate of the strongly convex setting and the $O(1/k)$ rate of the general convex one by relaxing strong convexity of the objective with Hölderian error bounds [243] and also by relaxing strong convexity of the domain with uniform convexity [149].

2.8 Extensions

2.8.1 Block coordinate Frank-Wolfe method

The block coordinate FW (BCFW) was introduced in [158] for block product domains of the form $\Omega = \Omega^{(1)} \times ... \times \Omega^{(m)} \subseteq \mathbb{R}^{n_1+...+n_m}$, and applied to structured SVM training. The algorithm operates by selecting a random block and performing a FW step in that block. Formally, for $s \in \mathbb{R}^{m_i}$ let $s^{(i)} \in \mathbb{R}^n$ be the vector with all blocks equal to 0 except for the i-th block equal to s. We can write the direction of the BCFW as

$$\begin{aligned} d_k &= s_k^{(i)} - x_k \\ s_k &\in \operatorname*{arg\,min}_{s \in \Omega^{(i)}} \nabla f(x_k)^\intercal s^{(i)} \end{aligned} \tag{2.8.1}$$

for a random index $i \in [1:n]$.
In [158], a convergence rate of

$$\mathbb{E}[f(x_k)] - f^* \leq \frac{2Km}{k+2m} \tag{2.8.2}$$

is proved, for $K = h_0 + \kappa_f^{\otimes}$, with $\kappa_f^{\otimes}$ the product domain curvature constant, defined as $\kappa_f^{\otimes} = \sum \kappa_f^{\otimes,i}$ where $\kappa_f^{\otimes,i}$ are the curvature constants of the objective fixing the blocks outside $\Omega^{(i)}$:

$$\kappa_f^{\otimes,i} := \sup \left\{ 2\tfrac{f(x+\alpha d^{(i)}) - f(x) - \alpha \nabla f(x)^\intercal d^{(i)}}{\alpha^2} : d \in \Omega - x,\, x \in \Omega,\, \alpha \in (0,1] \right\} . \tag{2.8.3}$$

An asynchronous and parallel generalization for this method was given in [234]. This version assumes that a cloud oracle is available, modeling a set of worker nodes each sending information to a server at different times. This information consists of an index i and the following LMO on $\Omega^{(i)}$:

$$s_{(i)} \in \operatorname*{arg\,min}_{s \in \Omega^{(i)}} \nabla f(x_{\widetilde{k}})^\intercal s^{(i)} . \tag{2.8.4}$$

The algorithm is called asynchronous because $\widetilde{k}$ can be smaller than k, modeling a delay in the information sent by the node. Once the server has collected a minibatch S of τ distinct indexes (overwriting repetitions), the descent direction is defined as

$$d_k = \sum_{i \in S} s_{(i)}^{(i)} , \tag{2.8.5}$$

If the indices sent by the nodes are i.i.d., then under suitable assumptions on the delay, a convergence rate of

$$\mathbb{E}[f(x_k)] - f^* \leq \frac{2mK_\tau}{\tau^2 k + 2m} \tag{2.8.6}$$

can be proved, where $K_\tau = m\kappa_{f,\tau}^{\otimes}(1+\delta) + h_0$ for δ depending on the delay error, with $\kappa_{f,\tau}^{\otimes}$ the average curvature constant in a minibatch keeping all the components not in the minibatch fixed.

In [197], several improvements are proposed for the BCFW, including an adaptive criterion to prioritize blocks based on their FW gap, and block coordinate versions of the AFW and the PFW variants.

In [214], a multi plane BCFW approach is proposed in the specific case of the structured SVM, based on caching supporting planes in the primal, corresponding to block linear minimizers in the dual. In [28], the duality for structured SVM between BCFW and stochastic subgradient descent is exploited to define a learning rate schedule for neural networks based only on one hyper parameter. The block coordinate approach is extended to the generalized FW in [26], with coordinates however picked in a cyclic order.

2.8.2 Variants for the min-norm point problem

Consider the min-norm point (MNP) problem

$$\min_{x \in \Omega} \|x\|_* \,, \tag{2.8.7}$$

with Ω a closed convex subset of $\mathbb{R}^n$ and $\|\cdot\|_*$ a norm on $\mathbb{R}^n$. In [238], a FW variant is introduced to solve the problem when Ω is a polytope and $\|\cdot\|_*$ is the standard Euclidean norm $\|\cdot\|$. Similarly to the variants introduced in Section 2.6.3, it generates a sequence of active sets $\{A_k\}$ with $s_k \in A_{k+1}$. At the step k the norm is minimized on the affine hull $\text{aff}(A_k)$ of the current active set A_k, that is

$$v_k = \arg\min_{y \in \text{aff}(A_k)} \|y\| \,. \tag{2.8.8}$$

The descent direction d_k is then defined as

$$d_k = v_k - x_k \,, \tag{2.8.9}$$

and the stepsize is given by a tailored line search that allows to remove some of the atoms in the set A_k (see, e.g. [157, 238]). Whenever x_{k+1} is in the relative interior of $\text{conv}(A_k)$, the FW vertex is added to the active set (that is, $s_k \in A_{k+1}$). Otherwise, at least one of the vertices not appearing in a convex representation of x_k is removed. This scheme converges linearly when applied to generic smooth strongly convex objectives (see, e.g., [157]).

In [122], a FW variant is proposed for minimum norm problems of the form

$$\min\{\|x\|_* : f(x) \leq 0,\, x \in K\} \tag{2.8.10}$$

with K a convex cone, f convex with L-Lipschitz gradient. In particular, the optimization domain is $\Omega = \{x \in \mathbb{R}^n : f(x) \leq 0\} \cap K$. The technique proposed in the article applies the standard FW method to the problems

$$\min\{f(x) : \|x\|_* \leq \delta_k,\, x \in K\} \,,$$

with $\{\delta_k\}$ an increasing sequence convergent to the optimal value $\bar{\delta}$ of the problem (2.8.10). Let $\Omega(\delta) = \{x \in \mathbb{R}^n : \|x\|_* \leq \delta\} \cap K$ for $\delta \geq 0$, and let

$$\mathrm{LM}(r) \in \operatorname*{arg\,min}_{x \in \Omega(1)} r^\intercal x \,,$$

so that by homogeneity for every k the linear minimization oracle on $C(\delta_k)$ is given by

$$\mathrm{LMO}_{\Omega(\delta_k)}(r) = \delta_k \mathrm{LM}(r) \,. \tag{2.8.11}$$

For every k, applying the FW method with suitable stopping conditions an approximate minimizer x_k of $f(x)$ over $\Omega(\delta_k)$ is generated, with an associated lower bound on the objective, an affine function in y:

$$f_k(y) := f(x_k) + \nabla f(x_k)^\intercal (y - x_k) \,. \tag{2.8.12}$$

Then the function

$$\ell_k(\delta) := \min_{y \in \Omega(\delta)} f_k(y) = f_k(\delta \mathrm{LM}(g_k)) \quad \text{with } g_k = \nabla f(x_k) \tag{2.8.13}$$

is decreasing and affine in δ and satisfies

$$\ell_k(\delta) = \min_{y \in \Omega(\delta)} f_k(y) \leq F(\delta) := \min_{y \in \Omega(\delta)} f(y) \,. \tag{2.8.14}$$

Therefore, for

$$\bar{\ell}_k(\delta) = \max_{i \in [1:k]} \ell_i(\delta) \leq F(\delta)$$

the quantity δ_{k+1} can be defined as $\min\{\delta \geq 0 : \bar{\ell}_k(\delta) \leq 0\}$, hence $F(\delta_{k+1}) \geq 0$. A complexity bound of $O(\frac{1}{\varepsilon} \ln(\frac{1}{\varepsilon}))$ was given to achieve precision ε applying this method, with $O(1/\varepsilon)$ iterations per subproblem and length of the sequence $\{\delta_k\}$ at most $O(\ln(1/\varepsilon))$ (see [122, Theorem 2] for details).

2.8.3 Variants for optimization over the trace norm ball

The FW method has found many applications for optimization problems over the trace norm ball. In this case, as explained in Example 2.4.4, linear optimization can be obtained by computing the top left and right singular vectors of the matrix $-\nabla f(X_k)$, an operation referred to as 1-SVD (see [10]) .

In the work [103], the FDFW is applied to the matrix completion problem (2.4.9), thus generating a sequence of matrices $\{X_k\}$ with $\|X_k\|_* \leq \delta$ for every k. The method can be implemented efficiently exploiting the fact that for X on the boundary of the nuclear norm ball, there is a simple expression for the face $\mathcal{F}(X)$. For $X \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(X) = k$ let $UDV^\intercal$ be the thin SVD of X, so that $D \in \mathbb{R}^{k\times k}$ is the diagonal matrix of non zero singolar values for X, with corresponding left and right singular vectors in the columns of $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$ respectively. If $\|X\|_* = \delta$ then the minimal face of the domain containing X is the set

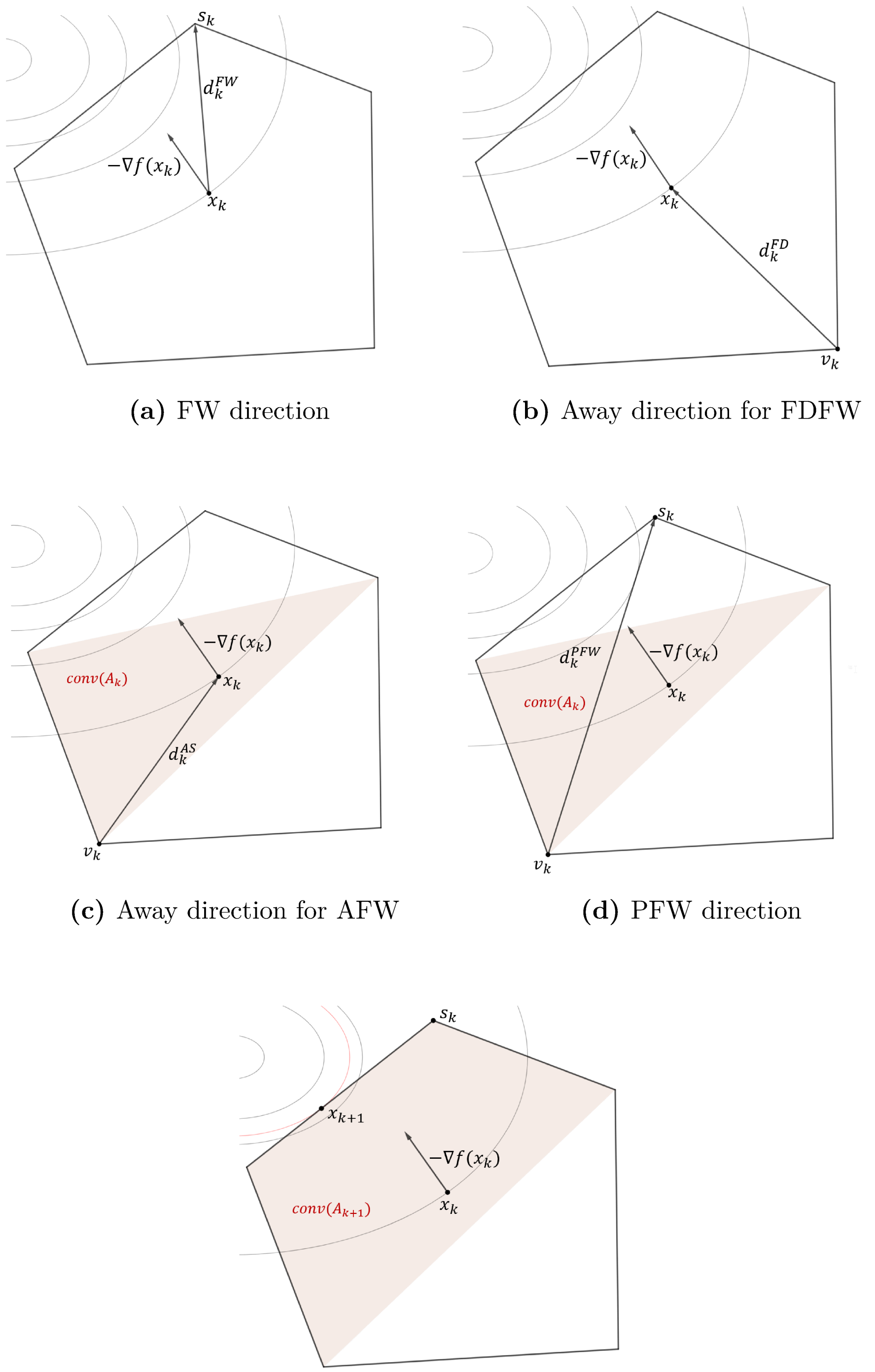
$$\mathcal{F}(X) = \{X \in \mathbb{R}^{m\times n} : X = UMV^\intercal \text{ for } M = M^\intercal \text{ psd with } \|M\|_* = \delta\}\,, \tag{2.8.15}$$

where psd stands for positive semidefinite.
It is not difficult to see that we have $\mathrm{rank}(X_k) \leq k + 1$ for every $k \in \mathbb{N}$, as well. Furthermore, the thin SVD of the current iterate X_k can be updated efficiently both after FW steps and after in face steps. The convergence rate of the FDFW in this setting is still $O(1/k)$.
In the recent work [232], an unbounded variant of the FW method is applied to solve a generalized version of the trace norm ball optimization problem:

$$\min_{X\in\mathbb{R}^{m\times n}} \{f(X) : \|\mathbb{P}XQ\|_* \leq \delta\} \tag{2.8.16}$$

with $\mathbb{P}, Q$ singular matrices. The main idea of the method is to decompose the domain in the sum $S+T$ between the kernel T of the linear function $\varphi_{\mathbb{P},Q}(X) = \mathbb{P}XQ$ and a bounded set $S \subset T^\perp$. Then gradient descent steps in the unbounded component T are alternated to FW steps in the bounded component S. The authors apply this approach to the generalized LASSO as well, using the AFW for the bounded component.

In [10], a variant of the classic FW using k-SVD (computing the top k left and right singular vectors for the SVD) is introduced, and it is proved that it converges linearly for strongly convex objectives when the solution has rank at most k. In [189], the FW step is combined with a proximal gradient step for a quadratic problem on the product of the nuclear norm ball with the ℓ_1 ball. Approaches using an equivalent formulation on the spectrahedron introduced in [137] are analyzed in [91, 106].



(a) FW direction

(b) Away direction for FDFW

(c) Away direction for AFW

(d) PFW direction

(e) EFW iteration

Figure 2.1: FW variants

Chapter 3

A unifying framework for the study of Frank-Wolfe variants

The study of Frank-Wolfe variants is often complicated by the presence of different kinds of "good" and "bad" steps. In this chapter, we aim to simplify the convergence analysis of specific variants by getting rid of such a distinction between steps, and to improve existing rates by ensuring a non-trivial bound at each iteration. In order to do this, we define the Short Step Chain (SSC) procedure, which skips gradient computations in consecutive short steps until proper conditions are satisfied. This algorithmic tool allows us to give a unified analysis and convergence rates in the general smooth non convex setting, as well as a linear convergence rate under a Kurdyka-Lojasiewicz (KL) property. While the KL setting has been widely studied for proximal gradient type methods, to our knowledge, it has never been analyzed before for the Frank-Wolfe variants considered in this chapter. An angle condition, ensuring that the directions selected by the methods have the steepest slope possible up to a constant, is used to carry out our analysis. We prove that such a condition is satisfied, when considering minimization problems over a polytope, by the away step Frank-Wolfe, the pairwise Frank-Wolfe, and the Frank-Wolfe method with in face directions. [1]

[1] This chapter is based on the article "Avoiding bad steps in Frank Wolfe variants" in *Computational Optimization and Applications, 2022* [209].

3.1 Motivation

In this chapter, we explain how to overcome an annoying issue affecting the analysis of some FW variants, and provide a unifying framework for the study of those methods. The issue we deal with is the presence of "bad iterations", i.e., iterations where we cannot show good progress. This happens when we are forced to take a short step along the search direction to guarantee feasibility of the iterate. The number of short steps typically needs to be upper bounded in the convergence analysis with "ad hoc" arguments (see, e.g., [103] and [157]). The main idea behind our method is to chain several short steps by skipping gradient updates until proper conditions are met.

3.1.1 Related work

FW variants. As seen in Chapter 2, the main drawback of the classic FW algorithm is its slow $O(1/k)$ convergence rate for convex objectives, which has motivated the study of variants with faster rates, starting at least with the work of Wolfe [237] (see [153] and [157] for recent references). For smooth strongly convex objectives, the convergence rates of many of these "improved directions" FW variants is linear on polytopes (see Section 2.7.1). Furthermore, in [148] it was proved that the convergence rate of an AFW variant is adaptive to Hölderian error bound conditions interpolating between the general convex case and the strongly convex one.

In addition to considering new directions, the works [58] and [59] propose strategies to skip the LMO computation from time to time by caching linear minimizers, while the recent work [153] for optimization on polytopes applies recursively a FW variant to smaller polytopes. However, to our knowledge, no strategy to avoid short steps has been discussed in these previous works.

A different approach, adopted in the general smooth convex setting, is to use FW variants to approximate projections. In particular, the conditional gradient sliding method uses the FW method to approximate projections on the feasible set within a projected gradient scheme (see, e.g., [124] and [161]). Another approach introduced in [79] for smooth convex objectives implicitly uses the Non Negative Matching Pursuit (NNMP) algorithm to compute an approximate projection of the negative gradient on the tangent cone. To our knowledge, however, conditional gradient sliding approaches always lead to a sublinear $O(1/\varepsilon)$ LMO complexity, and the approach in [79] does not lead to any improvement on the $O(1/\varepsilon)$ worst case

gradient complexity of the classic FW.
Outside the projection free setting, in [187] a procedure making multiple steps without updating the gradient (in a fashion similar to our SSC) is defined.
In the non convex setting, for the classic FW algorithm a convergence rate of $O(1/\sqrt{k})$ was proved in [156] and then extended to other variants in [47] and [205].

KL property. The KL property (see, e.g., [12], [36] and [37]) has been extensively applied to compute the convergence rates of proximal subgradient type methods (see, e.g., [12], [13], [38], [233] and [242]). Furthermore, for convex objectives, it has been proved that Hölderian error bound conditions are a particular case of this property [38]. However, we are not aware of previous applications to the Frank-Wolfe variants under study in this chapter.

Angle condition. The analysis of unconstrained descent methods often relies on some version of an angle condition, imposing an upper bound on the angle between the negative gradient and the descent direction selected by the method (see, e.g., [2], [114] and [249]). However, due to the presence of short steps and full FW steps, these analyses do not extend to our setting in a straightforward way.
In Section 3, we present an angle condition for optimization over a convex set. While to our knowledge this extension is novel for first order optimization methods, analogous conditions can be found in the context of direct search methods for linearly constrained derivative free optimization (see, e.g., [152] and [168]), imposed on the smallest angle between the negative gradient and a search direction. Finally, we remark that a variant of our condition was somehow used, but not stated explicitly, in [27] and [157] within the context of smooth strongly convex optimization over polytopes.

3.1.2 Contributions

Our main contributions are twofold:

- We formulate an angle condition for projection free methods, and prove that it leads to linear convergence in the number of "good steps" for non convex objectives satisfying a KL inequality. We show that this condition applies to the AFW, the PFW and the FDFW on polytopes. First, we give linear rates for good steps in Proposition 3.3.6. Then, we give global asymptotical rates under the assumption that the number of bad steps between two good steps is bounded in Proposition 3.3.7. We apply this result to FW variants in Corollary 3.3.8.

- We define the SSC procedure, which can be applied to all the FW variants listed in the first point, and show that it gets improvements on known rates (see Table 1 in Section 3.4). In particular, we prove that it leads to global linear convergence rates with no bad steps (see Lemma 3.4.11 and Corollary 3.4.15) under a global KL inequality and the angle condition. We then prove that we have local linear convergence rates and asymptotical linear convergence rates under a local KL property as well (see Theorem 3.4.13 and Corollary 3.4.14). This, to our knowledge, is the first (bad step free) linear convergence rate for FW variants under the KL inequality. In the general smooth non convex case, we further prove, under the angle condition, a $O(1/\sqrt{k})$ convergence rate with respect to a specific measure of non-stationarity for the iterates, that is the projection of the negative gradient on the convex cone of feasible directions (see Theorem 3.4.8, Corollary 3.4.9 and Remark 3.4.10).

While here we apply our framework only to the AFW, the PFW, and the FDFW on polytopes, we remark that our results hold for projection free methods on generic convex sets. In an extended version of this chapter [208] we show applications on convex sets with smooth boundary for FW variants and methods using orthographic retractions (see also [4], [22], [167] and references therein).
The reasons why eliminating bad steps truly makes a difference in our context are the following:

- it rules out impractical convergence rates due to a large number of bad steps. An interesting example is given by the rate guarantee reported in [157] for the pairwise Frank-Wolfe (PFW) variant on the $N-1$ dimensional simplex. This guarantee is indeed more loose than for the other variants, because there is no satisfactory bound on the number of such problematic steps (there is a best known bound of $3N!$ bad steps for each good step);

- it eliminates the dependence of the convergence rates on the support of the starting point (see, e.g., [139] and [153]). This dependence can significantly affect the performance of FW variants on smooth non convex optimization problems [84].

Finally, we mention that bad steps lead to a slow active set identification for the AFW. This will be discussed more in depth in Chapter 4.

The structure of the chapter is as follows. In Section 2, we define some notation and state some preliminary results from convex analysis. In Section 3, we introduce the angle condition for first-order projection free methods, show examples of FW variants satisfying the condition and prove linear convergence in the number of good steps. We define the SSC procedure in Section 4, where we also state the main convergence results. Preliminary numerical results are reported in Section 3.6.

3.2 Tangent cones and the KL condition

We consider the following constrained optimization problem:

$$\min \{f(x) \mid x \in \Omega\} \ . \tag{3.2.1}$$

In the rest of the chapter Ω is a compact and convex set and $f \in C^1(\Omega)$ with L-Lipschitz gradient. We define D as the diameter of Ω, and for $a, b \in \mathbb{R} \cup \{\pm\infty\}$ we denote as $[a < f(x) < b]$ the set $\{x \in \Omega \mid f(x) \in (a, b)\}$, with analogous definitions for non strict inequalities. We define $B_R(C)$ as the neighborhood $\{x \in \mathbb{R}^n \mid \text{dist}(C, x) < R\}$ of C of radius R and in particular $B_R(x)$ as the open euclidean ball of radius R and center x, $\bar{B}_R(x)$ as its closure. When C is closed and convex we define as $\pi(C, \cdot)$ the projection on C. If C is a cone then we denote with C^* its polar.

We now state some elementary properties related to the tangent and the normal cones, where for $x \in \Omega$ we denote with $T_\Omega(x)$ and $N_\Omega(x)$ the tangent and the normal cone to Ω in x respectively. The next proposition (from [211], Theorem 6.9) characterizes these cones for closed convex subsets of $\mathbb{R}^n$.

Proposition 3.2.1. *Let Ω be a closed convex set. For every point $x \in \Omega$ we have*

$$\begin{aligned} T_\Omega(x) &= \text{cl}\{w \mid \exists \lambda > 0 \text{ with } x + \lambda w \in \Omega\} \,, \\ \text{int}(T_\Omega(x)) &= \{w \mid \exists \lambda > 0 \text{ with } x + \lambda w \in \ \text{int}(\Omega)\} \,, \\ N_\Omega(x) &= T_\Omega(x)^* = \{v \in \mathbb{R}^n \mid (v, y - x) \le 0 \ \forall \ y \in \Omega\} \,. \end{aligned}$$

We have the following formula connecting the supremum of a linear function "slope" along feasible directions to the tangent and the normal cone:

Proposition 3.2.2. *If Ω is a closed convex subset of $\mathbb{R}^n$, $x \in \Omega$ then for every $g \in \mathbb{R}^n$*

$$\max\left\{0, \sup_{h \in \Omega \setminus \{x\}} \left(g, \frac{h - x}{\|h - x\|}\right)\right\} = \text{dist}(N_\Omega(x), g) = \|\pi(T_\Omega(x), g)\| \,.$$

Before giving the proof, we recall a useful well known result:

Proposition 3.2.3. *Let C be a closed convex cone. For every $y \in \mathbb{R}^n$*

$$\operatorname{dist}(C^*, y) = \sup_{c \in C} \hat{c}^\top y \,.$$

As stated in [60] this is an immediate consequence of the Moreau-Yosida decomposition:

$$y = \pi(C, y) + \pi(C^*, y) \,.$$

Proof of Proposition 3.2.2. First, by continuity of the scalar product we have

$$\sup_{h \in \Omega / \{x\}} \left(g, \frac{h - x}{\|h - x\|} \right) = \sup_{h \in T_\Omega(x) \setminus \{0\}} (g, \hat{h}) \,. \tag{3.2.2}$$

Since $N_\Omega(x) = T_\Omega(x)^*$ the first equality is exactly the one of Proposition 3.2.3 if $g \notin N_\Omega(x)$, and it is trivial since both terms are clearly 0 if $g \in N_\Omega(x)$.
It remains to prove

$$\operatorname{dist}(N_\Omega(x), g) = \|\pi(T_\Omega(x), g)\| \,, \tag{3.2.3}$$

which is true by the Moreau - Yosida decomposition. □

On polytopes, a geometric interpretation of Proposition 3.2.2 is that the smallest angle between g and a descent direction d feasible in x is achieved for $d = \pi(T_\Omega(x), g)$. In the rest of the chapter to simplify notations we often use $\pi_x(g)$ as a shorthand for $\|\pi(T_\Omega(x), g)\|$. Then, by Proposition 3.2.2, first order stationarity conditions in x for the gradient $-g$ become equivalent to $\pi_x(g) = 0$.
In the computation of the convergence rates, we often make the following assumption.

Assumption 3.1. Given a stationary point $x^* \in \mathbf{\Omega}$, there exists $\eta, \delta > 0$ such that for every $x \in [f(x^*) < f < f(x^*) + \eta] \cap B_\delta(x^*)$

$$\pi_x(-\nabla f(x)) \geq \sqrt{2\mu}(f(x) - f(x^*))^{\frac{1}{2}} \,. \tag{3.2.4}$$

We refer the reader to the extended version [208] of this chapter for a study of convergence rates under a more general inequality, interpolating between (3.2.4) and the generic non convex case. Let now i_Ω be the indicator function of $\mathbf{\Omega}$ so that $i_\Omega(x) = 0$ in $\mathbf{\Omega}$ and $i_\Omega(x) = +\infty$ otherwise. It can easily be seen that (3.2.4) is a special case of the KL inequality (see, e.g., [12], [13] and [38]) with exponent $\frac{1}{2}$

$$\operatorname{dist}(0, \partial f_\Omega(x)) \geq \sqrt{2\mu}(f_\Omega(x) - f_\Omega(x^*))^{\frac{1}{2}} \tag{3.2.5}$$

for $f_\Omega = f + i_\Omega$, using that

$$\pi_x(-\nabla f(x)) = \mathrm{dist}(-\nabla f(x), N_\Omega(x)) = \mathrm{dist}(0, \partial(f + i_\Omega)(x)), \tag{3.2.6}$$

with the last equality following by Proposition 3.2.2. For convex objectives, condition (3.2.4) is therefore implied by the Holderian error bound $f(x) - f(x^*) \geq \rho\, \mathrm{dist}(x, \mathcal{X}^*)^2$, for $\mathcal{X}^*$ set of solutions of Problem (3.2.1) (see [38, Corollary 6]), which in turn is implied by $\mu-$ strong convexity (see, e.g., [146]). For non convex objectives, Assumption 3.1 is implied by the Luo Tseng error bound [180] under some mild separability conditions for stationary points (see [170, Theorem 4.1]). This error bound is known to hold in a variety of convex and non convex settings (see Section 3.5 and references in [170]).

We now show that under suitable assumptions our KL condition is implied by the classic Polyak-Lojasiewicz (PL) inequality from [176] and [203]. We first recall the PL property as it is used in [146]:

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*). \tag{3.2.7}$$

with f^* optimal value of f with non empty solution set $\mathcal{X}^*$.

Proposition 3.2.4. *If f is convex, the optimal solution set $\mathcal{X}^*$ of f is contained in Ω and* (3.2.7) *holds, then* (3.2.4) *holds for every $x \in \Omega$.*

Proof. By [146, Theorem 2] the PL property is equivalent, for convex objectives, to the unconstrained quadratic growth condition:

$$f(x) - f^* \geq \frac{\mu}{2}\, \mathrm{dist}(x, \mathcal{X}^*)^2 \tag{3.2.8}$$

In turn, given that by the assumption $\mathcal{X}^* \subset \Omega$ the set $\mathcal{X}^*$ is the solution set for f_Ω as well, (3.2.8) implies the global non smooth Holderian error bound condition from [38] with $\varphi(t) = \sqrt{\frac{2t}{\mu}}$, and by [38, Corollary 6] this is equivalent to the KL property (3.2.4) holding globally on Ω. □

Remark 3.2.5. We remark that without the assumption $\mathcal{X}^* \subset \Omega$ the implication is no longer true even for convex objectives, a counter example being Ω equal to the unitary ball and $f((x^{(1)}, ..., x^{(n)})) = (x^{(1)} - 1)^2$. At the same time, the KL property we used does not imply the PL property in general, since the latter only deals with unconstrained minima.

3.3 An angle condition

Let $\mathcal{A}$ be a first-order optimization method defined for smooth functions on a closed subset Ω of $\mathbb{R}^n$. We assume that given first-order information $(x_k, \nabla f(x_k))$ the method always selects x_{k+1} along a feasible descent direction, so that for $(x, g) \in \Omega \times \mathbb{R}^n$ we can define

$$\mathcal{A}(x, g) \subset T_\Omega(x) \cap \{y \in \mathbb{R}^n \mid g^\top y > 0\} \cup \{0\}$$

as the possible descent directions selected by $\mathcal{A}$ when $x = x_k$, $g = -\nabla f(x_k)$ for some k (see Algorithm 1). When x is first-order stationary, we set $\mathcal{A}(x, g) = \{0\}$, otherwise we always assume $0 \notin \mathcal{A}(x, g) \neq \emptyset$.

We want to formulate an angle condition for the descent directions selected by $\mathcal{A}$, with respect to the infimum of the angles achieved with feasible descent directions. In order to do that, we define the directional slope lower bound as

$$\mathrm{DSB}_{\mathcal{A}}(\Omega, x, g) = \inf_{d \in \mathcal{A}(x,g)} \frac{g^\top d}{\pi_x(g)\|d\|}$$

if $0 \notin \mathcal{A}(x, g)$. Otherwise x is stationary for $-g$, $\pi_x(g) = 0$ and we set $\mathrm{DSB}_{\mathcal{A}}(\Omega, x, g) = 1$. Then with this definition it immediately follows $\mathrm{DSB}_{\mathcal{A}}(\Omega, x, g) \leq 1$ by Proposition 3.2.2. Notice also that when $x \in \mathrm{int}(\Omega)$ then $\mathrm{DSB}_{\mathcal{A}}(\Omega, x, g)$ is simply a lower bound on $\cos(\theta_{g,d})$ with θ the angle between g and a descent direction d:

$$\mathrm{DSB}_{\mathcal{A}}(\Omega, x, g) = \inf_{d \in \mathcal{A}(x,g)} \frac{g^\top d}{\|g\|\|d\|} \tag{3.3.1}$$

and thus imposing $\mathrm{DSB}_{\mathcal{A}}(\Omega, x, g) \geq \tau$ we retrieve the angle condition [2, equation (20)]. We remark that the RHS of (3.3.1) defining the unconstrained angle condition is also considered in the constrained setting in [79] (referred to as alignment condition), as a tool to evaluate potential descent directions. However, without $\pi_x(g)$ in the denominator no uniform lower bound can be given for the RHS, and therefore no worst case linear convergence rate (the rate given in [79, Corollary 3.6] is in fact $O(1/k)$).

Given a subset P of Ω we can finally define the slope lower bound

$$\mathrm{SB}_{\mathcal{A}}(\Omega, P) = \inf_{\substack{g \in \mathbb{R}^n \\ x \in P}} \mathrm{DSB}_{\mathcal{A}}(\Omega, x, g) = \inf_{\substack{g : \pi_x(g) \neq 0 \\ x \in P}} \mathrm{DSB}_{\mathcal{A}}(\Omega, x, g)\,.$$

For simplicity if $P = \Omega$ we write $\mathrm{SB}_{\mathcal{A}}(\Omega)$ instead of $\mathrm{SB}_{\mathcal{A}}(\Omega, \Omega)$.

We now show a few examples of Frank-Wolfe variants satisfying the following *angle condition*

$$\mathrm{SB}_{\mathcal{A}}(\Omega) = \tau > 0, \tag{3.3.2}$$

i.e. cases where the slope lower bound is strictly greater than 0.

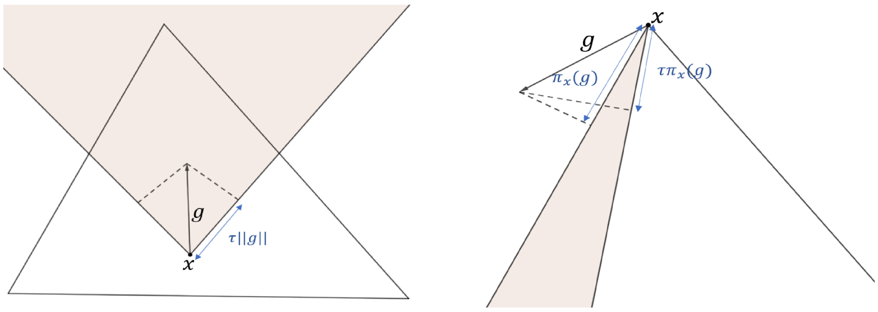


Figure 3.1: In red, cone of feasible descent directions satisfying the angle condition in x for the negative gradient g when x is in the interior (left) and when x is on the boundary (right).

3.3.1 Frank-Wolfe variants over polytopes and the angle condition

We now consider the AFW, PFW and FDFW and show that the angle condition is satisfied when Ω is a polytope. The AFW and PFW depend on a set of "elementary atoms" A such that $\Omega = \mathrm{conv}(A)$. Given A, for a base point $x \in \Omega$ we can define

$$S_x = \{S \subset A \mid x \text{ is a proper convex combination of all the elements in } S\}\,,$$

the family of possible active sets for x. In the rest of the chapter A is always clear from the context and for simplicity we write PFW, AFW instead of PFW_A, AFW_A. For $x \in \Omega$, $S \in S_x$, d^{PFW} is a PFW direction with respect to the active set S and gradient $-g$ iff

$$d^{\mathrm{PFW}} = s - q \text{ with } s \in \arg\max_{s\in\Omega} s^\top g \text{ and } q \in \arg\min_{q\in S} q^\top g\,. \tag{3.3.3}$$

Similarly, given $x \in \Omega$, $S \in S_x$, d^{AFW} is an AFW direction with respect to the active set S and gradient $-g$ iff

$$d^{\mathrm{AFW}} \in \arg\max\{g^\top d \mid d \in \{d^{\mathrm{FW}}, d^{AS}\}\}\,, \tag{3.3.4}$$

where d^{FW} is a classic Frank-Wolfe direction

$$d^{\mathrm{FW}} = s - x \text{ with } s \in \operatorname*{arg\,max}_{s \in \Omega} s^\top g \,, \tag{3.3.5}$$

and d^{AS} is the away direction

$$d^{\mathrm{AS}} = x - q \text{ with } q \in \operatorname*{arg\,min}_{q \in S} q^\top g \,. \tag{3.3.6}$$

The FDFW from [103], [116] (sometimes referred to as Decomposition invariant Conditional Gradient (DiCG) when applied to polytopes [110], [24]) relies only on the current point x and the current gradient $-g$ to choose a descent direction and, unlike the AFW and the PFW, does not need to keep track of the active set.

The in face direction is defined as

$$d^F = x_k - x_F \text{ with } x_F \in \arg\min\{g^\top y \mid y \in \mathcal{F}(x)\}$$

for $\mathcal{F}(x)$ the minimal face of Ω containing x. The selection criterion is then analogous to the one used by the AFW:

$$d^{\mathrm{FD}} \in \operatorname{argmax}\{g^\top d \mid d \in \{d^F, d^{\mathrm{FW}}\}\} \,. \tag{3.3.7}$$

We write $\mathrm{SB}_{\mathrm{FD}}, \mathrm{DSB}_{\mathrm{FD}}$ instead of $\mathrm{SB}_{\mathrm{FDFW}}, \mathrm{DSB}_{\mathrm{FDFW}}$ in the rest of the chapter. When Ω is a polytope and $|A| < \infty$, the angle condition holds for the directions and the related FW variants we introduced. Before stating a lower bound for $\mathrm{SB}_{\mathcal{A}}(\Omega)$ in this setting we need to recall the pyramidal width constant $\mathrm{PWidth}(A)$ introduced in [157]. We refer the reader to [206] and references therein for a discussion of various properties of this and related parameters.

We use here a characterization of $\mathrm{PWidth}(A)$ proved in [200]:

$$\mathrm{PWidth}(A) = \min_{\mathcal{F} \in \mathrm{pfaces}(\Omega)} \mathrm{dist}(\mathcal{F}, \mathrm{conv}(A \setminus \mathcal{F})) \,, \tag{3.3.8}$$

with $\mathrm{pfaces}(\Omega)$ the set of proper faces of Ω. We now introduce one key property of $\mathrm{PWidth}(A)$ which relates it to the angle along the PFW direction. While we give a self contained proof of the lemma relying only on (3.3.8), we remark that the lemma can also be proved using [157, Theorem 3].

We first need this preliminary lemma relating maximal stepsize length with PWidth. For $y \in \Omega$, $d \in \mathbb{R}^n$, let $\alpha^{\max}(y, d)$ the maximal feasible stepsize from y in the direction d.

Lemma 3.3.1. *Let x be a proper convex combination of atoms in $A' \subset A$, and $d \neq 0$ feasible direction in x. Then, for some $y \in \mathrm{conv}(A')$, we have*

$$\hat{\alpha}^{\max}(y, d) \geq \frac{\mathrm{PWidth}(A)}{\|d\|} \,. \tag{3.3.9}$$

Proof. Let $y \in \arg\max_{z \in \mathrm{conv}(A')} \hat{\alpha}^{\max}(z, d)$, and let $A'' \subset A'$ be such that y is a proper convex combination of elements in A''. Furthermore, let $\mathcal{F}_y$ be the minimal face containing the maximal feasible step point $\bar{y} := y + \hat{\alpha}^{\max}(y, d)$. We claim that $\mathcal{F}_y \cap A'' = \emptyset$. In fact, for $p \in A'' \cap \mathcal{F}_y$ we can consider an homothety of center p and factor $1 + \epsilon$ mapping y in $y_\epsilon \in \mathrm{conv}(A'')$ and $\bar{y}$ in $\bar{y}_\epsilon \in \mathcal{F}_y$ with

$$\bar{y}_\epsilon = y_\epsilon + (1 + \epsilon)\hat{\alpha}^{\max}(y, d)d \,.$$

But then we would have $\hat{\alpha}(\bar{y}_\epsilon, d) \geq (1 + \epsilon)\hat{\alpha}(\bar{y}, d)$, in contradiction with the maximality of $\hat{\alpha}(\bar{y}, d)$. Therefore

$$\hat{\alpha}^{\max}(y, d) \geq \mathrm{dist}(A'', \mathcal{F}_y) \geq \min_{\mathcal{F} \in \mathrm{pfaces}(\Omega)} \mathrm{dist}(\mathcal{F}, \mathrm{conv}(A \setminus \mathcal{F})) = \mathrm{PWidth}(A) \,, \tag{3.3.10}$$

where we used $A'' \cap \mathcal{F} = \emptyset$ in the second inequality, and [200, Theorem 2] in the equality. □

We can now prove the main Lemma.

Lemma 3.3.2. *We have the following lower bound*

$$\frac{g^\top d^{\mathrm{PFW}}}{\|\pi(T_\Omega(x), g)\|} \geq \mathrm{PWidth}(A) \,.$$

Proof. We use s, q and S as in (3.3.3). For z in Ω and d feasible direction in z we define as $\hat{\alpha}^{\max}(z, d)$ the maximal feasible stepsize in the direction d. Let $p = \pi(T_\Omega(x), g)$, and let y be a maximizer of $\hat{\alpha}^{\max}(y, p)$ for $y \in S$. We have

$$\begin{aligned} & g^\top d^{\mathrm{PFW}} = g^\top((s - y) + (y - q)) \geq g^\top(s - y) \geq g^\top((y + \hat{\alpha}^{\max}(y, p)p) - y) \\ \geq & \frac{\mathrm{PWidth}(A)}{\|p\|} g^\top p = \mathrm{PWidth}(A)\|p\| \,, \end{aligned} \tag{3.3.11}$$

where we used Lemma 3.3.1 in the third inequality, and $g^\top p = \|p\|^2$ as it follows by the Moreau-Yosida decomposition in the last equality. □

In order to define an angle condition for the FDFW, we use the following upper bound on PWidth(A), independent from the particular set A chosen to represent $\mathbf{\Omega}$:

$$\mathrm{PFWidth}(\Omega) = \min_{\substack{\mathcal{F}_1,\mathcal{F}_2 \in \mathrm{pfaces}(\Omega) \\ \mathcal{F}_1 \cap \mathcal{F}_2 = \emptyset}} \mathrm{dist}(\mathcal{F}_1, \mathcal{F}_2)\,. \tag{3.3.12}$$

Proposition 3.3.3. $\mathrm{SB}_{\mathrm{PFW}}(\Omega) \geq \tau_p := \frac{\mathrm{PWidth}(A)}{D}$, $\mathrm{SB}_{\mathrm{AFW}}(\Omega) \geq \frac{\tau_p}{2}$, $\mathrm{SB}_{\mathrm{FD}}(\Omega) \geq \frac{\tau_v}{2} := \frac{\mathrm{PFWidth}(\Omega)}{2D}$.

Proof. Let g be such that $\pi_x(g) \neq 0$. We have

$$\begin{aligned} \mathrm{DSB}_{\mathrm{PFW}}(\Omega, x, g) &= \inf_{d^{\mathrm{PFW}} \in \mathrm{PFW}(x,g)} \frac{g^\top d^{\mathrm{PFW}}}{\|d^{\mathrm{PFW}}\| \|\pi(T_\Omega(x), g)\|} \\ &\geq \frac{g^\top d^{\mathrm{PFW}}}{D\|\pi(T_\Omega(x), g)\|} \geq \frac{\mathrm{PWidth}(A)}{D}\,, \end{aligned}$$

where we used Lemma 3.3.2 in the last inequality.
Hence $\mathrm{SB}_{\mathrm{PFW}}(\Omega) \geq \frac{\mathrm{PWidth}(A)}{D}$ follows by taking the inf on the LHS for $x \in \Omega$ and g such that $\pi_x(g) \neq 0$ in (3.3.1). The inequality $\mathrm{SB}_{\mathrm{AFW}}(\Omega) \geq \frac{\mathrm{PWidth}(A)}{2D}$ is a corollary since

$$g^\top d^{\mathrm{AFW}} \geq \frac{1}{2} g^\top d^{\mathrm{PFW}}\,,$$

as it follows immediately from the definitions (see also [157, equation (6)]).
The angle condition for the FDFW can be proved analogously to the angle condition for the AFW, where in Lemma 3.3.1 the RHS can be improved with PFWidth(Ω) instead of PWidth(A) using that the active set A' can be taken as the set of vertices of a face. □

Remark 3.3.4. Results analogous to the ones in Proposition 3.3.3 can be proven relatively to the vertex facial distance vf(Ω) from [27]. More precisely, assuming $A = V(\Omega)$, for $V(\Omega)$ set of vertices of Ω, and that the AFW and the PFW keep active sets of size at most $\bar{s}$, we have $\mathrm{SB}_{\mathrm{PFW}}(\Omega) \geq \frac{\mathrm{vf}(\Omega)}{\bar{s}D}$, $\mathrm{SB}_{\mathrm{AFW}}(\Omega) \geq \frac{\mathrm{vf}(\Omega)}{2\bar{s}D}$ as a consequence of [27, Lemma 3.1]. Furthermore, for the FDFW we have $\mathrm{SB}_{\mathrm{FD}}(\Omega, \Omega_{\bar{s}}) \geq \frac{\mathrm{vf}(\Omega)}{2\bar{s}D}$, with $x \in \Omega_{\bar{s}} \subset \Omega$ iff there exists $S \in S_x$ such that $|S| \leq \bar{s}$.

3.3.2 Linear convergence for good steps under the angle condition

Consider now a method following the scheme described by Algorithm 1, and with Lipschitz constant dependent stepsize as defined by (2.5.4):

$$\alpha_k = \min\left(\bar{\alpha}_k, \alpha_k^{\max}\right)\,, \tag{3.3.13}$$

with

$$\bar{\alpha}_k = \frac{-\nabla f(x_k)^\top d_k}{L\|d_k\|^2}. \tag{3.3.14}$$

The following lemma shows that at every iteration a sufficient decrease condition is satisfied, independently from the method $\mathcal{A}$, when using stepsize (3.3.14).

Lemma 3.3.5. *If $\alpha_k \leq \bar{\alpha}_k$, thus in particular for the stepsize* (3.3.13)*, we have:*

$$f(x_k) - f(x_{k+1}) \geq \frac{L}{2}\|x_k - x_{k+1}\|^2 . \tag{3.3.15}$$

Proof. By the standard descent lemma [31, Proposition 6.1.2],

$$f(x_{k+1}) = f(x_k + \alpha_k d_k) \leq f(x_k) + \alpha_k \nabla f(x_k)^\top d_k + \alpha_k^2 \frac{L}{2}\|d_k\|^2 , \tag{3.3.16}$$

and in particular

$$f(x_k) - f(x_{k+1}) \geq -\alpha_k \nabla f(x_k)^\top d_k - \alpha_k^2 \frac{L}{2}\|d_k\|^2 \geq \frac{L}{2}\alpha_k^2\|d_k\|^2 = \frac{L}{2}\|x_{k+1} - x_k\|^2 , \tag{3.3.17}$$

where we used $\alpha_k \leq \bar{\alpha}_k$ in the last inequality. This proves (3.3.15). □

Assume now that the method $\mathcal{A}$ satisfies the angle condition (3.3.2). In the following proposition, we prove a general linear convergence rate in the number of *good steps*, (recall from Chapter 2 that these are the steps satisfying $\alpha_k = \bar{\alpha}_k$ or full FW steps), under the assumption that the method $\mathcal{A}$ satisfies the angle condition (3.3.2), and that the KL inequality (3.2.4) holds for the objective function f in Problem (3.2.1).

Proposition 3.3.6. *Let us assume that $\mathcal{A}$ satisfies the angle condition* (3.3.2)*, and the objective function f in Problem* (3.2.1) *satisfies condition* (3.2.4) *in x_k and x_{k+1}.*

- *If $\alpha_k = \bar{\alpha}_k$ then*

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\mu}{L}\tau^2\right)(f(x_k) - f(x^*)) . \tag{3.3.18}$$

- *If the step k is a full FW step then*

$$f(x_{k+1}) - f(x^*) \leq \left(1 + \frac{\mu}{L}\right)^{-1}(f(x_k) - f(x^*)) . \tag{3.3.19}$$

Proof. Let $p_k = \|\pi(T_\Omega(x_{k+1}), -\nabla f(x_{k+1}))\|$ and $\tilde{p}_k = \|\pi(T_\Omega(x_{k+1}), -\nabla f(x_k))\|$. We have

$$\begin{aligned} |p_k - \tilde{p}_k| &= |\|\pi(T_\Omega(x_{k+1}), -\nabla f(x_{k+1}))\| - \|\pi(T_\Omega(x_{k+1}), -\nabla f(x_k))\|| \\ &\leq \| - \nabla f(x_{k+1}) + \nabla f(x_k)\| \leq L\|x_{k+1} - x_k\| \,, \end{aligned} \tag{3.3.20}$$

where we used the 1-Lipschitzianity of projections in the first inequality.

If $\alpha_k = \bar{\alpha}_k$ then

$$\begin{aligned} f(x_{k+1}) =& f(x_k + \bar{\alpha}_k d_k) \leq f(x_k) - \frac{1}{2L}\left(\frac{\nabla f(x_k)^\top d_k}{\|d_k\|}\right)^2 \leq f(x_k) - \frac{\tau^2}{2L}p_{k-1}^2 \\ \leq& f(x_k) - \frac{\mu\tau^2}{L}(f(x_k) - f(x^*)) \,, \end{aligned} \tag{3.3.21}$$

where we used (3.3.16) in the first inequality, $\mathrm{SB}^f_{\mathcal{A}}(\Omega) = \tau$ in the second one, and condition (3.2.4) in the third one.

If the step k is a full FW step then $\tilde{p}_k = 0$ because $x_{k+1} \in \arg\min_{y\in\Omega} \nabla f(x_k)^\top y \Leftrightarrow -\nabla f(x_k) \in N_\Omega(x_{k+1}) \Leftrightarrow \|\pi(T_\Omega(x_{k+1}), -\nabla f(x_k))\| = 0$, where the last equivalence is true by Proposition 3.2.2. Then

$$f(x_{k+1}) - f(x^*) \leq \frac{p_k^2}{2\mu} \leq \frac{(\tilde{p}_k + L\|x_{k+1} - x_k\|)^2}{2\mu} = \frac{L^2}{2\mu}\|x_{k+1} - x_k\|^2 \leq \frac{L}{\mu}(f(x_k) - f(x_{k+1})) \,, \tag{3.3.22}$$

where we used (3.2.4) in the first inequality, (3.3.20) in the second, $\tilde{p}_k = 0$ and (3.3.17) in the last inequality. Then (3.3.17) and (3.3.19) follow by rearranging (3.3.21) and (3.3.22) respectively. □

We finally report an asymptotic rate under the additional assumption that bad steps between two good steps are limited.

Proposition 3.3.7. *Assume that the number of bad steps between two good steps is limited and that $\mathcal{A}$ satisfies the angle condition* (3.3.2). *Then:*

- *$\{x_k\}$ converges to the set of stationary points, and $f(x_k)$ is decreasing and convergent to $f^* \in \mathbb{R}$;*
- *if Assumption 3.1 holds for every stationary point in the level set $[f(x) = f^*]$, we have the asymptotic convergence rate:*

$$f(x_k) - f(x^*) \leq M\bar{q}^{\gamma_g(k)} \,, \tag{3.3.23}$$

for some $M > 0$, $\gamma_g(k)$ number of good steps among the first k steps and

$$\bar{q} = \max\left(\left(1 + \frac{\mu}{L}\right)^{-1}, \left(1 - \frac{\mu}{L}\tau^2\right)\right). \tag{3.3.24}$$

Proof. Let $k(j)$ be the subsequence of iterates associated to good steps, so that by assumption $k(j+1) - k(j)$ is bounded, and define $\tilde{k}(j) = k(j) - 1$ if $\alpha_{k(j)} = \bar{\alpha}_{k(j)}$, $\tilde{k}(j) = k(j)$ otherwise. Notice that $\tilde{k}(j+1) - \tilde{k}(j)$ is also bounded. By (3.3.17) we have that $\{f(x_k)\}$ is decreasing and thus convergent to $f^* \in \mathbb{R}$, and also that $\|x_k - x_{k+1}\| \to 0$. With the notation used in Proposition 3.3.6 we now claim $p_{\tilde{k}(j)} \to 0$. In fact if $\alpha_{k(j)} = \bar{\alpha}_{k(j)}$ then

$$p_{\tilde{k}(j)}^2 = p_{k-1}^2 \le \frac{2L}{\tau^2}(f(x_k) - f(x_{k+1})) \to 0\,, \tag{3.3.25}$$

where we used (3.3.21) in the inequality, and if $k(j)$ is a full FW step then

$$p_{\tilde{k}(j)} \le p_{k(j)} \le \tilde{p}_{k(j)} + L\|x_{k(j)+1} - x_{k(j)}\| = L\|x_{k(j)+1} - x_{k(j)}\| \to 0\,, \tag{3.3.26}$$

where we used (3.3.20) in the first inequality and $\tilde{p}_{k(j)} = 0$ in the equality.
We therefore have $p_{\tilde{k}(j)} \to 0$. Equivalently, thanks to (3.2.6) we have
$\text{dist}(0, \partial f_\Omega(x_{\tilde{k}(j)})) \to 0$, so if x^* is a limit point of $x_{\tilde{k}(j)}$ by lower semicontinuity of the subdifferential we must have $0 \in \partial f_\Omega(x^*)$, i.e., x^* is stationary. In particular, by compactness $\{x_{\tilde{k}(j)}\}$ must converge to the set of stationary points. By the boundedness of $\|x_{k+1} - x_k\|$ and $\tilde{k}(j+1) - \tilde{k}(j)$ we also have that the set of limit points of $\{x_k\}$ coincides with the set of limit points of $\{x_{\tilde{k}(j)}\}$, and in particular it is a subset of stationary points contained in $[f(x) = f^*]$.
Let $\Omega^* \subset [f(x) = f^*]$ be the set of limit points of $\{x_k\}$. By compactness (see [39, Lemma 6]), we have that for some fixed $\varepsilon, \eta > 0$, the KL property holds for every $x^* \in \Omega^*$ with parameters ε and η. Then for k large enough $x_k \in B_\delta(x^*) \cap [f(x^*) < f < f(x^*) + \eta]$ for some $x^* \in \Omega$, and the asymptotic rates follow by Proposition 3.3.6. □

For the three FW variants described before we can now give an asymptotic linear convergence rate in the number of good steps. We refer the reader to Table 1 for bounds on this number.

Corollary 3.3.8. *Let us assume that the objective function f satisfies Assumption 3.1 for every stationary point in the level set $[f(x) = f^*]$ and $\Omega = \text{conv}(A)$ with*

$|A| < +\infty$ *in Problem* (3.2.1)*. Then the AFW, the PFW and the FDFW converge at a rate*

$$f(x_k) - f(x^*) \leq M\bar{q}_{gs}^{\gamma_g(k)}, \tag{3.3.27}$$

for some $M > 0$*, with* $\gamma_g(k)$ *the number of good steps among the first* k *steps,*

$$\bar{q}_{gs} = \max\left(1 - \frac{\mu}{L}\left(\frac{\mathrm{PWidth}(A)}{2D}\right)^2, \left(1 + \frac{\mu}{L}\right)^{-1}\right) \tag{3.3.28}$$

for the AFW,

$$\bar{q}_{gs} = 1 - \frac{\mu}{L}\left(\frac{\mathrm{PWidth}(A)}{D}\right)^2 \tag{3.3.29}$$

for the PFW, and

$$\bar{q}_{gs} = \max\left(1 - \frac{\mu}{L}\left(\frac{\mathrm{PFWidth}(\Omega)}{2D}\right)^2, \left(1 + \frac{\mu}{L}\right)^{-1}\right) \tag{3.3.30}$$

for the FDFW.

Proof. For the AFW and the FDFW the rates (3.3.28) and (3.3.30) for good steps follow directly from (3.3.18) and (3.3.19) together with the bound on τ given in Proposition 3.3.3. Since the PFW never performs full FW steps, its rate (3.3.29) for good steps follow directly from (3.3.18) together with the bound on τ given in Proposition 3.3.3. Finally, given that the number of bad steps between two good steps is limited for all these methods (see [153, 157]), we have all the assumptions to apply Proposition 3.3.7. □

3.4 First order projection free methods with SSC procedure

We introduce here the SSC procedure, and prove convergence rates both under the KL inequality (3.2.4) and in the generic non convex case.

3.4.1 The SSC procedure

The SSC procedure chains consecutive short steps, thus skipping updates for the gradient (and possibly for related information, like linear minimizers), until proper stopping conditions are met. Such a procedure, whose detailed scheme is given in Algorithm 4, can be easily embedded in a first-order approach (see Algorithm 3).

Algorithm 3 First-order method with SSC

```
$x_0 \in \Omega$, $k = 0$.
while $x_k$ is not stationary do
    $g = -\nabla f(x_k)$.
    $x_{k+1} = \mathrm{SSC}(x_k, g)$.
    $k = k + 1$.
end while
```

Algorithm 4 $\mathrm{SSC}(\bar{x}, g)$

```
 1: Initialization. $y_0 = \bar{x}$, $j = 0$.
    Phase I
 2: select $d_j \in \mathcal{A}(y_j, g)$, $\alpha_{\max}^{(j)} \in \alpha_{\max}(y_j, d_j)$
 3: if $d_j = 0$ then
 4:     return $y_j$
 5: end if
    Phase II
 6: compute $\beta_j$ with (3.4.2)
 7: let $\alpha_j = \min(\alpha_{\max}^{(j)}, \beta_j)$
 8: $y_{j+1} = y_j + \alpha_j d_j$
 9: if $\alpha_j = \beta_j$ then
10:     return $y_{j+1}$
11: end if
12: $j = j + 1$, go to Step 2
```

Given that the gradient $-g$ is constant during the SSC, this procedure is an application of $\mathcal{A}$ for the minimization of the linearized objective $f_g(z) = -g^\top(z - \bar{x}) + f(\bar{x})$ with particular stepsizes and stopping criterion. More specifically, after a stationarity check (Phase I), the stepsize α_j is computed by taking the minimum between the maximal stepsize $\alpha_{\max}^{(j)}$ (which we always assume to be greater than 0) and an auxiliary stepsize β_j. Here $\alpha_{\max}(y_j, d_j)$ denotes the set of possible maximal stepsizes used by $\mathcal{A}$ from y_j in the direction d_j. The point y_{j+1} generated in Phase II is always feasible since $\alpha_j \leq \alpha_{\max}^{(j)}$ is always smaller than the maximal feasible stepsize along the direction d_j. Notice that if the method $\mathcal{A}$ used in the SSC performs a FW step (see equation (3.3.5) for the definition of FW step), then the SSC terminates, with $\alpha_j = \beta_j$ or with y_{j+1} global minimizer of f_g.

The auxiliary step size β_j is defined as the maximal feasible stepsize for the trust region

$$\Omega_j = \bar{B}_{\|g\|/2L}(\bar{x} + \frac{g}{2L}) \cap \bar{B}_{g^\top \hat{d}_j/L}(\bar{x}) \tag{3.4.1}$$

when $y_j \in \Omega_j$, otherwise the method stops returning y_j. Summarizing,

$$\beta_j = \begin{cases} 0 & \text{if } y_j \notin \Omega_j\,, \\ \beta_{\max}(\Omega_j, y_j, d_j) & \text{if } y_j \in \Omega_j\,, \end{cases} \tag{3.4.2}$$

where $\beta_{\max}(\Omega_j, y_j, d_j) = \max\{\beta \in \mathbb{R}_{\geq 0} \mid y_j + \beta d_j \in \Omega_j\}$ is the maximal feasible stepsize in the direction d_j starting from y_j with respect to Ω_j. Since Ω_j is the intersection of two balls there is a simple closed form expression for β_j. In particular, using that $y_0 = \bar{x}$, if $d_0 \neq 0$ we have

$$\beta_0 = \frac{g^\top \hat{d}_0}{L\|d_0\|}\,,$$

which corresponds to (3.3.13) in the non maximal case, and where $\beta_0 > 0$ since $d_0 \neq 0$ is by assumption a descent direction for $-g$.

Remark 3.4.1. When the Lipschitz constant L is not available, it can be approximated adaptively in the following way. At the step k we start with an estimate $\tilde{L} = L_k$ of the Lipschitz constant. Then, we compute x_k^+ with the procedure SSC$(x_k, -\nabla f(x_k))$, and repeat setting $\tilde{L} := 2\tilde{L}$ until

$$f(x_k) - f(x_k^+) \geq \frac{1}{2}\nabla f(x_k)^\top (x_k - x_k^+) \tag{3.4.3}$$

holds. When this happens, we set $x_{k+1} = x_k^+$ and $L_{k+1} = \tilde{L}$. The linear convergence results we will see later in this section can be extended in a straightforward way when L is approximated with this adaptive scheme.

Employing the trust region Ω_j in the definition of β_j guarantees the sufficient decrease condition

$$f(y_j) \leq f(x_k) - \frac{L}{2}\|x_k - y_j\|^2 \tag{3.4.4}$$

and monotonicity of the true objective f during the SSC.

To see why (3.4.4) holds, notice that the second ball $\bar{B} = \bar{B}_{\|g\|/2L}(x_k + \frac{g}{2L})$ appearing in the definition of Ω_j does not depend on j, so that since $y_0 \in \bar{B}$ we have $y_j \in \bar{B}$ for every $j \in [0 : T]$, with T maximal iteration index of the SSC. This is enough to obtain (3.4.4) because for every $z \in \bar{B}$ we have

$$f(z) \leq f(\bar{x}) - g^\top(z - \bar{x}) + \frac{L}{2}\|z - \bar{x}\|^2 \leq f(\bar{x}) - \frac{L}{2}\|\bar{x} - z\|^2 , \tag{3.4.5}$$

where the first inequality is the standard descent lemma and the second follows from the definition of $\bar{B}$.

We prove that the true objective f is monotone decreasing in the next lemma.

Lemma 3.4.2. *Let us assume $y_j \in \bar{B}_{g^\top \hat{d}_j/L}(\bar{x})$. Then for every $\beta \in [0, \beta_j]$ we have*

$$\frac{d}{d\beta} f(y_j + \beta d_j) \leq 0 ,$$

and thus in particular $f(y_j + \beta_j d_j) \leq f(y_j)$.

Proof. We have

$$\begin{aligned}
&\frac{d}{d\beta} f(y_j + \beta d_j) = \|d_j\| \nabla f(y_j + \beta d_j)^\top \hat{d}_j \\
=&\|d_j\|((\nabla f(y_j + \beta d_j) + g) - g)^\top \hat{d}_j = \|d_j\|((\nabla f(y_j + \beta d_j) + r)^\top \hat{d}_j - g^\top \hat{d}_j) \\
\leq&\|d_j\|(L\|\bar{x} - y_j - \beta d_j\| - g^\top \hat{d}_j) \leq 0 ,
\end{aligned}$$

where we used $g = -\nabla f(\bar{x})$ and the Lipschitzianity of ∇f in the first inequality and

$$y_j + \beta d_j \in \bar{B}_{g^\top \hat{d}_j/L}(\bar{x})$$

in the second. □

The next result illustrates how the sequence $\{x_k\}$ generated by Algorithm 3 satisfies certain descent conditions. This is an adaptation to our setting of the ones used in the analysis of many proximal type gradient methods (see [12], [13], [38] and references therein). A subtle difference is the introduction of an "hidden sequence" $\{\tilde{x}_k\}$ to control the projection of the negative gradient on the tangent cone.

Proposition 3.4.3. *Let us consider the sequence $\{x_k\}$ generated by Algorithm 3 and assume that*

- *the angle condition* (3.3.2) *holds;*
- *the SSC condition terminates in a finite number of steps.*

Then

$$f(x_k) - f(x_{k+1}) \geq \frac{L}{2}\|x_k - x_{k+1}\|^2 \,, \tag{3.4.6}$$

$$\|x_k - x_{k+1}\| \geq K\|\pi(T_\Omega(\tilde{x}_k), -\nabla f(\tilde{x}_k))\| \tag{3.4.7}$$

for some $\tilde{x}_k \in \{y_j\}_{j=0}^T$ such that $f(x_{k+1}) \leq f(\tilde{x}_k) \leq f(x_k) - \frac{L}{2}\|x_k - \tilde{x}_k\|^2$, $\|\tilde{x}_k - x_k\| \leq \|x_{k+1} - x_k\|$ and for $K = \tau/(L(1+\tau))$.

Proof. Let $B_j = \bar{B}_{g^\top \hat{d}_j/L}(x_k)$ and let T be such that $x_{k+1} = y_T$.
Inequality (3.4.4) applied with $j = T$ gives (3.4.6). Moreover, by taking $\tilde{x}_k = y_{\tilde{T}}$ for some $\tilde{T} \in [0 : T]$ the conditions

$$f(x_{k+1}) \leq f(\tilde{x}_k) \leq f(x_k) - \frac{L}{2}\|x_k - \tilde{x}_k\|^2 \tag{3.4.8}$$

are satisfied by Lemma 3.4.2 and (3.4.4).
Let now $p_j = \|\pi(T_\Omega(y_j), -\nabla f(y_j))\|$ and $\tilde{p}_j = \|\pi(T_\Omega(y_j), g)\| = \|\pi(T_\Omega(y_j), -\nabla f(x_k))\|$. We have

$$|p_j - \tilde{p}_j| \leq L\|y_j - x_k\| \,, \tag{3.4.9}$$

reasoning as for (3.3.20). We now distinguish four cases according to how the SSC terminates.
Case 1: $T = 0$ or $d_T = 0$. Since there are no descent directions $x_{k+1} = y_T$ must be stationary for the gradient g. Equivalently, $\tilde{p}_T = \|\pi(T_\Omega(x_{k+1}), g)\| = 0$. We can now write

$$\|x_{k+1} - x_k\| \geq \frac{1}{L}(|p_T - \tilde{p}_T|) = \frac{p_T}{L} > Kp_T \,,$$

where we used (3.4.9) in the first inequality and $\tilde{p}_T = 0$ in the equality. Finally, it is clear that if $T = 0$ then $d_0 = 0$, since y_0 must be stationary for $-g$.
Before examining the remaining cases we remark that if the SSC terminates in Phase II then $\alpha_{T-1} = \beta_{T-1}$ must be maximal w.r.t. the conditions $y_T \in B_{T-1}$ or $y_T \in \bar{B}$. If $\alpha_{T-1} = 0$ then $y_{T-1} = y_T$, and in this case we cannot have $y_{T-1} \in \partial \bar{B}$, otherwise the SSC would terminate in Phase II of the previous cycle. Therefore necessarily $y_T = y_{T-1} \in \text{int}(B_{T-1})^c$ (Case 2). If $\beta_{T-1} = \alpha_{T-1} > 0$ we must have

$y_{T-1} \in \Omega_{T-1} = B_{T-1} \cap \bar{B}$, and $y_T \in \partial B_{T-1}$ (case 3) or $y_T \in \partial \bar{B}$ (case 4) respectively.
Case 2: $y_{T-1} = y_T \in \mathrm{int}(B_{T-1})^c$. We can rewrite the condition as

$$g^\top \hat{d}_{T-1} \leq L\|y_{T-1} - x_k\| = L\|y_T - x_k\| \,. \tag{3.4.10}$$

Thus

$$p_T = p_{T-1} \leq \tilde{p}_{T-1} + L\|y_T - x_k\| \leq \frac{1}{\tau} g^\top \hat{d}_{T-1} + L\|y_T - x_k\| \leq \left(\frac{L}{\tau} + L\right) \|y_T - x_k\| \,, \tag{3.4.11}$$

where in the equality we used $y_T = y_{T-1}$, the first inequality follows from (3.4.9) and again $y_T = y_{T-1}$, the second from $\frac{g^\top \hat{d}_T}{\tilde{p}_T} \geq \mathrm{DSB}_{\mathcal{A}}(\Omega, y_T, g) \geq \mathrm{SB}_{\mathcal{A}}(\Omega) = \tau$, and the third from (3.4.10). Then $\tilde{x}_k = x_{k+1} = y_T$ satisfies the desired conditions.
Case 3: $y_T = y_{T-1} + \beta_{T-1} d_{T-1}$ and $y_T \in \partial B_{T-1}$. Then from $y_{T-1} \in B_{T-1}$ it follows

$$L\|y_{T-1} - x_k\| \leq g^\top \hat{d}_{T-1} \,, \tag{3.4.12}$$

and $y_T \in \partial B_{T-1}$ implies

$$g^\top \hat{d}_{T-1} = L\|y_T - x_k\| \,. \tag{3.4.13}$$

Combining (3.4.12) with (3.4.13) we obtain

$$L\|y_{T-1} - x_k\| \leq L\|y_T - x_k\| \,. \tag{3.4.14}$$

Thus

$$p_{T-1} \leq \tilde{p}_{T-1} + L\|y_{T-1} - x_k\| \leq \frac{1}{\tau} g^\top \hat{d}_{T-1} + L\|y_{T-1} - x_k\| \leq \left(\frac{L}{\tau} + L\right) \|y_T - x_k\| \,,$$

where we used (3.4.13), (3.4.14) in the last inequality and the rest follows reasoning as for (3.4.11). In particular we can take $\tilde{x}_k = y_{T-1}$, where $\|\tilde{x}_k - x_k\| \leq \|x_{k+1} - x_k\|$ by (3.4.14).
Case 4: $y_T = y_{T-1} + \beta_{T-1} d_{T-1}$ and $y_T \in \partial \bar{B}$.
The condition $x_{k+1} = y_T \in \bar{B}$ can be rewritten as

$$L\|x_{k+1} - x_k\|^2 - g^\top (x_{k+1} - x_k) = 0 \,. \tag{3.4.15}$$

For every $j \in [0 : T]$ we have

$$x_{k+1} = y_j + \sum_{i=j}^{T-1} \alpha_i d_i \,. \tag{3.4.16}$$

We now want to prove that for every $j \in [0 : T]$

$$\|x_{k+1} - x_k\| \geq \|y_j - x_k\| . \tag{3.4.17}$$

Indeed, we have

$$\begin{aligned} L\|x_{k+1} - x_k\|^2 = g^\top(x_{k+1} - x_k) = g^\top(y_j - x_k) + \sum_{i=j}^{T-1} \alpha_i g^\top d_i \\ \geq g^\top(y_j - x_k) \geq L\|y_j - x_k\|^2 , \end{aligned}$$

where we used (3.4.15) in the first equality, (3.4.16) in the second, $g^\top d_j \geq 0$ for every j in the first inequality and $y_j \in \bar{B}$ in the second inequality.
We also have

$$\begin{aligned} \frac{g^\top(x_{k+1} - x_k)}{\|x_{k+1} - x_k\|} = \frac{g^\top \sum_{j=0}^{T-1} \alpha_j d_j}{\| \sum_{j=0}^{T-1} \alpha_j d_j \|} \geq \frac{g^\top \sum_{j=0}^{T-1} \alpha_j d_j}{\sum_{j=0}^{T-1} \alpha_j \|d_j\|} \\ \geq \min \left\{ \frac{g^\top d_j}{\|d_j\|} \mid 0 \leq j \leq T-1 \right\} . \end{aligned} \tag{3.4.18}$$

Thus for $\tilde{T} \in \arg\min \left\{ \frac{g^\top d_j}{\|d_j\|} \mid 0 \leq j \leq T-1 \right\}$

$$g^\top \hat{d}_{\tilde{T}} \leq \frac{g^\top(x_{k+1} - x_k)}{\|x_{k+1} - x_k\|} = L\|x_{k+1} - x_k\| , \tag{3.4.19}$$

where we used (3.4.18) in the first inequality and (3.4.15) in the second.
We finally have

$$p_{\tilde{T}} \leq \tilde{p}_{\tilde{T}} + L\|y_{\tilde{T}} - x_k\| \leq \frac{1}{\tau} g^\top \hat{d}_{\tilde{T}} + L\|y_{\tilde{T}} - x_k\| \leq \left(\frac{L}{\tau} + L \right) \|x_{k+1} - x_k\| ,$$

where we used (3.4.17), (3.4.19) in the last inequality and the rest follows reasoning as for (3.4.11). In particular $\tilde{x}_k = y_{\tilde{T}}$ satisfies the desired properties, where $\|\tilde{x}_k - x_k\| \leq \|x_{k+1} - x_k\|$ by (3.4.17). □

3.4.2 SSC for Frank-Wolfe variants

In this section, we show how to apply our results to the PFW, the AFW and the FDFW on polytopes, i.e., we prove finite termination of the SSC procedure when one of these methods is considered in Algorithm 3. We also give worst case and average worst case bounds for the number of iterations of the SSC. We start by proving a general termination criterion.

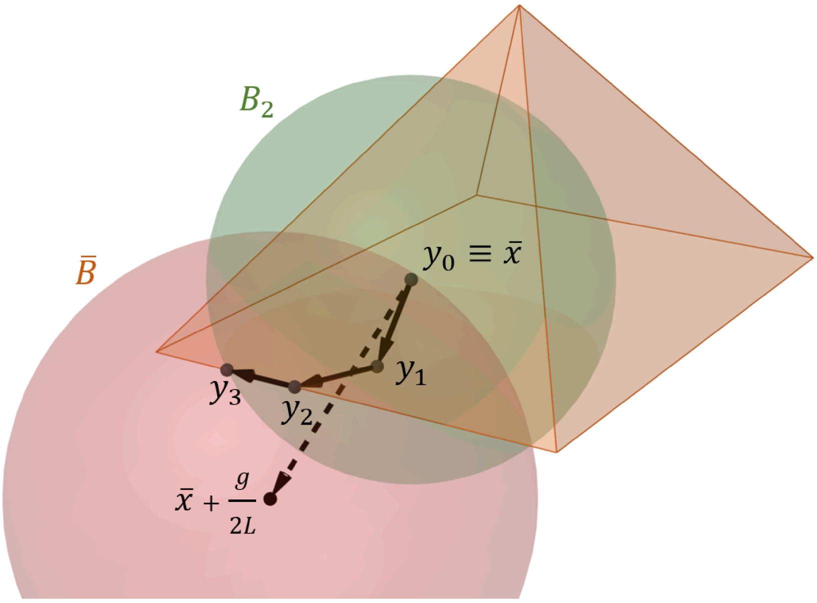


Figure 3.2: Instance of SSC with FDFW. $B_2 = \bar{B}_{g^\top \hat{d}_2/L}(\bar{x})$.

Lemma 3.4.4. *Assume that the method $\mathcal{A}$ applied to any linear function $L_g(x) = -g^\top x$ on the feasible set Ω and with every stepsize maximal always terminates in at most T iterations with an optimal solution, i.e. generates a sequence $\{y_j\}_{j\in[0,T']}$ with $T' \leq T$ and $y_{T'} \in \arg\min_{x\in\Omega} L_g(x)$. Then the SSC with the method $\mathcal{A}$ on the feasible set Ω always terminates in at most T iterations.*

Proof. Assume by contradiction that the SSC does at least $T+1$ iterations, generating the sequence $\{y_j\}_{j\in[0:T+1]}$ before terminating. Notice that in this case the SSC must always do maximal steps for $j \in [0 : T]$, because it terminates at step 9 when $\alpha_j = \beta_j$ and in particular if $\alpha_j < \alpha_{\max}^{(j)}$. Then for some $T' \leq T$ we must have that $y_{T'} \in \arg\min_{x\in\Omega} L_g(x)$, which gives a contradiction because in this case the method can't find a feasible descent direction in Phase I and terminates returning $y_{T'}$. □

Remark 3.4.5. Using the same line of reasoning, it is not difficult to prove that the SSC always terminates if the method $\mathcal{A}$ applied to linear objectives and with stepsizes always maximal generates a (possibly finite) sequence $\{y_j\}$ satisfying

$$\liminf \pi_{y_j}(g) = 0\,. \tag{3.4.20}$$

We now denote with $\{S^{(j)}\}$ the sequence of active sets generated by the AFW and the PFW method in the SSC, and with y_j proper convex combination of the elements in $S^{(j)}$. Furthermore, for the FDFW we assume that the maximal stepsize is given by feasibility conditions as in [103]:

$$\alpha_{\max}(x, d) = \{\alpha^{\max}(x, d)\}\,. \tag{3.4.21}$$

Notice that after a maximal in face step from y_j we have $\dim(\mathcal{F}(y_{j+1})) < \dim(\mathcal{F}(y_j))$ because y_{j+1} lies on the boundary of $\mathcal{F}(y_j)$.

Proposition 3.4.6. *The SSC always terminates in at most:*

- $|A|$ *iterations for the AFW,*
- $|A| - 1$ *iterations for the PFW,*
- $\dim(\Omega) + 1$ *iterations for the FDFW.*

Proof. By Lemma 3.4.4 we just need to bound the maximum number of iterations if the method performs always maximal steps for a linear objective $L_g(x)$. The AFW can do at most $|A| - 1$ consecutive maximal away steps, since at every such step the number of active atoms decreases by one. Analogously, the FDFW can do at most $\dim(\Omega)$ consecutive maximal in face steps, since at every such steps the dimension of the minimal face containing the current iterate decreases by one. The respective bound follows Lemma 3.4.4 by noticing that in the linear case the methods terminate after a full FW step. For the PFW, the linearity of the objective implies that only atoms in $\bar{A} := \arg\max_{a \in A} g^\top x$ can be added to the support, and only atoms in $A \setminus \bar{A}$ can be dropped from the support. In particular, once an atom is dropped from the active set it cannot be added again, and since at every maximal step the PFW drops an atom from the active set its maximal number of iterations is $|A \setminus \bar{A}| \leq |A| - 1$. □

We now define and give a bound on the worst case average number of SSC iterations. Let $T(k)$ be the number of points generated by the SSC at the step k. Then we define the worst case average number of SSC iterations as the supremum of

$$\lim_{k \to \infty} \frac{1}{k} \sum_{i=0}^{k-1} T(i) \tag{3.4.22}$$

over all the possible realizations of Algorithm 3 (of course under specific assumptions on $\mathcal{A}$).

The proof of the following result uses analogous arguments to the ones in [157, Theorem 8] to bound the number of bad steps.

Proposition 3.4.7. *Assume that the linear minimizer is not changed during the SSC. Then, for an infinite sequence $\{x_k\}$, the worst case average number of iterations is*

- *2 for the AFW and the PFW,*
- $\Delta(\Omega) + 1$ *for the FDFW.*

Proof. Let $T(k)$ be the number of iterates generated by the SSC at the step k in Phase II. For the AFW and the PFW, reasoning as in the proof of Proposition 3.4.6 we obtain that if the SSC does $T(k)$ iterations, the number of active vertices decreases by at least $T(k) - 2$. Then on the one hand

$$|S^{(k)}| - |S^{(0)}| \geq 1 - |S^{(0)}|, \tag{3.4.23}$$

while on the other hand

$$\begin{aligned} |S^{(k)}| - |S^{(0)}| &= \sum_{i=0}^{k-1}(|S^{(i+1)}| - |S^{(i)}|) \\ &\leq 2k - \sum_{i=0}^{k-1} T(i)\,. \end{aligned} \tag{3.4.24}$$

Combining (3.4.23) and (3.4.24) and rearranging, we obtain:

$$\frac{1}{k}\sum_{i=0}^{k-1} T(i) \leq 2 + \frac{|S^{(0)}| - 1}{k}, \tag{3.4.25}$$

and the desired result follows by taking the limit for $k \to \infty$.

For the FDFW, notice that at every iteration the SSC performs a sequence of maximal in face steps terminated either by a Frank Wolfe step, after which $\mathcal{F}(y_j)$ can increase of at most $\Delta(\Omega)$, or by a non maximal in face step, after which $\mathcal{F}(y_j)$ stays the same. In both cases, we have

$$\dim(\mathcal{F}(x_{k+1})) - \dim(\mathcal{F}(x_k)) \leq \Delta(\Omega) - T(k) + 1. \tag{3.4.26}$$

Then,

$$\dim \mathcal{F}(x_k) - \dim \mathcal{F}(x_0) \geq -\dim \mathcal{F}(x_0)\,, \tag{3.4.27}$$

and

$$\begin{aligned} \dim \mathcal{F}(x_k) - \dim \mathcal{F}(x_0) &= \sum_{i=0}^{k-1}(\dim(\mathcal{F}(x_{i+1}) - \dim(\mathcal{F}(x_i)))) \\ &\leq k\Delta(\Omega) + k - \sum_{i=0}^{k-1} T(i)\,. \end{aligned} \tag{3.4.28}$$

The conclusion follows as for the AFW and the PFW. □

3.4.3 Convergence rates

Smooth non convex objectives

We first prove, in the generic smooth non convex case, convergence to the set of stationary points with a rate of $O(\frac{1}{\sqrt{k}})$ for $\|\pi(T_\Omega(\tilde{x}_i), -\nabla f(\tilde{x}_i))\|$.

Theorem 3.4.8. *Let us consider the sequence* $\{x_k\}$ *generated by Algorithm 3 and assume that*

- *the angle condition* (3.3.2) *holds;*
- *the SSC procedure always terminates in a finite number of steps.*

Then $\{f(x_k)\}$ *is decreasing,* $f(x_k) \to \tilde{f}^* \in \mathbb{R}$ *and the limit points of* $\{x_k\}$ *are stationary. Furthermore, for any sequence* $\{\tilde{x}_k\}$ *satisfying the conditions of Proposition 3.4.3, we have* $\|\tilde{x}_k - x_k\| \to 0$*, and*

$$\min_{0\le i\le k} \|\pi(T_\Omega(\tilde{x}_i), -\nabla f(\tilde{x}_i))\| \le \min_{0\le i\le k} \frac{\|x_{i+1} - x_i\|}{K} \le \sqrt{\frac{2(f(x_0) - \tilde{f}^*)}{K^2 L(k+1)}}\,, \tag{3.4.29}$$

for $K = \tau/(L(1+\tau))$.

Proof. The sequence $\{f(x_k)\}$ is decreasing by (3.4.6). Thus by compactness $f(x_k) \to \tilde{f}^* \in \mathbb{R}$ and in particular $f(x_k) - f(x_{k+1}) \to 0$. So that by (3.4.6) also $\|x_{k+1} - x_k\| \to 0$. Let $\{x_{k(i)}\} \to \tilde{x}^*$ be any convergent subsequence of $\{x_k\}$. For $\{\tilde{x}_k\}$ chosen as in the proof of Proposition 3.4.3 we have $\|\tilde{x}_k - x_k\| \le \|x_{k+1} - x_k\|$ because $\tilde{x}_k = y_T = x_k$ in case 1 and case 2, by (3.4.14) in case 3, and by (3.4.17) in case 4. Therefore

$$\|\tilde{x}_{k(i)} - x_{k(i)}\| \le \|x_{k(i)+1} - x_{k(i)}\| \to 0\,.$$

Furthermore, $\|\pi(T_\Omega(\tilde{x}_{k(i)}), -\nabla f(\tilde{x}_{k(i)})))\| \le \frac{\|x_{k(i)+1} - x_{k(i)}\|}{K} \to 0$ again by Proposition 3.4.3, so that $\tilde{x}_{k(i)} \to \tilde{x}^*$ with $\|\pi(T_\Omega(\tilde{x}_{k(i)}), -\nabla f(\tilde{x}_{k(i)}))\| \to 0$. Then

$$\|\pi(T_\Omega(\tilde{x}^*), -\nabla f(\tilde{x}^*))\| = 0$$

and $\tilde{x}^*$ is stationary.

The first inequality in (3.4.29) follows directly from (3.4.7). As for the second, we have

$$\begin{aligned}
&\frac{k+1}{K^2}(\min_{0\le i\le k} \|x_{i+1} - x_i\|)^2 = \frac{k+1}{K^2} \min_{0\le i\le k} \|x_{i+1} - x_i\|^2 \\
\le &\frac{1}{K^2} \sum_{i=0}^{k} \|x_i - x_{i+1}\|^2 \le \frac{2}{LK^2} \sum_{i=0}^{k} (f(x_{i+1}) - f(x_i)) \le \frac{2(f(x_0) - \tilde{f}^*)}{LK^2}\,,
\end{aligned}$$

Algorithm	Article	LMO c.r.	Gradient c.r.	Gap
NCGS	[205]	$O\left(\frac{1}{k^{0.25}}\right)$	$O\left(\frac{1}{\sqrt{k}}\right)$	$\min_{0\le i\le k} \pi(x_i)$
AFW, FW	[47], [156]	$O\left(\frac{1}{\sqrt{k}}\right)$	$O\left(\frac{1}{\sqrt{k}}\right)$	$\min_{0\le i\le k} G(x_i)$
AFW, PFW, FDFW + SSC	Ours	$O\left(\frac{1}{\sqrt{k}}\right)$	$O\left(\frac{1}{\sqrt{k}}\right)$	$\min_{0\le i\le k} \|\pi(T_\Omega(\tilde{x}_i), -\nabla f(\tilde{x}_i))\|$

Table 3.2: Comparison between convergence rates in the generic smooth non convex case. See also Remark 3.4.10. $\pi(x) = \|x - \pi\left(\Omega, x - \frac{\nabla f(x)}{2L}\right)\|$, G is the FW gap (see Section 2.6.1).

where we used (3.4.6) in the first inequality, $\{f(x_i)\}$ decreasing together with $f(x_i) \to \tilde{f}^*$ in the second and the thesis follows by rearranging terms. □

We now give a corollary for Theorem 3.4.8 specialized to the FW variants described in Section 3.3.1 (see also Table 3.2).

Corollary 3.4.9. *Let us assume that* $\Omega = \text{conv}(A)$*, with* $|A| < +\infty$ *in Problem* (3.2.1). *Then the sequence* $\{x_k\}$ *generated by Algorithm 3 with AFW (PFW or FDFW) in the SSC converges at a rate given by equation* (3.4.29)*, with* $\tau = \tau_p/2$ *(*τ_p *or* $\tau_v/2$*, respectively).*

Proof. Finite termination of the SSC follows by Proposition 3.4.6, and the angle condition is satisfied by Proposition 3.3.3. Thus we have all the assumptions to apply Theorem 3.4.8. □

Remark 3.4.10. Notice that in Table 3.2 we use the Frank Wolfe gap (see Section 2.6.1) as a measure of convergence. By combining equation (3.2.3) with (2.6.8), we obtain, for any $y \in \Omega$

$$G(y) \le D\|\pi(T_\Omega(y), -\nabla f(y))\|\,. \tag{3.4.30}$$

Taking into account equation (3.4.30), it is easy to see that our rate is an improvement of the ones proved in [156] and [47] (see Table 3.2). Furthermore, we do not need to start from a vertex to avoid dependence from the support of $\{x_0\}$ like in [47, Theorem 5.1]. Finally, our method improves the conditional gradient sliding rate (NCGS) not only in LMO but also in gradients, given that from $\Omega - \{y\} \subset T_\Omega(y)$ it follows $\pi(y) \le \|\pi(T_\Omega(y), -\nabla f(y))\|/2L$ for every $y \in \Omega$.

Objectives with KL property

As a consequence of Proposition 3.4.3, we have linear convergence rates for the general algorithmic scheme reported in Algorithm 3 under the KL inequality (3.2.4), the angle condition (3.3.2), and finite termination of the SSC procedure. In the next results (Lemma 3.4.11, Theorem 3.4.13 and Corollary 3.4.14), we always assume the following:

- the angle condition (3.3.2) holds;
- the SSC procedure always terminates in a finite number of steps.

Lemma 3.4.11. *Let us consider the sequence $\{x_k\}$ generated by Algorithm 3 and assume that the objective function f satisfies condition* (3.2.4)*, with $f(x^*)$ fixed, in every feasible point generated by the algorithm. Then, for $q = \left(1 + \frac{\mu}{L}\frac{\tau^2}{(1+\tau)^2}\right)^{-1}$ we have $f(x_k) \to f(x^*)$, with*

$$f(x_k) - f(x^*) \leq q^k(f(x_0) - f(x^*)) , \tag{3.4.31}$$

and $x_k \longrightarrow \tilde{x}^*$ with

$$\|x_k - \tilde{x}^*\| \leq \frac{\sqrt{2-2q}(f(x_0) - f(\tilde{x}^*))}{\sqrt{L}(1-\sqrt{q})} q^{\frac{k}{2}} , \tag{3.4.32}$$

for $\tilde{x}^*$ stationary point such that $f(\tilde{x}^*) = f(x^*)$.

In order to prove Lemma 3.4.11 we first need a technical Lemma based on Karamata's inequality ([143], [144]) for concave functions. We now recall the inequality. Given $A, B \in \mathbb{R}^N$ it is said that A majorizes B, written $A \succ B$, if

$$\begin{aligned} \sum_{i=1}^{j} A_i &\geq \sum_{i=1}^{j} B_i \text{ for } j \in [1:N] , \\ \sum_{i=1}^{N} A_i &= \sum_{i=1}^{N} B_i . \end{aligned}$$

If h is concave and $A \succ B$ by Karamata's inequality

$$\sum_{i=1}^{N} h(A_i) \leq \sum_{i=1}^{N} h(B_i) .$$

We can now state and prove the technical lemma.

Lemma 3.4.12. *Let $\{\tilde{f}_i\}_{i \in [0:j]}$ be a sequence of nonnegative numbers such that $\tilde{f}_{i+1} \le q\tilde{f}_i$ for some $q < 1$. Then*

$$\sum_{i=0}^{j-1} \sqrt{\tilde{f}_i - \tilde{f}_{i+1}} \le \frac{\sqrt{\tilde{f}_0(1-q)}}{1-\sqrt{q}} \,. \tag{3.4.33}$$

Proof. Let $\bar{j} = \max\{i \ge 0 \mid \tilde{f}_j \le q^i \tilde{f}_0\}$, so that by (3.4.41) we have $\bar{j} \ge j$. Define $w^*, v \in \mathbb{R}^{\bar{j}+1}_{\ge 0}$ by

$$\begin{aligned} v &= (\tilde{f}_0 - q\tilde{f}_0, ..., q^{\bar{j}-1}\tilde{f}_0 - q^{\bar{j}}\tilde{f}_0, q^{\bar{j}}\tilde{f}_0 - \tilde{f}_j)\,, \\ w^* &= (\tilde{f}_0 - \tilde{f}_1, ..., \tilde{f}_{j-1} - \tilde{f}_j, 0, ..., 0)\,. \end{aligned} \tag{3.4.34}$$

Then for $0 \le l < \bar{j}$ we have

$$\sum_{i=0}^{l} v_i = \tilde{f}_0 - q^{l+1}\tilde{f}_0 \le \tilde{f}_0 - \tilde{f}_{\min(l+1,j)} = \sum_{i=0}^{l} w_i^* \,, \tag{3.4.35}$$

where we used $q^{l+1}\tilde{f}_0 \ge \tilde{f}_{l+1}$ for $l \le j-1$ and $q^{l+1}\tilde{f}_0 \ge \tilde{f}_j$ for $j \le l < \bar{j}$ in the inequality. Furthermore, for $l = \bar{j}$ we have

$$\sum_{i=0}^{l} v_i = \tilde{f}_0 - \tilde{f}_j = \sum_{i=0}^{l} w_i^* \,. \tag{3.4.36}$$

Now if w is the permutation in descreasing order of w^*, clearly thanks to (3.4.35), and (3.4.36) we have $w \succ v$. Then

$$\begin{aligned} \sum_{i=0}^{j-1} \sqrt{\tilde{f}_i - \tilde{f}_{i+1}} &= \sum_{i=0}^{\bar{j}+1} \sqrt{w_i^*} = \sum_{i=0}^{\bar{j}+1} \sqrt{w_i} \le \sum_{i=0}^{\bar{j}+1} \sqrt{v_i} \\ &\le \sqrt{\tilde{f}_0} \sum_{i=0}^{+\infty} \sqrt{q^i - q^{i+1}} = \frac{\sqrt{\tilde{f}_0(1-q)}}{1-\sqrt{q}} \,, \end{aligned} \tag{3.4.37}$$

where the first inequality follows from Karamata's inequality. □

Proof of Lemma 3.4.11. If the sequence $\{x_k\}$ is finite, with $x_m = \tilde{x}_m$ stationary for some $m \ge 0$, we define $x_k = x_m$ for every $k \ge m$, so that we can always assume $\{x_k\}$ infinite. Notice that with this convention the sufficient decrease condition (3.4.6) is still satisfied for every k. Let $f_k = f(x_k) - f(x^*)$. $\{f_k\}$ is monotone decreasing by (3.4.6), and nonnegative since (3.2.4) holds for every x_k.

We want prove $f_{k+1} \leq q f_k$. This is clear if $f_{k+1} = 0$. Otherwise using the notation of Proposition 3.4.3 we have

$$f_k - f_{k+1} \geq \frac{L}{2} \|x_k - x_{k+1}\|^2 \geq \frac{LK^2}{2} \|\pi(T_\Omega(\tilde{x}_k), -\nabla f(\tilde{x}_k))\| , \tag{3.4.38}$$

where we used (3.4.6) in the first inequality, (3.4.7) in the second. Since $\tilde{x}_k \in \{y_j\}_{j=0}^T$ by Proposition 3.4.3, we can apply (3.2.4) in $\tilde{x}_k$ to obtain

$$\frac{LK^2}{2} \|\pi(T_\Omega(\tilde{x}_k), -\nabla f(\tilde{x}_k))\|^2 \geq \mu L K^2 (f(\tilde{x}_k) - f(x^*)) \geq \mu L K^2 f_{k+1}. \tag{3.4.39}$$

Concatenating (3.4.38), (3.4.39) and rearranging we obtain

$$f_{k+1} \leq (1 + \mu L K^2)^{-1} f_k = q f_k \, . \tag{3.4.40}$$

Thus by induction for any $i \geq 0$

$$f_{k+i} \leq q^i f_k \, , \tag{3.4.41}$$

which implies in particular (3.4.31).
We can now bound the length of the tails of $\{x_k\}$:

$$\begin{aligned} \sum_{i=0}^{+\infty} \|x_{k+i} - x_{k+i+1}\| &\leq \sqrt{\frac{2}{L}} \sum_{i=0}^{+\infty} \sqrt{f_{k+i} - f_{k+i+1}} \\ &\leq \frac{\sqrt{2 f_k (1-q)}}{\sqrt{L}(1-\sqrt{q})} \leq \frac{\sqrt{2 f_0 (1-q)}}{\sqrt{L}(1-\sqrt{q})} q^{\frac{k}{2}} \, , \end{aligned} \tag{3.4.42}$$

where we used (3.4.6) in the first inequality, Lemma 3.4.12 with $\{\tilde{f}_i\} = \{f_{k+i}\}$ and for $j \rightarrow +\infty$ in the second inequality, and (3.4.41) in the third. In particular $x_k \rightarrow \tilde{x}^*$ with

$$\|x_k - \tilde{x}^*\| \leq \sum_{j=0}^{+\infty} \|x_{k+j} - x_{k+j+1}\| = \frac{\sqrt{2 f_0 (1-q)}}{\sqrt{L}(1-\sqrt{q})} q^{\frac{k}{2}} \tag{3.4.43}$$

by (3.4.42). □

The KL assumption of Lemma 3.4.11 is trivially true if (3.2.4) holds globally for every x^* in the set of solutions of Problem (3.2.1); an analogous assumption is used in [146] for the PL property. By [38, Corollary 6], for convex objectives this assumption is satisfied in particular under a global quadratic Holderian error bound, thus, e.g., by strongly convex objectives.
Under mild assumptions on the stationary point x^*, we can also apply Lemma 3.4.11 locally on non convex objectives, thus adapting to our projection free setting the local results given in [13, Section 2.3] for proximal methods.

Theorem 3.4.13. *Let Assumption 3.1 hold at* x^*. *Further assume that* $x_k \in B_\delta(x^*) \Rightarrow f(x_{k+1}) \geq f(x^*)$. *Then, for some* $\tilde{\delta} > 0$, *if* $x_0 \in B_{\tilde{\delta}}(x^*)$ *the rates* (3.4.31) *and* (3.4.32) *hold.*

Proof. By continuity, for $\tilde{\delta} \to 0$ and $f_0 = f(x_0) - f(x^*)$ we have that

$$\max_{x_0 \in B_{\tilde{\delta}}(x^*) \cap [f \geq f(x^*)]} f_0 \to 0\,, \tag{3.4.44}$$

so we can take $\tilde{\delta} < \delta/2$ small enough in such a way that

$$\max_{x_0 \in B_{\tilde{\delta}}(x^*) \cap [f \geq f(x^*)]} \frac{\sqrt{2f_0(1-q)}}{L(1-\sqrt{q})} + \sqrt{\frac{2}{L}}\sqrt{f_0} < \frac{\delta}{2}\,. \tag{3.4.45}$$

Let now $x_0 \in B_{\tilde{\delta}}(x^*) \cap [f \geq f(x^*)]$, so that

$$\tilde{\delta} < \frac{\delta}{2} < \delta - \frac{\sqrt{2f_0(1-q)}}{L(1-\sqrt{q})} - \sqrt{\frac{2}{L}}\sqrt{f_0}\,, \tag{3.4.46}$$

where we use (3.4.45) in the second inequality. We now want to prove, by induction on k, $\{x_i\}_{i\in[0:k]} \subset B_\delta(x^*)$ with $f(x_{i+1}) \leq qf(x_i)$ for every $i \in [0:k]$ and $k \in \mathbb{N}$. To start with,

$$\sum_{i=0}^{k-1} \|x_i - x_{i+1}\| \leq \sqrt{\frac{2}{L}} \sum_{i=0}^{k-1} \sqrt{f_i - f_{i+1}} \leq \frac{\sqrt{2f_0(1-q)}}{\sqrt{L}(1-\sqrt{q})} \tag{3.4.47}$$

where we used (3.4.6) in the first inequality, and Lemma 3.4.12 (which we can apply thanks to the inductive assumption) in the second. But then

$$\begin{aligned}
\|x_{k+1} - x^*\| &\leq \|x_0 - x^*\| + \left(\sum_{i=0}^{k-1} \|x_i - x_{i+1}\|\right) + \|x_k - x_{k+1}\| \\
&\leq \tilde{\delta} + \frac{\sqrt{2f_0(1-q)}}{L(1-\sqrt{q})} + \sqrt{\frac{2}{L}}\sqrt{f_k - f_{k+1}} \\
&< \tilde{\delta} + \frac{\sqrt{2f_0(1-q)}}{L(1-\sqrt{q})} + \sqrt{\frac{2}{L}}\sqrt{f_k} < \delta\,,
\end{aligned} \tag{3.4.48}$$

where we used (3.4.47) together with (3.4.6) in the second inequality, the assumption $x_k \in B_\delta(x^*) \Rightarrow f_{k+1} \geq 0$ in the third inequality, and (3.4.46) together with $f_0 \geq f_k$ in the last inequality.

We now have

$$\begin{aligned}
\|\tilde{x}_k - x^*\| &\leq \|x_0 - x^*\| + \left(\sum_{i=0}^{k-1} \|x_i - x_{i+1}\|\right) + \|x_k - \tilde{x}_k\| \\
&\leq \|x_0 - x^*\| + \left(\sum_{i=0}^{k-1} \|x_i - x_{i+1}\|\right) + \|x_k - x_{k+1}\| < \delta\,,
\end{aligned} \tag{3.4.49}$$

where we use $\|\tilde{x}_k - x_k\| \leq \|x_{k+1} - x_k\|$ in the second inequality and the last inequality follows as in (3.4.48). Thus $\tilde{x}_k \in B_\delta(x^*)$ as well, which is enough to prove (3.4.40) and complete the induction. We have thus obtained $\{\tilde{x}_k\}, \{x_k\} \subset B_\delta(x^*)$, and the conclusion follows exactly as in the proof of Lemma 3.4.11. □

It is not difficult to see that the assumption $x_k \in B_\delta(x^*) \Rightarrow f(x_{k+1}) \geq f(x^*)$ is true, e.g., if x^* is a minimizer on its connected component of the sublevel set $[f \leq f(x_0)]$.

As a corollary of Theorem 3.4.13, we can apply Lemma 3.4.11 and derive the following asymptotic rates.

Corollary 3.4.14. *Let us consider the sequence $\{x_k\}$ generated by Algorithm 3. Let Assumption 3.1 hold at every point of the limit set of $\{x_k\}$. Then, for some positive constants M and $\tilde{M}$, $\{x_k\} \to x^*$, with the asymptotic rates:*

$$\begin{aligned} f(x_k) - f(x^*) &\leq Mq^k \,, \\ \|x_k - x^*\| &\leq \tilde{M} q^{\frac{k}{2}} \,. \end{aligned} \tag{3.4.50}$$

Proof. Let x^* be a limit point of $\{x_k\}$, and let $\tilde{\delta}$ be as in Theorem 3.4.13. First, for some $\bar{k} \in \mathbb{N}$ we must have $x_{\bar{k}} \in B_{\tilde{\delta}}(x^*)$. Furthermore, for every $k \in \mathbb{N}$ we have $f(x_k) \geq f(x^*)$ because $f(x_k)$ is non increasing and converges to $f(x^*)$. Thus we have all the necessary assumptions to obtain the asymptotic rates by applying Theorem 3.4.13 to $\{y_k\} = \{x_{\bar{k}+k}\}$. □

Similarly to what we did for Theorem 3.4.8, here we give a corollary for Lemma 3.4.11 related to the FW variants described in Section 3.3.1.

Corollary 3.4.15. *Let us assume that the objective function f satisfies condition* (3.2.4) *on every point generated by the algorithm, with $f(x^*)$ fixed, and that* $\Omega = \mathrm{conv}(A)$ *with* $|A| < +\infty$ *in Problem* (3.2.1). *Then the sequence $\{x_k\}$ generated by Algorithm 3 with AFW (PFW or FDFW) in the SSC converges at the rates given by Lemma 3.4.11, with $\tau = \tau_p/2$ (τ_p or $\tau_v/2$, respectively).*

Proof. Finite termination of the SSC follows by Proposition 3.4.6, and the angle condition is satisfied by Proposition 3.3.3. Thus we have all the assumptions to apply Lemma 3.4.11. □

For comparison, we now recall some well-known result related to global linear convergence rates for the FW variants under analysis.

Algorithm	Article	Objective	$\gamma(k)$	I_b	q_{gs}	h_k/h_0 upper bound	T_{avg}
AFW	[157]	SC	$k/2$	$\|S_0\| - 1$	$1 - \frac{\mu}{L}\frac{\tau_p^2}{4}$	$\left(1 - \frac{\mu}{L}\frac{\tau_p^2}{4}\right)^{\frac{k}{2}}$	-
PFW	[157]	SC	$k/(3\|A\|! + 1)$	-	$1 - \frac{\mu}{L}\tau_p^2$	$\left(1 - \frac{\mu}{L}\tau_p^2\right)^{\frac{k}{3\|A\|!+1}}$	-
FDFW^2	[153]	SC	$k/(\Delta(\Omega) + 1)$	$\dim(\mathcal{F}(x_0))$	$1 - \frac{\mu}{L}\frac{\tau_v^2}{4}$	$\left(1 - \frac{\mu}{L}\frac{\tau_v^2}{4}\right)^{\frac{k}{\Delta(\Omega)+1}}$	-
AFW + SSC	Ours	NC, KL	k	-	$\left(1 + \frac{\mu}{L}\frac{\tau_p^2}{(2+\tau_p)^2}\right)^{-1}$	$\left(1 + \frac{\mu}{L}\frac{\tau_p^2}{(2+\tau_p)^2}\right)^{-k}$	2
PFW + SSC	Ours	NC, KL	k	-	$\left(1 + \frac{\mu}{L}\frac{\tau_p^2}{(1+\tau_p)^2}\right)^{-1}$	$\left(1 + \frac{\mu}{L}\frac{\tau_p^2}{(1+\tau_p)^2}\right)^{-k}$	2
FDFW + SSC	Ours	NC, KL	k	-	$\left(1 + \frac{\mu}{L}\frac{\tau_v^2}{(1+\tau_v)^2}\right)^{-1}$	$\left(1 + \frac{\mu}{L}\frac{\tau_v^2}{(1+\tau_v)^2}\right)^{-k}$	$\Delta(\Omega) + 1$

Table 1: Comparison between the rates of the standard and SSC version of some FW variants for $\Omega = \text{conv}(A)$ with $|A| < \infty$. SC = strongly convex, NC = non convex, KL = KL property. $\gamma(k)$: lower bound on the number of good steps after k steps, counting from the first good step. I_b: bound on the number of bad steps before the first good step. q_{gs}: rate in good steps. h_k/h_0 upper bound: worst case rate assuming no initial bad steps, equal to $q_{gs}^{\gamma(k)}$. $\Delta(\Omega)$ = maximum increase in face dimension $\mathcal{F}(x_{k+1}) - \mathcal{F}(x_k)$ after a FW step. S_0 = active set for x_0. T_{avg} = worst case average iteration number of the SSC (see Proposition 3.4.7)

Proposition 3.4.16. *Let us assume that the objective function f is $\mu-$strongly convex and $\Omega = \text{conv}(A)$ with $|A| < +\infty$ in Problem (3.2.1). Let $\{x_k\}$ be a sequence generated by the AFW (PFW or FDFW), with stepsize given by exact line search. If the initial active set is $S_0 = \{x_0\}$ for the AFW ($S_0 = \{x_0\}$ for the PFW, $\dim(\mathcal{F}(x_0)) = 0$ for the FDFW), then*

$$f(x_k) - f^* \leq q_{gs}^{\gamma(k)}(f(x_0) - f^*)\,, \tag{3.4.51}$$

for $\gamma(k)$ and q_{gs} given in Table 1.

Proof. For the AFW and the PFW the result follows directly from [157, Theorem 1], with the exception of the good steps rate for the PFW, which can be obtained by applying the bound [157, Equation 10] in [157, Equation 5]. For the FDFW the result follows from [153, Theorem 1] (where the method is referred to as DiCG), with the bound $\mu\text{PWidth}(V(\Omega))^2$ on the geometric strong convexity constant implied by [157, Theorem 6] improved to $\mu\text{PFWidth}(\Omega)^2$ as in Proposition 3.3.3. □

For all the examples where an upper bound on $\tau_p = \frac{\text{PWidth}(A)}{D}$ is known (see [206], [200] and references therein) when $\dim(\text{conv}(A)) \to \infty$ then $\tau_p \to 0$ and our rates for the SSC converge to the rates without SSC for good steps in Table 1.

While we are not able to prove this limit in general, for all polytopes with dimension greater or equal to 2, except low dimensional simplices (see Example 3.4.17), we still have $\tau_p \leq \frac{1}{2}$ (because $\mathrm{PdirW}(A, g, x) + \mathrm{PdirW}(A, -g, x) \leq D$ for x in the relative interior of $\mathrm{conv}(A)$ and $\pm g$ feasible and orthogonal to $\mathrm{conv}(S)$ for some $S \in S_x$). Using this together with Example 3.4.17 for simplices, it is easy to check that the rates in Corollary 3.4.15 (SSC based FW variants) are strict improvements on the known worst case rates (standard FW variants) reported in Proposition 3.4.16, with a limited number of exceptions. These are the trivial one dimensional case and simplices with low dimension (≤ 4 for the PFW, and ≤ 8 for the AFW using the loose bounds in Example 3.4.17) combined with objectives having condition number μ/L sufficiently close to 1.

Example 3.4.17. If $W(\mathrm{conv}(A))$ is the width of $\mathrm{conv}(A)$ (see [157, Section 3]) then it follows directly from the definition of PWidth that $W(\mathrm{conv}(A)) \geq \mathrm{PWidth}(A)$, with equality for $A = \{e_1, ..., e_n\}$ (see [157] and [200]). Let now $A = \{a_1, ..., a_n\}$ be a set of n affinely independent points in $\mathbb{R}^{n-1}$. We claim that, for $r_n = \sqrt{1 - \frac{1}{n}}$ circumradius of the $n - 1$ dimensional unit simplex Δ_{n-1}

$$\mathrm{PWidth}(A)/D \leq r_n^{-1} W(\Delta_{n-1}) = \begin{cases} 2r_n^{-1}\sqrt{\frac{1}{n}} & \text{for } n \text{ even,} \\ 2r_n^{-1}\sqrt{\frac{1}{n-1/n}} & \text{for } n \text{ odd.} \end{cases} \tag{3.4.52}$$

To see this, assume without loss of generality $D = 1$ and $0 \in \mathrm{int}(\Omega)$ for $\Omega = \mathrm{conv}(A)$. Then if $A_S = \{\hat{a}_1, ..., \hat{a}_n\}$ we have $W(\mathrm{conv}(A_S)) \geq W(\mathrm{conv}(A))$. We can conclude

$$\frac{\mathrm{PWidth}(A)}{D} = \mathrm{PWidth}(A) \leq W(\mathrm{conv}(A)) \leq W(\mathrm{conv}(A_S)) \leq r_n^{-1} W(\Delta_{n-1}), \tag{3.4.53}$$

where in the last inequality we used that regular simplices maximize the width among simplices with fixed inradius (see, e.g., [9] and [115]).

Remark 3.4.18. The two main assumptions we make on the algorithm in this section are the angle condition and finite termination of the SSC. When the angle condition fails, like for the FW method when the solution is on the boundary, we expect the method to exhibit the zigzagging behaviour mentioned in Section 2.6.2. As for finite termination, given the very mild convergence properties necessary to achieve it discussed in Remark 3.4.5, when it is violated the algorithm might not converge at all even without SSC.

3.5 Examples

We now discuss some examples of objectives satisfying the KL property and sets where the angle condition can be satisfied with an explicit bound, relevant to practical optimization problems.

3.5.1 KL property

The KL property of Assumption 3.1 is satisfied for Problem (3.2.1) in the following cases:

- f is composite strongly convex, i.e. $f(x) = g(Bx)$ with g strongly convex, and Ω is a polytope [170, Proposition 4.1],

- f is composite strongly convex as in the previous point, Ω is the l^p ball for $p \in [1,2]$, and $\inf_{x\in\Omega} f(x) > \inf_{x\in\mathbb{R}^n} g(Bx)$ [170, Proposition 4.2],

- f is (non convex) quadratic, i.e. $f(x) = x^\top Qx + b^\top x + c$, and Ω is a polytope, [170, Corollary 5.2],

- f is non convex quadratic and does not satisfy the degeneracy condition of [138, equation (30)], and Ω is the unit sphere [138, Theorem 3.13].

- f is a nonlinear least square objective with full row rank Jacobian, and x^* is in the interior of Ω (see [82, Theorem 2] for a special case that easily generalizes to the desired property).

3.5.2 Angle condition bounds

Bounds using PWidth

For the unit simplex and the unit cube explicit $\Theta(1/\sqrt{n})$ values were given in [200, Example 1 and 2]. With analogous arguments it can be proved that the PWidth of the l_1 ball is $1/\sqrt{n}$. By Proposition (3.3.3), this implies that the angle condition can be lower bounded with $\tau = \Theta(1/\sqrt{n})$ for the unit simplex and the l_1 ball, and with $\tau = \Theta(1/n)$ for the unit cube.

Bounds using facial distance vf

For a polytope $\Omega = \{x \in \mathbb{R}^n \mid Ax \leq b\}$ with $A \in \mathbb{R}^{m \times n}$ the facial distance can be defined as (see [27]):

$$\mathrm{vf}(\Omega) = \min_{\substack{v \in V(\Omega) \\ i : (a^{(i)})^\top v < b_i}} \frac{b_i - (a^{(i)})^\top v}{\|a^{(i)}\|} . \tag{3.5.1}$$

It is the easy to bound $\mathrm{vf}(\Omega)$ on some specific class of polytopes and, consequently, give an explicit bound for the angle condition (see also [24]). For instance, if the matrix A is totally unimodular (i. e. all the vertices are integral for b integral), we have the following properties.

Proposition 3.5.1. *If the matrix A is totally unimodular and b is integral, then for $\bar{a} = \max_{i \in [1:m]} \|a_i\|$:*

- *for the AFW or the PFW, if the size of the active set stays bounded by $\bar{s}$, then*

$$\mathrm{SB}_{\mathrm{AFW}}(\Omega) \geq \frac{1}{2\bar{s}\bar{a}D}, \quad \mathrm{SB}_{\mathrm{PFW}}(\Omega) \geq \frac{1}{\bar{s}\bar{a}D}; \tag{3.5.2}$$

- *for the FDFW,*

$$\mathrm{SB}_{\mathrm{FD}}(\Omega) \geq \frac{1}{2D\bar{a}(\dim(\Omega) + 1)} \geq \frac{1}{2D\bar{a}(n+1)}. \tag{3.5.3}$$

Proof. If A is totally unimodular then for $i \in [1:m], v \in V$ such that $b_i - (a^{(i)})^\top v > 0$ we have

$$\frac{b_i - (a^{(i)})^\top v}{\|a_i\|} \geq \frac{1}{\|a_i\|} \tag{3.5.4}$$

since the numerator on the LHS must be at least one. By applying (3.5.4) to the RHS of (3.5.1) we obtain

$$\mathrm{vf}(\Omega) \geq \min_{i \in [1:m]} \frac{1}{\|a_i\|} = \frac{1}{\bar{a}} . \tag{3.5.5}$$

Then the thesis follows for the AFW and the PFW directly from the bounds of Remark 3.3.4. For the FDFW, the second part of (3.5.3) is trivially true since $\dim(\Omega) \leq n$, and the first follows by the bound given in Remark 3.3.4, using that by the Caratheodory theorem for every feasible point x there exists $S \in S_x$ with $|S| \leq \dim(\Omega) + 1$. □

The bound of Proposition 3.5.1 allows us to bound the angle condition for the min cost flow polytope with integral capacities:

$$\Omega = \{x \in \mathbb{R}^n \mid Ax \leq b,\ 0 \leq x \leq c\}\,, \tag{3.5.6}$$

with b, c integral and A incidence matrix of a directed graph G.

Corollary 3.5.2. *Consider a directed graph G with incidence matrix $A \in \mathbb{R}^{m\times n}$ and maximum degree of a vertex d. Then if Ω is given as in* (3.5.6)*:*

$$\mathrm{SB}_{\mathrm{FD}}(\Omega) \geq \frac{1}{2\sqrt{d}(n+1)\|c\|} \tag{3.5.7}$$

Proof. By the capacity constraints, the diameter of Ω is at most $\|c\|$. Then the result follows easily from Proposition 3.5.1 by noticing that Ω can be rewritten as $\{x \in \mathbb{R}^n \mid \tilde{A}x \leq b\}$ for $\tilde{A} = (A; I; -I)$ totally unimodular (see, e.g., [223]) with maximum norm of a row equal to $\sqrt{d}$. □

Bounds on sets with smooth boundary

On convex sets with smooth boundary the angle condition can be satisfied with constant arbitrarily close to 1 using orthographic retractions [208, Section 6.3]. Furthermore, on sublevel sets of smooth and strongly convex functions the FDFW satisfies the angle condition with constant equal to the condition number of the function divided by 2 [208, Section 6.2].

3.5.3 Applications

There is a number of practical optimization problems with the feasible sets and objectives discussed above. To start with, the LASSO problem, the minimum enclosing ball problem, training linear support vector machines and finding maximal cliques in graphs can all be formulated as convex quadratic optimization problems [48] on the l_1 ball or the simplex. The trust region subproblem is a non convex quadratic problem on the unit sphere (see [138]). The min cost flow problem with a quadratic objective is also of practical interest [220]. Many other examples can be found in [170].

3.6 Numerical tests

We tested the SSC on the AFW and the PFW methods, applied to a quadratic (non convex) relaxation of the maximum clique problem proposed in [40].

More precisely, let A be the adjacency matrix of a graph G. In [40] it is proved that there is a one to one correspondence between the maximal cliques of G and the local minima of the function $f : \Delta_{n-1} \to \mathbb{R}$ defined by

$$f(x) = -x^\top Ax - \frac{1}{2}\|x\|^2. \tag{3.6.1}$$

Therefore, we consider instances of Problem (3.2.1) with objective (3.6.1) and feasible set the $n-1$ dimensional unit simplex, that is $\Omega = \Delta_{n-1}$.
The graph instances we use are taken from the DIMACS benchmark [140]. To have a fair comparison for both the AFW and the PFW we use the stepsize given by

$$\alpha_k = \min\{\alpha_k^{\max}, -\frac{\nabla f(x_k)^\top d_k}{L\|d_k\|^2}\} \tag{3.6.2}$$

with $\alpha_k^{\max}$ determined by boundary conditions. In this way the new point computed by the methods coincides with the first point computed in the SSC procedure of their multistep versions.
We reported in Tables 3, 4 the results for the most challenging instances, aggregated on 100 runs starting from random points. The SSC clearly improves the CPU times while keeping the solution quality. Indeed in these problems the SSC allows the methods to identify the support of a local minimum in fewer iterations, so that the slow initial convergence phase is skipped (see Figures 3.3, 3.4).

Remark 3.6.1. While discussing the optimization of the SSC for specific problems is beyond the scope of this thesis, we remark that the method can still be useful even when both gradient updates and LMO are very cheap, as it is often the case with Frank Wolfe variants. For instance, in the case of quadratic problems on the simplex we deal with in this section, if the SSC does s AFW steps, the resulting point can be written as an affine combination of the starting point together with at most s vertices. The gradient updates can then be performed in parallel at once, as a matrix-vector multiplication where the vector has at most $s+1$ non zero components. Without SSC, such updates must be performed sequentially. Beside this, without SSC the objective value must be computed at every iteration rather than only at the end of the SSC.

Table 3: Max clique found, average clique size, standard deviation of clique sizes and average CPU time for AFW and SSC + AFW on max clique instances from the DIMACS benchmark.

	AFW				**SSC + AFW**			
Instance	Max	Mean	Std	CPU time	Max	Mean	Std	CPU time
C2000.5	14	11.7	0.89	2.800	14	11.6	1.00	0.082
C2000.9	67	60.2	2.20	3.135	65	60.0	2.05	0.200
C4000.5	16	12.8	0.94	23.487	16	12.5	0.92	0.429
MANN_a81	1080	1080.0	0.00	31.156	1080	1080.0	0.00	25.047
keller6	45	38.4	2.41	13.713	43	37.8	2.22	0.413

Table 4: Max clique found, average clique size, standard deviation of clique sizes and average CPU time for PFW and SSC + PFW on max clique instances from the DIMACS benchmark.

	PFW				**SSC + PFW**			
Instance	Max	Mean	Std	CPU time	Max	Mean	Std	CPU time
C2000.5	14	11.8	0.86	2.811	14	12.1	0.86	0.077
C2000.9	67	62.3	1.83	3.031	68	62.0	1.77	0.150
C4000.5	15	12.7	0.92	23.423	16	13.4	0.95	0.379
MANN_a81	1080	1080.0	0.00	19.867	1080	1080.0	0.00	15.442
keller6	44	37.3	2.68	13.515	45	35.6	2.83	0.258

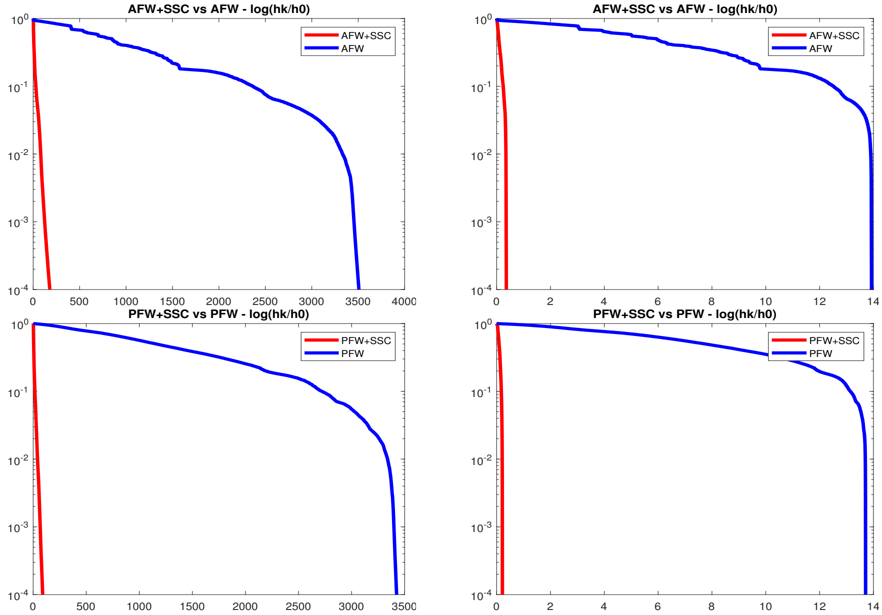


Figure 3.3: Iteration number and CPU time vs $\log(h_k/h_0)$ in the first and the second column respectively for the instance keller6

Data availability. The data analysed during the current study are available in the 2nd DIMACS implementation challenge repository,
`http://archive.dimacs.rutgers.edu/pub/challenge/graph/benchmarks/clique/`

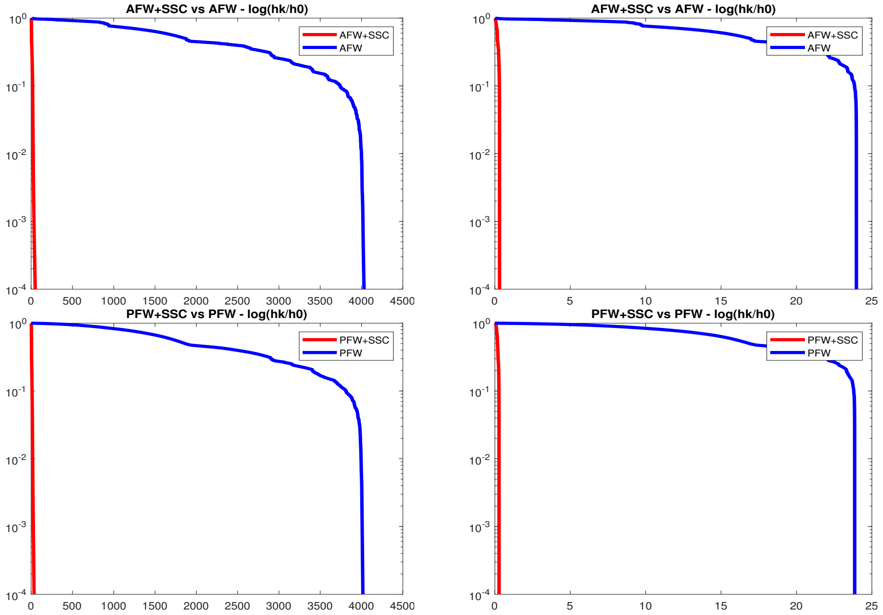


Figure 3.4: Iteration number and CPU time vs $\log(h_k/h_0)$ in the first and the second column respectively for the instance C4000.5

Chapter 4

Active Set Identification properties of the Away-Step Frank–Wolfe Algorithm

In this chapter, we study active set identification results for the AFW in different settings. We first prove a local identification property that we apply, in combination with a convergence hypothesis, to get an active set identification result. We then prove, in the nonconvex case, a novel $O(1/\sqrt{k})$ convergence rate result and active set identification for different step sizes (under suitable assumptions on the set of stationary points). By exploiting those results, we also give explicit active set complexity bounds for both strongly convex and nonconvex objectives. While we initially consider the probability simplex as feasible set, we subsequently show how to adapt some of our results to generic polytopes. [1]

4.1 Active set identification and FW variants

Identifying a surface containing a solution (and/or the support of sparse solutions) represents a relevant task in optimization, since it allows to reduce the dimension of the problem at hand and to apply a more sophisticated method in the end (see, e.g., [29, 33, 83, 88, 118–120]). This is the reason why, in the last decades, identification properties of optimization methods have been the subject of extensive

[1] This chapter is based on "Active Set Complexity of the Away-Step Frank–Wolfe Algorithm" in *SIAM Journal on Optimization, vol. 30, iss. 3, pp. 2470-2500, 2020* [48].

studies.

Beside its slow convergence rate discussed in Chapter 2, the classic FW approach has another relevant drawback with respect to other algorithms: even when dealing with the simplest polytopes, it cannot identify the active set in finite time (see, e.g., [46]). Due to the renewed interest in the method, it has hence become a relevant issue to determine whether some FW variants admit active set identification properties similar to those of other first order methods. In this chapter we focus on the AFW and analyze active set identification properties for problems of the form

$$\min \{f(x) \mid x \in \Delta_{n-1}\},$$

where the objective f is a differentiable function with Lipschitz regular gradient and the feasible set is the probability simplex. When the algorithm converges to a stationary point x^* we say that it identifies the active set if it correctly determines all the binding constraints. The active set complexity is then defined as the number of iterations after which every sequence generated by the algorithm identifies this subset of constraints. In the chapter, we extend this active set complexity definition to include sequences convergent to certain subsets of stationary points. We also extend some of the active set complexity results to general polytopes.

4.1.1 Contributions

It is a classic result that on polytopes and under strict complementarity conditions the AFW with exact line search identifies the face containing the minimum in finite time for strongly convex objectives [116]. More general active set identification properties for Frank-Wolfe variants have recently been analyzed in [46], where the authors proved active set identification for sequences convergent to a stationary point, and AFW convergence to a stationary point for C^2 objectives with a finite number of stationary points and satisfying a technical convexity-concavity assumption (this assumption is essentially a generalization of a property related to quadratic possibly indefinite functions). The main contributions of this chapter with respect to [46] are twofold:

- First, we give quantitative local and global active set identification complexity bounds under suitable assumptions on the objective. The key element in the computation of those bounds is a quantity that we call "active set radius".

This radius determines a neighborhood of a stationary point for which the AFW at each iteration identifies an active constraint (if there is any not yet identified one). In particular, to get the active set complexity bound it is sufficient to know how many iterations it takes for the AFW sequence to enter this neighborhood.

- Second, we analyze the identification properties of AFW without the technical convexity-concavity $\boldsymbol{C}^2$ assumption used in [46] (we consider general nonconvex objectives with Lipschitz gradient instead). More specifically, we prove active set identification under different conditions on the step size and some additional hypotheses on the support of stationary points.

In order to prove our results, we consider step sizes dependent on the Lipschitz constant of the gradient (see, e.g., [22], [134] and references therein). By exploiting the affine invariance property of the AFW (see, e.g., [136]), we also extend some of the results to generic polytopes. In our analysis we see how the AFW identification properties are related to the value of Lagrangian multipliers on stationary points. This, to the best of our knowledge, is the first time that some active set complexity bounds are given for a variant of the FW algorithm.

The chapter is organized as follows: after presenting the AFW method for optimization on the simplex and some preliminaries in Section 4.2, we study the local behaviour of this algorithm regarding the active set in Section 4.3. In Section 4.4 we provide active set identification results in a quite general context, and apply these to the strongly convex case for obtaining complexity bounds. Section 4.5 treats the nonconvex case, giving both global and local active set complexity bounds. Finally, in Section 4.6 we extend some of our results to generic polytopes.

4.1.2 Related work

In [60] the authors proved that the projected gradient method and other converging sequential quadratic programming methods identify quasi-polyhedral faces under some nondegeneracy conditions. In [61] those results were extended to the case of exposed faces in polyhedral sets without the nondegeneracy assumptions. This extension is particularly relevant to our work since the identification of exposed faces in polyhedral sets is the framework that we use in studying the AFW on polytopes. In [240] the results of [60] were generalized to certain nonpolyhedral surfaces called "C^p identifiable" contained in the boundary of convex sets. A key insight in these early works was the openness of a generalized normal cone defined for the identifiable surface containing a nondegenerate stationary point. This openness guarantees that, in a neighborhood of the stationary point, the projection of the gradient identifies the related surface. It turns out that for linearly constrained sets the generalized normal cone is related to positive Lagrangian multipliers on the stationary point.

A generalization of [60] to nonconvex sets was proved in [62], while an extension to nonsmooth objectives was first proved in [123]. Active set identification results have also been proved for a variety of projected gradient, proximal gradient and stochastic gradient related methods (see for instance [218] and references therein).

Recently, explicit active set complexity bounds have been given for some of the methods listed above. Bounds for proximal gradient and block coordinate descent method were analyzed in [196] and [195] under strong convexity assumptions on the objective. A more systematic analysis covering many gradient related proximal methods (like, e.g., accelerated gradient, quasi Newton and stochastic gradient proximal methods) was carried out in [218].

As for FW-like methods, in addition to the results in [116] and [46] discussed earlier, identification results have been proved in [78] for fully corrective variants on the probability simplex. However, since fully corrective variants require computing the minimum of the objective on a given face at each iteration, they are not suited for nonconvex problems.

4.2 Preliminaries

In this chapter, $f : \Delta_{n-1} \to \mathbb{R}$ is a function with gradient having Lipschitz constant L. The constant L is also used as Lipschitz constant for ∇f with respect to the norm $\|\cdot\|_1$. This does not require any additional hypothesis on f since $\|\cdot\|_1 \geq \|\cdot\|$, so that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \leq L\|x - y\|_1$$

for every $x, y \in \Delta_{n-1}$. $\mathcal{X}^*$ is the set of points satisfying first order optimality conditions for the minimization of f on Δ_{n-1}, that is $\nabla f(x)^\top d \geq 0$ for every d feasible direction in x. We call $\mathcal{X}^*$ the set of stationary points (see, e.g., [30]).

We define dist_1 in the same way of the Euclidean distance dist but with respect to $\|\cdot\|_1$ instead of $\|\cdot\|$. We now introduce the multiplier functions, which were recently used in [83] to define an active set strategy for minimization over the probability simplex.

For every $x \in \Delta_{n-1}$, $i \in [1:n]$ the multiplier function $\lambda_i : \Delta_{n-1} \to \mathbb{R}$ is defined as

$$\lambda_i(x) = \nabla f(x)^\top (e_i - x),$$

or in vector form

$$\lambda(x) = \nabla f(x) - x^\top \nabla f(x) e \ .$$

For every $x \in \mathcal{X}^*$ these functions coincide with the Lagrangian multipliers of the constraints $x_i \geq 0$.

We define the *the extended support* in $x \in \mathcal{X}^*$ as

$$I(x) = \{i \in [1:n] \mid \lambda_i(x) = 0\} \,,$$

and with $I^c(x) = \{1, ...n\} \setminus I(x)$ the set of binding constraints in x. By first order optimality conditions (for minimization) we have $\lambda_i(x) \geq 0$ for every $i \in [1:n]$ and therefore

$$\lambda_i(x) > 0 \ \forall \ i \in I^c(x) \,.$$

We use the notation $a_k \to A$ for the convergence of a sequence $\{a_k\}$ to the set A as equivalent to $\text{dist}(a_k, A) \to 0$.

Keeping in mind that

$$\Delta_{n-1} = \text{conv}(\{e_i, \ i = 1, \dots, n\}),$$

we can assume that $\text{LMO}_{\Delta_{n-1}}(r)$ always returns a vertex of the probability simplex, that is

$$\text{LMO}_{\Delta_{n-1}}(r) = e_{\hat{\imath}}$$

with $\hat{\imath} \in \arg\min_i r_i$.

4.2.1 FW and AFW on the probability simplex

Algorithm 1 is the classical FW method on the probability simplex. At each iteration, this first order method generates a descent direction that points from the current iterate x_k to a vertex s_k minimizing the scalar product with the gradient, and then moves along this search direction of a suitable step size if stationarity conditions are not satisfied.

Algorithm 5 Frank–Wolfe method on the probability simplex

```
Initialize $x_0 \in \Delta_{n-1}$, $k := 0$
Set $s_k := e_{\hat{\imath}}$, with $\hat{\imath} \in \arg\min_i \nabla_i f(x_k)$ and $d_k^{\mathcal{FW}} := s_k - x_k$
if $x_k$ is stationary then
    STOP
end if
Choose the step size $\alpha_k \in (0, 1]$ with a suitable criterion
Update: $x_{k+1} := x_k + \alpha_k d_k^{\mathcal{FW}}$
Set $k := k + 1$. Go to Step 2
```

The away step variant for the unit simplex is instead reported in Algorithm 2. When the AFW performs an away step, we have that either the support of the current iterate stays the same or decreases of one (we get rid of the component whose index is associated to the away direction in case $\alpha_k = \alpha_k^{\max}$). On the other hand, when the algorithm performs a Frank Wolfe step, only the vertex given by the LMO is eventually added to the support of the current iterate. These two properties are fundamental for the active set identification of the AFW.

Algorithm 6 Away–step Frank–Wolfe on the probability simplex

```
$x_0 \in \Delta_{n-1}$, $k := 0$
Set $s_k := e_{\hat{\imath}}$, with $\hat{\imath} \in \arg\min_i \nabla_i f(x_k)$ and $d_k^{\mathcal{FW}} := s_k - x_k$
if $x_k$ is stationary then
    STOP
end if
Let $v_k := e_{\hat{\jmath}}$, with $\hat{\jmath} \in \arg\max_{j \in S_k} \nabla_j f(x_k)$, $S_k := \{j : (x_k)_j > 0\}$ and $d_k^{\mathcal{A}} := x_k - v_k$
if $-\nabla f(x_k)^\top d_k^{\mathcal{FW}} \geq -\nabla f(x_k)^\top d_k^{\mathcal{A}}$ then
    $d_k := d_k^{\mathcal{FW}}$, and $\alpha_k^{\max} := 1$
else
    $d_k := d_k^{\mathcal{A}}$, and $\alpha_k^{\max} := (x_k)_i/(1 - (x_k)_i)$
end if
Choose the step size $\alpha_k \in (0, \alpha_k^{\max}]$ with a suitable criterion
Update: $x_{k+1} := x_k + \alpha_k d_k$
$k := k + 1$. Go to step 2.
```

4.2.2 Technical results related to step sizes

In order to obtain convergence results we of course need some lower bound on the step size. In particular, we lower bound α_k with the Lipschitz constant dependent step size $\bar{\alpha}_k$ introduced in Section 2.5:

$$\bar{\alpha}_k = \min\left(\alpha_k^{\max}, \frac{-\nabla f(x_k)^\top d_k}{L\|d_k\|^2}\right) , \tag{4.2.1}$$

We now prove several properties related to the step size given in (4.2.1). First, we prove that it is always a lower bound on the step size obtained by the exact line search. We then prove that

$$\alpha_k \geq \min(\alpha_k^{\max}, c\frac{p_k}{L\|d_k\|^2}) \text{ for some } c > 0,$$

for the Armijo line search and if we impose the weak Wolfe conditions, setting $\alpha_k = \alpha_k^{\max}$ whenever they cannot be satisfied. When $c \geq 1$ then (4.2.1) is of course a lower bound for the step size α_k, and when $c < 1$ we can still recover (4.2.1) by considering $\tilde{L} = \frac{L}{c}$ instead of L as Lipschitz constant.

Lemma 4.2.1. *Consider a sequence* $\{x_k\}$ *in* Δ_{n-1} *such that* $x_{k+1} = x_k + \alpha_k d_k$ *with* $\alpha_k \in \mathbb{R}_{\geq 0}$, $d_k \in \mathbb{R}^n$. *Let* $\bar{\alpha}_k$ *be defined as in* (4.2.1), *let* $p_k = -\nabla f(x_k)^\top d_k$ *and assume* $p_k > 0$. *Then:*

1. *If* $0 \leq \alpha_k \leq 2p_k/(\|d_k\|^2 L)$, *the sequence* $\{x_k\}$ *has the property* (4.5.33).
2. *If* $\alpha_k = \bar{\alpha}_k$ *then* (4.5.3) *is satisfied with* $\rho = \frac{1}{2}$. *Additionally, we have*

$$f(x_k) - f(x_{k+1}) \geq L\frac{\|x_{k+1} - x_k\|^2}{2} . \tag{4.2.2}$$

3. *If* α_k *is given by exact line search, then* $\alpha_k \geq \bar{\alpha}_k$ *and* (4.5.3) *is again satisfied with* $\rho = \frac{1}{2}$.

If $\alpha_k \leq \alpha_k^{\max}$ the condition of point 1 implies $0 \leq \alpha_k \leq 2\bar{\alpha}_k$.

Proof. By the standard descent lemma [31, Proposition 6.1.2] we have

$$f(x_k) - f(x_k + \alpha d_k) \geq \alpha p_k - \alpha^2\frac{L\|d_k\|^2}{2} . \tag{4.2.3}$$

It is immediate to check

$$\alpha\nabla f(x_k)^\top d_k + \alpha^2\frac{L\|d_k\|^2}{2} \leq 0 , \tag{4.2.4}$$

for every $0 \leq \alpha \leq \frac{2p_k}{L\|d_k\|^2}$.

$$\alpha p_k - \alpha^2 \frac{L\|d_k\|^2}{2} \geq \alpha p_k/2 \geq \alpha^2 \frac{L\|d_k\|^2}{2} \tag{4.2.5}$$

for every $0 \leq \alpha \leq \frac{p_k}{L\|d_k\|^2}$.

1. For every $x \in \operatorname{conv}(x_k, x_{k+1}) \subseteq \left\{x_k + \alpha d_k \mid 0 \leq \alpha \leq \frac{2p_k}{L\|d_k\|^2}\right\}$, we have

$$f(x) = f(x_k + \alpha d_k) \leq f(x_k) + \alpha \nabla f(x_k)^\top d_k + \alpha^2 \frac{L\|d_k\|^2}{2} \leq f(x_k) \ ,$$

where we used (4.2.3) in the first inequality and (4.2.4) in the second inequality.
2. We have

$$f(x_k) - f(x_{k+1}) = f(x_k) - f(x_k + \bar{\alpha}_k d_k) \geq \bar{\alpha}_k p_k/2 \ ,$$

where we have the hypotheses to apply (4.2.5) since $0 \leq \bar{\alpha}_k \leq \frac{p_k}{L\|d_k\|^2}$. Again by (4.2.5)

$$f(x_k) - f(x_{k+1}) = f(x_k) - f(x_k + \bar{\alpha}_k d_k) \geq \bar{\alpha}_k^2 \frac{L\|d_k\|^2}{2} = L\frac{\|x_k - x_{k+1}\|^2}{2} \ .$$

3. If $\alpha_k = \alpha_k^{\max}$ then there is nothing to prove since $\bar{\alpha}_k \leq \alpha_k^{\max}$. Otherwise we have

$$0 = \frac{\partial}{\partial \alpha} f(x_k + \alpha d_k)|_{\alpha = \alpha_k} = d_k^\top (\nabla f(x_k + \alpha_k d_k)) \tag{4.2.6}$$

and therefore

$$\begin{aligned} -d_k^\top \nabla f(x_k) &= -d_k^\top \nabla f(x_k) + d_k^\top \nabla f(x_k + \alpha_k d_k) = -d_k^\top (\nabla f(x_k) - \nabla f(x_k + \alpha_k d_k)) \\ &\leq L\|d_k\|\|x_k - (x_k + \alpha_k d_k)\| = \alpha_k L \|d_k\|^2 \ , \end{aligned} \tag{4.2.7}$$

where we used (4.2.6) in the first equality and the Lipschitz condition in the inequality. From (4.2.7) it follows

$$\alpha_k \geq \frac{-d_k^\top \nabla f(x_k)}{L\|d_k\|^2} \geq \bar{\alpha}_k$$

and this proves the first claim. As for the second,

$$f(x_k) - f(x_k + \alpha_k d_k) \geq f(x_k) - f(x_k + \bar{\alpha}_k d_k) \geq \frac{\bar{\alpha}_k}{2} p_k \ ,$$

where the first inequality follows from the definition of exact line search and the second by point 2 of the lemma. □

Corollary 4.2.2. *Under the hypotheses of Lemma 4.2.1, assume that $f(x_k)$ is monotonically decreasing and assume that for some subsequence $k(j)$ we have $x_{k(j)+1} = x_{k(j)} + \bar{\alpha}_{k(j)} d_{k(j)}$. Then*

$$\|x_{k(j)} - x_{k(j)+1}\| \to 0 \ .$$

Proof. By (4.2.2) we have

$$f(x_{k(j)}) - f(x_{k(j)+1}) \geq \frac{L}{2} \|x_{k(j)} - x_{k(j)+1}\|^2$$

and the conclusion follows by monotonicity and boundedness. □

We now briefly recall the Armijo line search and the Wolfe conditions with a couple of adaptations to our setting. For the Armijo search we impose the usual condition of sufficient decrease

$$f(x_k) - f(x_k + \alpha_k d_k) \geq c_1 \alpha_k p_k \tag{4.2.8}$$

and assume that the tentative step sizes are given by $\beta_k^{(0)} = \alpha_k^{\max}$, $\beta_k^{(j+1)} = \gamma \beta_k^{(j)}$ for $c_1, \gamma \in (0,1)$.

Lemma 4.2.3. *If α_k is determined by the Armijo line search described above then*

$$\alpha_k \geq \min(\alpha_k^{\max}, 2\gamma(1-c_1)\frac{p_k}{L\|d_k\|^2}) \geq \min\{1, 2\gamma(1-c_1)\}\bar{\alpha}_k \tag{4.2.9}$$

with $\bar{\alpha}_k = \min(\alpha_k^{\max}, \frac{p_k}{L\|d_k\|^2})$ as in (4.2.1)*, and* (4.5.3) *holds with $\rho = c_1 \min\{1, 2\gamma(1-c_1)\} < 1$.*

Proof. From the upper bound on f given in (4.2.3) it follows

$$f(x_k) - f(x_k + \alpha d_k) \geq c_1 \alpha p_k \quad \text{for } \alpha \in [0, 2(1-c_1)\frac{p_k}{L\|d_k\|^2}] \tag{4.2.10}$$

and

$$\alpha_k > 2\gamma(1-c_1)\frac{p_k}{L\|d_k\|^2}.$$

Therefore

$$\alpha_k \geq \min(\alpha_k^{\max}, 2\gamma(1-c_1)\frac{p_k}{L\|d_k\|^2}) \geq \min\{1, 2\gamma(1-c_1)\}\bar{\alpha}_k, \tag{4.2.11}$$

which proves (4.2.9). We also have

$$f(x_k) - f(x_k + \alpha_k d_k) \geq c_1 \alpha_k p_k \geq c_1 \min\{1, 2\gamma(1-c_1)\}\bar{\alpha}_k p_k, \tag{4.2.12}$$

where we used the Armijo condition (4.2.8) in the first inequality and (4.2.9) in the second. Hence, by $c_1, \gamma \in (0,1)$ and $c_1(1-c_1) \leq \frac{1}{4}$, we get that equation (4.5.3) holds with $\rho = c_1 \min\{1, 2\gamma(1-c_1)\} < 1$. □

The weak Wolfe conditions [194] are (4.2.8) together with

$$-d_k^\top \nabla f(x_k + \alpha_k d_k) \le c_2 p_k \tag{4.2.13}$$

for some $c_2 \in (c_1, 1)$.

Lemma 4.2.4. *Assume $\alpha_k = \min(\alpha_k^{\max}, \tilde{\alpha}_k)$ with $\tilde{\alpha}_k$ satisfying the weak Wolfe conditions. Then*

$$\alpha_k \ge \min(\alpha_k^{\max}, (1-c_2)\frac{p_k}{L\|d_k\|^2}) \ge (1-c_2)\bar{\alpha}_k \tag{4.2.14}$$

and (4.5.3) *holds with $\rho = c_1(1-c_2) < 1$.*

Proof. **Case a)**: $\alpha_k = \alpha_k^{\max}$. Then trivially $\alpha_k \ge \bar{\alpha}_k$ and by point 2 of Lemma 4.2.1, equation (4.5.3) is satisfied with $\rho = \frac{1}{2}$.
Case b): the second weak Wolfe condition holds. We have

$$c_2 p_k \ge -d_k^\top \nabla f(x_k+\alpha_k d_k) = d_k^\top(-\nabla f(x_k)+(\nabla f(x_k)-\nabla f(x_k+\alpha_k d_k))) \ge p_k - \alpha_k L\|d_k\|^2 \tag{4.2.15}$$

where we used (4.2.13) in the first inequality. Rearranging (4.2.15) we obtain

$$\alpha_k \ge \frac{(1-c_2)p_k}{L\|d_k\|^2}\,. \tag{4.2.16}$$

As for part 1 we can now use the Armijo condition (4.2.8) to obtain (4.5.3) with $\rho = c_1(1-c_2)$:

$$f(x_k) - f(x_k + \alpha_k d_k) \ge c_1 \alpha_k p_k \ge c_1(1-c_2)\bar{\alpha}_k p_k, \tag{4.2.17}$$

where we used (4.2.16) in the second inequality. To conclude, since $\frac{1}{4} \ge c_1(1-c_1) > c_1(1-c_2)$ for $0 < c_1 < c_2 < 1$, the bound (4.5.3) holds in both cases with $\rho = c_1(1-c_2)$. □

4.2.3 Elementary inequalities

In several proofs we need some elementary inequalities concerning the euclidean norm $\|\cdot\|$ and the norm $\|\cdot\|_1$.

Lemma 4.2.5. *Given $\{x, y\} \subset \Delta_{n-1}$, $i \in [1:n]$ we have that*

1. *$\|e_i - x\| \le \sqrt{2}(e_i - x)_i$ holds; that*

2. *$(y-x)_i \le \|y-x\|_1/2$ holds; and*

3. if $\{x_k\}$ is a sequence generated on the probability simplex by the AFW then $\|x_{k+1} - x_k\|_1 \leq 2\|x_{k+1} - x_k\|$ for every k.

Proof. 1. $(e_i - x)_j = -x_j$ for $j \neq i$, $(e_i - x)_i = 1 - x_i = \sum_{j \neq i} x_j$. In particular

$$\|e_i - x\| = (\sum_{j \neq i} x_j^2 + (e_i - x)_i^2)^{\frac{1}{2}} \leq ((\sum_{j \neq i} x_j)^2 + (1 - x_i)^2)^{\frac{1}{2}} = \sqrt{2}(\sum_{j \neq i} x_j) = \sqrt{2}(e_i - x)_i$$

2. Since $\sum_{j \in [1:n]} x_j = \sum_{j \in [1:n]} y_j$ so that $\sum (x - y)_j = 0$ we have

$$(y - x)_i = \sum_{j \neq i} (x - y)_j$$

and as a consequence

$$\|y - x\|_1 = \sum_{j \in [1:n]} |(y - x)_j| \geq (y - x)_i + \sum_{j \neq i} (x - y)_j = 2(y - x)_i \ .$$

3. We have $x_{k+1} - x_k = \alpha_k d_k$ with $d_k = \pm(e_i - x_k)$ for some $i \in [1:n]$. By homogeneity it suffices to prove $\|d_k\| \geq \frac{1}{2}\|d_k\|_1$. We have

$$\|d_k\| \geq 1 - (x_k)_i = \frac{1}{2}(1 - (x_k)_i + \sum_{j \neq i} (x_k)_j) = \frac{1}{2}\|d_k\|_1 \ ,$$

where in the first equality we used $\sum_{i=1}^n (x_k)_i = 1$ (so that $1 - (x_k)_i = \sum_{j \neq i} (x_k)_j$) and in the second equality we used $0 \leq x_k \leq 1$. □

4.3 Local active set variables identification property of the AFW

In this section we prove a rather technical proposition which is the key tool to give quantitative estimates for the active set complexity. It states that when the sequence is close enough to a fixed stationary point at every step the AFW identifies one variable violating the complementarity conditions with respect to the multiplier functions on this stationary point (if it exists), and it sets the variable to 0 with an away step. The main difficulty is giving a tight estimate for how close the sequence must be to a stationary point for this identifying away step to take place.

A lower bound on the size of the nonmaximal away steps is needed in the following theorem, since otherwise for steps small enough the sequence can stay arbitrarily close to the starting point.

Let $\{x_k\}$ be the sequence of points generated by the AFW, and let x^* be a fixed point in $\mathcal{X}^*$. Since x^* does not vary in this section, we write for simplicity I and I^c instead of $I(x^*)$ and $I^c(x^*)$, respectively, in the rest of this section.
Note that by complementary slackness we have $x^*_j = 0$ for all $j \in I^c$.

Before proving the main theorem we need to prove the following lemma to bound the Lipschitz constant of the multipliers on stationary points.

Lemma 4.3.1. *Given $h > 0$, $x_k \in \Delta_{n-1}$ such that $\|x_k - x^*\|_1 \leq h$ let*

$$O_k = \{i \in I^c \mid (x_k)_i = 0\}$$

and assume that $O_k \neq I^c$. Let $\delta_k = \max_{i \in [1:n] \setminus O_k} \lambda_i(x^*)$. For every $i \in \{1, ..., n\}$:

$$|\lambda_i(x^*) - \lambda_i(x_k)| \leq h(L + \frac{\delta_k}{2}) \ . \tag{4.3.1}$$

Proof. By considering the definition of $\lambda(x)$, we can write

$$\begin{aligned}|\lambda_i(x_k) - \lambda_i(x^*)| &= |\nabla f(x_k)_i - \nabla f(x^*)_i + \nabla f(x^*)^\top (x^* - x_k) + (\nabla f(x^*) - \nabla f(x_k))^\top x_k| \\ &\leq |\nabla f(x^*)_i - \nabla f(x_k)_i + (\nabla f(x_k) - \nabla f(x^*))^\top x_k| + |\nabla f(x^*)^\top (x^* - x_k)| \ .\end{aligned} \tag{4.3.2}$$

By taking into account the fact that $x_k \in \Delta_{n-1}$ and gradient of f is Lipschitz continuous, we have

$$\begin{aligned}&|\nabla f(x_k)_i - \nabla f(x^*)_i + (\nabla f(x^*) - \nabla f(x_k))^\top x_k| = |(\nabla f(x^*) - \nabla f(x_k))^\top (x_k - e_i)| \\ &\leq \|\nabla f(x^*) - \nabla f(x_k)\|_1 \|x_k - e_i\|_\infty \leq Lh,\end{aligned} \tag{4.3.3}$$

where the last inequality is justified by the Hölder inequality with exponents $1, \infty$. We now bound the second term in the right-hand side of (4.3.2). Let

$$u_j = \max\{0, (x^* - x_k)_j\}, \ l_j = \max\{0, -(x^* - x_k)_j\} \ .$$

We have $\sum_{j \in [1:n]} x^*_j = \sum_{j \in [1:n]} (x_k)_j = 1$ since $\{x^*, x_k\} \subset\in \Delta_{n-1}$, so that

$$\sum_{j \in [1:n]} (x^* - x_k)_j = \sum_{j \in [1:n]} (u_j - l_j) = 0 \quad \text{and hence} \quad \sum_{j \in [1:n]} u_j = \sum_{i \in [1:n]} l_j.$$

Moreover, $h' \overset{\text{def}}{=} 2 \sum_{j \in [1:n]} u_j = 2 \sum_{j \in [1:n]} l_j = \sum_{j \in [1:n]} u_j + l_j = \sum_{j \in [1:n]} |x^*_j - (x_k)_j| \leq h$, hence

$$h'/2 = \sum_{j \in [1:n]} u_j = \sum_{j \in [1:n]} l_j \leq h/2 \ .$$

We can finally bound the second piece of (4.3.2), using $u_j = l_j = 0$ for all $j \in O_k$ (because $(x_k)_j = x^*_j = 0$):

$$\begin{aligned} |\nabla f(x^*)^\top (x^* - x_k)| &= |\nabla f(x^*)^\top u - \nabla f(x^*)^\top l| \leq \frac{h'}{2}(\nabla f(x^*)_M - \nabla f(x^*)_m) \\ &\leq \frac{h}{2}(\nabla f(x^*)_M - \nabla f(x^*)_m), \qquad (4.3.4) \end{aligned}$$

where $\nabla f(x_k)_M$ and $\nabla f(x_k)_m$ are respectively the maximum and minimum component of the gradient in $[1 : n] \setminus O_k$.
Now, considering inequalities (4.3.2), (4.3.3) and (4.3.4), we can write

$$|\lambda_i(x_k) - \lambda_i(x^*)| \leq Lh + \frac{h}{2}(\nabla f(x^*)_M - \nabla f(x^*)_m).$$

By taking into account the definition of δ_k and the fact that $\lambda(x^*)_j \geq 0$ for all j, we can write

$$\delta_k = \max_{i,j \in [1:n] \setminus O_k} (\nabla f(x^*)_i - \nabla f(x^*)_j) \geq \nabla f(x^*)_M - \nabla f(x^*)_m.$$

We can finally write

$$|\lambda_i(x_k) - \lambda_i(x^*)| \leq h(L + \frac{\delta_k}{2}),$$

thus concluding the proof. □

We now show a few simple but important results that connect the multipliers and the directions selected by the AFW algorithm. For a fixed x_k the multipliers $\lambda_i(x_k)$ are the values of the linear function $x \mapsto \nabla f(x_k)^\top x$ on the vertices of Δ_{n-1} (up to a constant), which in turn are the values considered in the AFW to select the direction. This basic observation is essentially everything we need for the next results.

Lemma 4.3.2. *Using the notation introduced in Algorithm 2, we have:*

(a) If $\max\{\lambda_i(x_k) \mid i \in S_k\} > \max\{-\lambda_i(x_k) \mid i \in [1 : n]\}$, *then the AFW performs an away step with* $d_k = d_k^{\mathcal{A}} = x_k - e_{\hat{\imath}}$ *for some* $i \in \text{argmax}\{\lambda_i(x_k) \mid i \in S_k\}$.

(b) For every $i \in [1 : n] \setminus S_k$ *if* $\lambda_i(x_k) > 0$ *then* $(x_{k+1})_i = (x_k)_i = 0$.

Proof. (a) By the definition of the away direction $d_k^{\mathcal{A}}$ it follows

$$d_k^{\mathcal{A}} \in \text{argmax}\{-\nabla f(x_k)^\top d \mid d = x_k - e_i, i \in S_k\}$$

which implies

$$d_k^{\mathcal{A}} = x_k - e_{\hat{\imath}} \quad \text{for some } \hat{\imath} \in \operatorname{argmax}\{-\nabla f(x_k)^\top (x_k - e_i) \mid i \in S_k\} = \operatorname{argmax}\{\lambda_i(x_k) \mid i \in S_k\} \,. \tag{4.3.5}$$

As a consequence of (4.3.5)

$$-\nabla f(x_k)^\top d_k^{\mathcal{A}} = \max\{-\nabla f(x_k)^\top d \mid d = x_k - e_i, i \in S_k\} = \max\{\lambda_i(x_k) \mid i \in S_k\} \,, \tag{4.3.6}$$

where the second equality follows from $\lambda_i(x_k) = -\nabla f(x_k)^\top d$ with $d = x_k - e_i$. Analogously

$$\begin{aligned} -\nabla f(x_k)^\top d_k^{\mathcal{FW}} &= \max\{-\nabla f(x_k)^\top d \mid d = e_i - x_k, i \in \{1, \ldots n\}\} \\ &= \max\{-\lambda_i(x_k) \mid i \in \{1, \ldots n\}\} \,. \end{aligned} \tag{4.3.7}$$

We can now prove that $-\nabla f(x_k)^\top d_k^{\mathcal{FW}} < -\nabla f(x_k)^\top d_k^{\mathcal{A}}$, so that the away direction is selected under assumption (a):

$$\begin{aligned} &-\nabla f(x_k)^\top d_k^{\mathcal{FW}} = \max\{-\lambda_i(x_k) \mid i \in \{1, \ldots n\}\} \\ &< \max\{\lambda_i(x_k) \mid i \in S_k\} = -\nabla f(x_k)^\top d_k^{\mathcal{A}}, \end{aligned}$$

where we used (4.3.6) and (4.3.7) for the first and the second equality respectively, and the inequality is true by hypothesis.

(b) By considering the fact that $(x_k)_i = 0$, we surely cannot choose the vertex e_i to define the away-step direction. Furthermore, since $\lambda(x_k)_i = \nabla f(x_k)^\top (e_i - x_k) > 0$, direction $d = e_i - x_k$ cannot be chosen as the Frank-Wolfe direction at step k as well. This guarantees that $(x_{k+1})_i = 0$. □

We can now prove the main theorem. The strategy is to split $[1:n]$ in three subsets I, $J_k \subset I^c$ and $O_k = I^c \setminus J_k$ and use Lemma 4.3.1 to control the variation of the multiplier functions on each of these three subsets. In the proof we examine two possible cases under the assumption of being close enough to a stationary point. If $J_k = \emptyset$, which means that the current iteration of the AFW has identified the support of the stationary point, then we show that the AFW chooses a direction contained in the support, so that also $J_{k+1} = \emptyset$.

If $J_k \neq \emptyset$, we show that in the neighborhood claimed by the theorem the largest multiplier in absolute value is always positive, with index in J_k, and big enough, so that the corresponding away step is maximal. This means that the AFW at the iteration $k+1$ identifies a new active variable.

Theorem 4.3.3. *If I^c is not the empty set, let us define*

$$\delta_{\min} = \min\{\lambda_i(x^*) \mid i \in I^c\} > 0, \;\; J_k = \{i \in I^c \mid (x_k)_i > 0\} \; .$$

Assume that for every k such that $d_k = d_k^{\mathcal{A}}$ the step size α_k is either maximal with respect to the boundary condition (that is $\alpha_k = \alpha_k^{\max}$) or $\alpha_k \geq \frac{-\nabla f(x_k)^\top d_k}{L\|d_k\|^2}$. If $\|x_k - x^*\|_1 < \frac{\delta_{\min}}{\delta_{\min}+2L} = r_*$ then

$$|J_{k+1}| \leq \max\{0, |J_k| - 1\} \; . \tag{4.3.8}$$

The latter relation also holds in case $I^c = \emptyset$ whence we put $r_* = +\infty$.

Proof. If $I^c = \emptyset$, or equivalently, if $\lambda(x^*) = 0$, then there is nothing to prove since $J_k \subset I^c = \emptyset \Rightarrow |J_k| = |J_{k+1}| = 0$.
So assume $I^c \neq \emptyset$. Recall that $\lambda_i(x^*) > 0$ for every $i \in I^c$, so that necessarily $\delta_{\min} > 0$.
For every $i \in [1 : n]$, by Lemma 4.3.1

$$\begin{aligned} \lambda_i(x_k) &\geq \lambda_i(x^*) - \|x_k - x^*\|_1 (L + \frac{\delta_k}{2}) \\ &> \lambda_i(x^*) - r_*(L + \frac{\delta_k}{2}) = \lambda_i(x^*) - \frac{\delta_{\min}(L + \frac{\delta_k}{2})}{2L + \delta_{\min}} \; . \end{aligned} \tag{4.3.9}$$

We now distinguish two cases.
Case 1: $|J_k| = 0$. Then $\delta_k = 0$ because $J_k \cup I = I$ and $\lambda_i(x^*) = 0$ for every $i \in I$. Relation (4.3.9) becomes

$$\lambda_i(x_k) \geq \lambda_i(x^*) - \frac{\delta_{\min} L}{2L + \delta_{\min}},$$

so that for every $i \in I^c$, since $\lambda_i(x^*) \geq \delta_{\min}$, we have

$$\lambda_i(x_k) \geq \delta_{\min} - \frac{\delta_{\min} L}{2L + \delta_{\min}} > 0 \; . \tag{4.3.10}$$

This means that for every $i \in I^c$ we have $(x_k)_i = 0$ by the Case 1 condition $J_k = \emptyset$ and $\lambda_i(x_k) > 0$ by (4.3.10). We can then apply part (b) of Lemma 4.3.2 and conclude $(x_{k+1})_i = 0$ for every $i \in I^c$. Hence $J_{k+1} = \emptyset = J_k$ and Theorem 4.3.3 is proved in this case.
Case 2. $|J_k| > 0$. For every $i \in \operatorname{argmax}\{\lambda_j(x^*) \mid j \in J_k\}$, we have

$$\lambda_i(x^*) = \max_{j \in J_k} \lambda_j(x^*) = \max_{j \in J_k \cup I} \lambda_j(x^*),$$

where we used the fact that $\lambda_j(x^*) = 0 < \lambda_i(x^*)$ for every $j \in I$. Then by the definition of δ_k, it follows

$$\lambda_i(x^*) = \delta_k.$$

Thus (4.3.9) implies

$$\lambda_i(x_k) > \lambda_i(x^*) - \frac{\delta_{\min}(L + \frac{\delta_k}{2})}{2L + \delta_{\min}} = \delta_k - \frac{\delta_{\min}(L + \frac{\delta_k}{2})}{2L + \delta_{\min}}, \tag{4.3.11}$$

where we used (4.3.9) in the inequality. But since $\delta_k \geq \delta_{\min}$ and the function $\delta_{\min} \mapsto -\frac{\delta_{\min}}{2L+\delta_{\min}}$ is decreasing in $\mathbb{R}_{>0}$ we have

$$\delta_k - \frac{\delta_{\min}(L + \frac{\delta_k}{2})}{2L + \delta_{\min}} \geq \delta_k - \frac{\delta_k(L + \frac{\delta_k}{2})}{2L + \delta_k} = \frac{\delta_k}{2} . \tag{4.3.12}$$

Concatenating (4.3.11) with (4.3.12), we finally obtain

$$\lambda_i(x_k) > \frac{\delta_k}{2} . \tag{4.3.13}$$

We now show that $d_k = x_k - e_{\hat{\jmath}}$ with $\hat{\jmath} \in J_k$.
For every $j \in I$, since $\lambda_j(x^*) = 0$, again by Lemma 4.3.1, we have

$$\begin{aligned} |\lambda_j(x_k)| &= |\lambda_j(x_k) - \lambda_j(x^*)| \leq \|x_k - x^*\|_1(L + \delta_k/2) \\ &< r_*(L + \delta_k/2) = \frac{\delta_{\min}(L + \frac{\delta_k}{2})}{2L + \delta_{\min}} \leq \delta_k/2, \end{aligned} \tag{4.3.14}$$

where we used $\|x_k - x^*\|_1 < r_*$, which is true by definition, in the first inequality, and rearranged (4.3.12) to get the last inequality. For every $j \in I^c$, by (4.3.9), we can write

$$\lambda_j(x_k) > \delta_{\min} - \frac{\delta_{\min}(L + \frac{\delta_k}{2})}{2L + \delta_{\min}} > -\frac{\delta_k}{2} .$$

Using this together with (4.3.14) and (4.3.11), we get $-\lambda_j(x_k) < \delta_k/2 < \lambda_h(x_k)$ for every $j \in [1 : n], h \in \text{argmax}\{\lambda_q(x^*) \mid q \in J_k\}$. So the hypothesis of Lemma 4.3.2 is satisfied and $d_k = d_k^{\mathcal{A}} = x_k - e_{\hat{\jmath}}$ with $\hat{\jmath} \in \text{argmax}\{\lambda_j(x_k) \mid j \in S_k\}$. We need to show $\hat{\jmath} \in J_k$. But $S_k \subseteq I \cup J_k$ and by (4.3.14) if $\hat{\jmath} \in I$ then $\lambda_l(x_k) < \delta_k/2 < \lambda_j(x_k)$ for every $j \in \text{argmax}\{\lambda_j(x^*) \mid j \in J_k\}$. If $\hat{\jmath} \in O_k$ then $(x_k)_{\hat{\jmath}} = 0$ and $\hat{\jmath} \notin S_k$. Hence we can conclude $\text{argmax}\{\lambda_j(x_k) \mid j \in S_k\} \subseteq J_k$ and $d_k = x_k - e_{\hat{\jmath}}$ with $\hat{\jmath} \in J_k$. In particular, by (4.3.13) we get

$$\max\{\lambda_j(x_k) \mid j \in J_k\} = \lambda_{\hat{\jmath}}(x_k) > \frac{\delta_k}{2} . \tag{4.3.15}$$

We now want to show that $\alpha_k = \alpha_k^{\max}$. Assume by contradiction $\alpha_k < \alpha_{\max}$. Then by the lower bound on the step size and (4.3.13)

$$\alpha_k \geq \frac{-\nabla f(x_k)^\top d_k}{L\|d_k\|^2} = \frac{\lambda_i(x_k)}{L\|d_k\|^2} \geq \frac{\delta_{\min}}{2L\|d_k\|^2}\,, \tag{4.3.16}$$

where in the last inequality we used (4.3.15) together with $\delta_k \geq \delta_{\min}$. Also, by Lemma 4.2.5

$$\begin{aligned} &\|d_k\| = \|e_{\hat{\jmath}} - x_k\| \leq \sqrt{2}(e_{\hat{\jmath}} - x_k)_{\hat{\jmath}} = -\sqrt{2}(d_k)_{\hat{\jmath}} \Rightarrow \frac{(d_k)_{\hat{\jmath}}}{\|d_k\|^2} \leq \frac{(d_k)_{\hat{\jmath}}}{\|d_k\|\sqrt{2}} \leq -1/2 \\ &(x_k)_{\hat{\jmath}} = (x_k - x^*)_{\hat{\jmath}} \leq \frac{\|x_k - x^*\|_1}{2} < \frac{r_*}{2} = \frac{\delta_{\min}}{4L + 2\delta_{\min}}. \end{aligned} \tag{4.3.17}$$

Finally, combining (4.3.17) with (4.3.16)

$$\begin{aligned} (x_{k+1})_{\hat{\jmath}} = (x_k)_{\hat{\jmath}} + (d_k)_{\hat{\jmath}}\alpha_k &< \frac{r_*}{2} - \frac{\|d_k\|^2}{2}\alpha_k \leq \frac{r_*}{2} - \frac{\|d_k\|^2}{2}\frac{\delta_{\min}}{2L\|d_k\|^2} \\ &= \frac{\delta_{\min}}{4L + 2\delta_{\min}} - \frac{\delta_{\min}}{4L} < 0, \end{aligned}$$

where we used (4.3.16) to bound α_k in the first inequality, (4.3.17) to bound $(x_k)_{\hat{\jmath}}$ and $\frac{(d_k)_{\hat{\jmath}}}{\|d_k\|^2}$. Hence $(x_{k+1})_{\hat{\jmath}} < 0$, contradiction. □

4.4 Active set complexity bounds

Before giving the active set complexity bounds in several settings it is important to clarify that by active set associated to a stationary point x^* we do not mean the set $\mathrm{supp}(x^*)^c = \{i \in [1:n] \mid (x^*)_i = 0\}\}$ but the set $I^c(x^*) = \{i \in [1:n] \mid \lambda_i(x^*) > 0\}$ of binding constraints. In general $I^c(x^*) \subset \mathrm{supp}(x^*)^c$ by complementarity conditions, with

$$\mathrm{supp}(x^*)^c = I^c(x^*) \Leftrightarrow \text{strict complementarity holds in } x^*. \tag{4.4.1}$$

The face $\mathcal{F}$ of Δ_{n-1} defined by the constraints with indices in $I^c(x^*)$ still has a nice geometrical interpretation: it is the face of Δ_{n-1} exposed by $-\nabla f(x^*)$.
It is at this point natural to require that the sequence $\{x_k\}$ converges to a subset A of $\mathcal{X}^*$ for which I^c is constant. This motivates the following definition:

Definition 4.4.1. A compact subset A of $\mathcal{X}^*$ is said to have the *support identification property (SIP)* if there exists an index set $I_A^c \subset [1:n]$ such that

$$I^c(x) = I_A^c \quad \text{for all } x \in A\,.$$

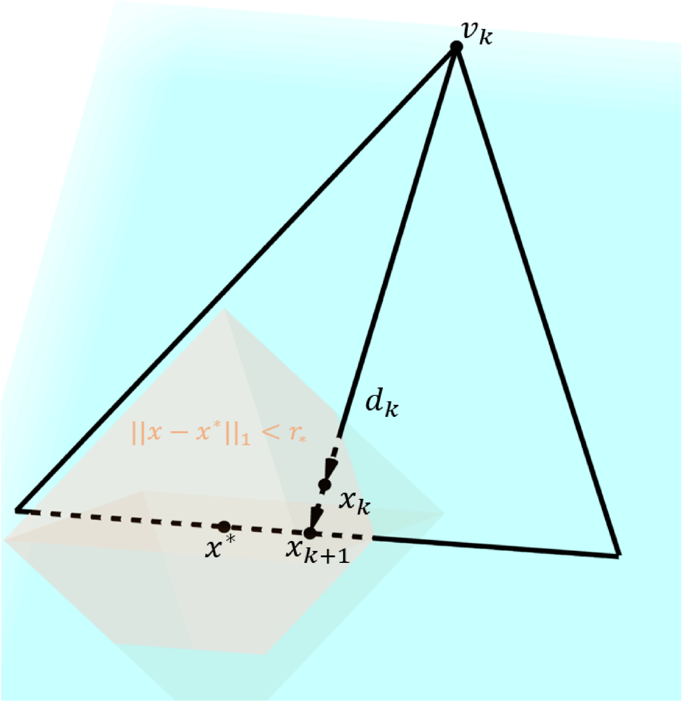


Figure 4.1: Away step identifies one active constraint

In other words, A has the SIP if and only if the set of binding constraints I^c is constant for x varying in A. The geometrical interpretation of Definition 4.4.1 is the following: for every point x in the subset A, the negative gradient $-\nabla f(x)$ exposes the same face. This is trivially true if A is a singleton so that the notion of subset with the SIP generalizes the one of stationary point. From the geometrical interpretation it is clear that A has the SIP also if it is contained in the relative interior of a face $\mathcal{F}$ of Δ_{n-1} and strict complementarity conditions hold for every point in A. In this case the negative gradient of the points in A always exposes $\mathcal{F}$. As a pathological example, for $f \equiv 0$ all the subsets of Δ_{n-1} have the SIP because every $x \in \Delta_{n-1}$ is stationary with $I^c(x) = \emptyset$.

We further define

$$\delta_{\min}(A) = \min\{\lambda_i(x) \mid x \in A,\ i \in I_A^c\} \ .$$

Notice that by the compactness of A we always have $\delta_{\min}(A) > 0$ if A enjoys the SIP. We can finally give a rigorous definition of what it means to solve the active set problem:

Definition 4.4.2. Consider an algorithm generating a sequence $\{x_k\}$ converging to a subset A of $\mathcal{X}^*$ enjoying the SIP. We say that this algorithm solves the active set problem in M steps if $(x_k)_i = 0$ for every $i \in I_A^c$, $k \geq M$. If, given a set of conditions on (A, f, x_0), M is the minimum number which has this property for every sequence generated by the algorithm, then we say that the active set complexity of the algorithm is M, under the given conditions.

We can now apply Theorem 4.3.3 to show that once a sequence is definitely close enough to a set enjoying the SIP, the AFW identifies the active set in at most $|I^c|$ steps.

Theorem 4.4.3. *Let $\{x_k\}$ be a sequence generated by the AFW, with step size $\alpha_k \geq \bar{\alpha}_k$. Let $\mathcal{X}^*$ be the set of stationary points of a function $f : \Delta_{n-1} \to \mathbb{R}$ with ∇f having Lipschitz constant L. Assume that there exists a compact subset A of $\mathcal{X}^*$ with the SIP such that $x_k \to A$. Then there exists M such that*

$$(x_k)_i = 0 \quad \textit{for every } k \geq M \textit{ and all } i \in I_A^c\,.$$

Proof. Let $J_k = \{i \in I_A^c \mid (x_k)_i > 0\}$ and choose $\bar{k}$ such that $\text{dist}_1(x_k, A) < \frac{\delta_{\min}(A)}{2L+\delta_{\min}(A)} = r_*$ for every $k \geq \bar{k}$. Then for every $k \geq \bar{k}$ there exists $y^* \in A$ with $\|x_k - y^*\|_1 < r_*$. But since by hypothesis for every $y^* \in A$ the support of the multiplier function is I_A^c with $\delta_{\min}(A) \leq \lambda_i(y^*)$ for every $i \in I_A^c$, we can apply Theorem 4.3.3 with y^* as fixed point and obtain that $|J_{k+1}| \leq \max(0, |J_k| - 1)$. This means that it takes at most $|J_{\bar{k}}| \leq |I_A^c|$ steps for all the variables with indices in I_A^c to be 0. Again by (4.3.8), we conclude by induction $|J_k| = 0$ for every $k \geq M = \bar{k} + |I_A^c|$, since $|J_{\bar{k}+|I_A^c|}| = 0$. □

The proof of Theorem 4.4.3 also gives a relatively simple upper bound for the complexity of the active set problem:

Proposition 4.4.4. *Under the assumptions of Theorem 4.4.3, the active set complexity is at most*

$$\min\{\bar{k} \in \mathbb{N}_0 \mid \text{dist}_1(x_k, A) < r_* \forall k \geq \bar{k}\} + |I_A^c|,$$

where $r_* = \frac{\delta_{\min}(A)}{2L+\delta_{\min}(A)}$.

We now report an explicit bound for the strongly convex case, and analyze in depth the nonconvex case in Section 4.5. From strong convexity of f, it is easy to see that the following inequality holds for every x on Δ_{n-1}:

$$f(x) \geq f(x^*) + \frac{u_1}{2}\|x - x^*\|_1^2, \tag{4.4.2}$$

with $u_1 > 0$.

Corollary 4.4.5. *Let $\{x_k\}$ be the sequence of points generated by AFW with $\alpha_k \geq \bar{\alpha}_k$. Assume that f is strongly convex and let*

$$h_k \leq q^k h_0, \tag{4.4.3}$$

with $q < 1$ and $h_k = f(x_k) - f_*$, be the convergence rate related to the AFW (see [157], Theorem 8). Then the active set complexity is

$$\max\left(0, \left\lceil \frac{\ln(h_0) - \ln(u_1 r_*^2/2)}{\ln(1/\mathrm{q})} \right\rceil\right) + |I^c| \ . \tag{4.4.4}$$

Proof. Notice that by the linear convergence rate (4.4.3), and the fact that $q < 1$, the number of steps needed to reach the condition

$$h_k \leq \frac{u_1}{2} r_*^2 \tag{4.4.5}$$

is at most

$$\bar{k} = \max\left(0, \left\lceil \frac{\ln(h_0) - \ln(u_1 r_*^2/2)}{\ln(1/q)} \right\rceil\right) \ .$$

We claim that if condition (4.4.5) holds then it takes at most $|I^c|$ steps for the sequence to be definitely in the active set.
Indeed, if $q^k h_0 \leq \frac{u_1}{2} r_*^2$ then necessarily $x_k \in B_1(x^*, r_*)$ by (4.4.2), and by monotonicity of the bound (4.4.3) we then have $x_{k+h} \in B_1(x^*, r_*)$ for every $h \geq 0$. Once the sequence is definitely in $B_1(x^*, r_*)$ by (4.3.8) it takes at most $|J_{\bar{k}}| \leq |I^c|$ steps for all the variables with indices in I^c to be 0. To conclude, again by (4.3.8) since $|J_{\bar{k}+|I^c|}| = 0$ by induction $|J_m| = 0$ for every $m \geq \bar{k} + |I^c|$. □

Remark 4.4.6. In Corollary 4.4.5, if we assume the linear rate (4.4.3) (which may not hold in the nonconvex case), then the strong convexity of f can be replaced by the condition (4.4.2).

4.5 Active set complexity for nonconvex objectives

In this section, we focus on problems with nonconvex objectives. We first give a more explicit convergence rate for AFW in the nonconvex case, then we prove a general active set identification result for the method. Finally, we analyze both local and global active set complexity bounds related to AFW. A fundamental element in our analysis is the FW gap function $g : \Delta_{n-1} \to \mathbb{R}$ defined as

$$g(x) = \max_{i\in[1:n]} \{-\lambda_i(x)\} \ .$$

We clearly have $g(x) \geq 0$ for every $x \in \Delta_{n-1}$, with equality iff x is a stationary point. The reason why this function is called FW gap is evident from the relation

$$g(x_k) = -\nabla f(x_k)^\top d_k^{\mathcal{FW}}.$$

This is a standard quantity appearing in the analysis of FW variants (see, e.g., [136]) and is computed for free at each iteration of a FW-like algorithm. In [156], the author uses the gap to analyze the convergence rate of the classic FW algorithm in the nonconvex case. More specifically, a convergence rate of $O(\frac{1}{\sqrt{k}})$ is proved for the minimal FW gap up to iteration k:

$$g_k^* = \min_{0\leq i\leq k-1} g(x_i).$$

The results extend in a nice and straightforward way the ones reported in [192] for proving the convergence of gradient methods in the nonconvex case. Inspired by the analysis of the AFW method for strongly convex objectives reported in [200], we now study the AFW convergence rate in the nonconvex case with respect to the sequence $\{g_k^*\}$.

4.5.1 Global convergence

We start investigating the minimal FW gap, giving estimates of rates of convergence. In the next theorem and in the subsequent Corollary 4.5.2 we assume that the AFW starts from a vertex of the probability simplex. Thanks to the affine invariance properties of the AFW this is not a restrictive assumption. For a generic starting point one can indeed apply the same theorem to the AFW starting from e_{n+1} for $\tilde{f} : \Delta_n \to \mathbb{R}$ satisfying

$$\tilde{f}(y) = f(y_1 e_1 + \cdots + y_n e_n + y_{n+1} x_0), \tag{4.5.1}$$

where $x_0 \in \Delta_{n-1}$ is the desired starting point (see also Corollary 4.5.3). Formally, this leads to the computation of a sequence $\{y_k\}$ on Δ_n which can be mapped to a sequence $\{x_k\}$ on Δ_{n-1} by the affine transformation

$$p(y) = y_1 e_1 + \cdots + y_n e_n + y_{n+1} x_0 \, . \tag{4.5.2}$$

In Section 4.6, we discuss the invariance of the AFW under affine transformations in more detail.

Theorem 4.5.1. *Let $f^* = \min_{x\in\Delta_{n-1}} f(x)$, and let $\{x_k\}$ be a sequence generated by the AFW algorithm applied to f on Δ_{n-1}, with x_0 a vertex of Δ_{n-1}. Assume that the step size α_k is larger or equal than $\bar{\alpha}_k$ (as defined in* (4.2.1)*), and that*

$$f(x_k) - f(x_k + \alpha_k d_k) \geq \rho\bar{\alpha}_k \left(-\nabla f(x_k)^\top d_k\right) \tag{4.5.3}$$

for some fixed $\rho > 0$. Then for every $T \in \mathbb{N}$

$$g_T^* \leq \max\left(\sqrt{\frac{4L(f(x_0) - f^*)}{\rho T}}, \frac{4(f(x_0) - f^*)}{T}\right) . \tag{4.5.4}$$

Proof. Let $r_k = -\nabla f(x_k)$ and $g_k = g(x_k)$. We distinguish three cases.

Case 1. $\bar{\alpha}_k < \alpha_k^{\max}$. Then $\bar{\alpha}_k = \frac{-\nabla f(x_k)^\top d_k}{L\|d_k\|^2}$ and relation (4.5.3) becomes

$$f(x_k) - f(x_k + \alpha_k d_k) \geq \rho\bar{\alpha}_k r_k^\top d_k = \frac{\rho}{L\|d_k\|^2}(r_k^\top d_k)^2$$

and consequently

$$f(x_k) - f(x_{k+1}) \geq \frac{\rho}{L\|d_k\|^2}(r_k^\top d_k)^2 \geq \frac{\rho}{L\|d_k\|^2} g_k^2 \geq \frac{\rho g_k^2}{2L}, \tag{4.5.5}$$

where we used $r_k^\top d_k \geq g_k$ in the second inequality and $\|d_k\| \leq \sqrt{2}$ in the third one. As for S_k, by hypothesis we have either $d_k = d_k^{\mathcal{FW}}$ so that $d_k = e_i - x_k$ or $d_k = d_k^{\mathcal{A}} = x_k - e_i$ for some $i \in \{1, ..., n\}$. In particular $S_{k+1} \subseteq S_k \cup \{i\}$ so that $|S_{k+1}| \leq |S_k| + 1$.
Case 2: $\alpha_k = \bar{\alpha}_k = \alpha_k^{\max} = 1, d_k = d_k^{\mathcal{FW}}$. By the standard descent lemma [31, Proposition 6.1.2] applied to f with center x_k and $\alpha = 1$

$$f(x_{k+1}) = f(x_k + d_k) \leq f(x_k) + \nabla f(x_k)^\top d_k + \frac{L}{2}\|d_k\|^2 .$$

Since by the Case 2 condition $\min\left(\frac{-\nabla f(x_k)^\top d_k}{\|d_k\|^2 L}, 1\right) = \alpha_k = 1$ we have

$$\frac{-\nabla f(x_k)^\top d_k}{\|d_k\|^2 L} \geq 1 \text{, so} \quad -L\|d_k\|^2 \geq \nabla f(x_k)^\top d_k ,$$

hence we can write

$$f(x_k) - f(x_{k+1}) \geq -\nabla f(x_k)^\top d_k - \frac{L}{2}\|d_k\|^2 \geq -\frac{\nabla f(x_k)^\top d_k}{2} \geq \frac{1}{2} g_k . \tag{4.5.6}$$

Reasoning as in Case 1 we also have $|S_{k+1}| \leq |S_k| + 1$.
Case 3: $\alpha_k = \bar{\alpha}_k = \alpha_k^{\max}$, $d_k = d_k^{\mathcal{A}}$. Then $d_k = x_k - e_i$ for $i \in S_k$ and

$$(x_{k+1})_j = (1 + \alpha_k)(x_k)_j - \alpha_k(e_i)_j,$$

with $\alpha_k = \alpha_k^{\max} = \frac{(x_k)_i}{1-(x_k)_i}$. Therefore $(x_{k+1})_j = 0$ for $j \in \{1, ..., n\} \setminus S_k \cup \{i\}$ and $(x_{k+1})_j \neq 0$ for $j \in S_k \setminus \{i\}$. In particular $|S_{k+1}| = |S_k| - 1$.

For $i = 1, 2, 3$ let now $n_i(T)$ be the number of Case i steps done in the first T iterations of the AFW. We have by induction on the recurrence relation we proved for $|S_k|$

$$|S_T| - |S_0| \leq n_1(T) + n_2(T) - n_3(T) \ , \tag{4.5.7}$$

for every $T \in \mathbb{N}$.
Since $n_3(T) = T - n_1(T) - n_2(T)$ from (4.5.7) we get

$$n_1(T) + n_2(T) \geq \frac{T + |S_T| - |S_0|}{2} \geq \frac{T}{2} \ , \tag{4.5.8}$$

where we used $|S_0| = 1 \leq |S_T|$. Let now C_i^T be the set of iteration counters up to $T - 1$ corresponding to Case i steps for $i \in \{1, 2, 3\}$, which satisfies $|C_i^T| = n_i(T)$. We have by summing (4.5.5) and (4.5.6) for the indices in C_1^T and C_2^T respectively

$$\sum_{k \in C_1^T} f(x_k) - f(x_{k+1}) + \sum_{k \in C_2^T} f(x_{k+1}) - f(x_k) \geq \sum_{k \in C_1^T} \frac{\rho g_k^2}{2L} + \sum_{k \in C_2^T} \frac{1}{2} g_k \ . \tag{4.5.9}$$

We now lower bound the right-hand side of (4.5.9) in terms of g_T^* as follows:

$$\begin{aligned} &\sum_{k \in C_1^T} \frac{\rho g_k^2}{2L} + \sum_{k \in C_2^T} \frac{1}{2} g_k \geq |C_1^T| \min_{k \in C_1^T} \frac{\rho g_k^2}{2L} + |C_2^T| \min_{k \in C_2^T} \frac{g_k}{2} \\ \geq &(|C_1^T| + |C_2^T|) \min\left(\frac{\rho (g_T^*)^2}{2L}, \frac{g_T^*}{2}\right) = [n_1(T) + n_2(T)] \min\left(\rho\frac{(g_T^*)^2}{2L}, \frac{g_T^*}{2}\right) \\ \geq &\frac{T}{2} \min\left(\frac{\rho (g_T^*)^2}{2L}, \frac{g_T^*}{2}\right) \ . \end{aligned} \tag{4.5.10}$$

Since the left-hand side of (4.5.9) can clearly be upper bounded by $f(x_0) - f^*$ we have

$$f(x_0) - f^* \geq \frac{T}{2} \min\left(\frac{\rho (g_T^*)^2}{2L}, \frac{g_T^*}{2}\right) \ .$$

To finish, if $\frac{T}{2}\min\left(\frac{g_T^*}{2}, \frac{\rho(g_T^*)^2}{2L}\right) = \frac{Tg_T^*}{4}$ we then have

$$g_T^* \leq \frac{4(f(x_0) - f^*)}{T} \tag{4.5.11}$$

and otherwise

$$g_T^* \leq \sqrt{\frac{4L(f(x_0) - f^*)}{\rho T}}\ . \tag{4.5.12}$$

The claim follows by taking the max in the system formed by (4.5.11) and (4.5.12). □

In Section 4.2.2, we prove that condition (4.5.3) is satisfied by exact line search and Armijo line search as well. We also prove that it is satisfied if we impose the weak Wolfe conditions and take $\alpha_k^{\max}$ whenever the conditions are incompatible with the constraint $\alpha_k \leq \alpha_k^{\max}$.

When the step sizes coincide with the lower bounds $\bar{\alpha}_k$ or are obtained using exact line search, we have the following corollary:

Corollary 4.5.2. *Under the assumptions of Theorem 4.5.1, if $\alpha_k = \bar{\alpha}_k$ or if α_k is selected by exact line search then for every $T \in \mathbb{N}$*

$$g_T^* \leq \max\left(\sqrt{\frac{8L(f(x_0) - f^*)}{T}}, \frac{4(f(x_0) - f^*)}{T}\right)\ . \tag{4.5.13}$$

Proof. By points 2 and 3 of Lemma 4.2.1, relation (4.5.3) is satisfied with $\rho = \frac{1}{2}$ for both $\alpha_k = \bar{\alpha}_k$ and α_k given by exact line search, and we also have $\alpha_k \geq \bar{\alpha}_k$ in both cases. The conclusion follows directly from Theorem 4.5.1. □

Applying the trick of adding the starting point as a vertex allows us to drop the assumptions of starting from a vertex in Theorem 4.5.1.

Corollary 4.5.3. *Let $x_0 \in \Delta_{n-1}$, and let $\{y_k\}$ be a sequence generated by the AFW applied to the objective function $\tilde{f}$ defined in* (4.5.1) *with $y_0 = e_{n+1}$. Let $\{x_k\} = \{p(y_k)\}$. Then under the assumptions of Theorem 4.5.1 on α_k and f, the bound* (4.5.4) *and Corollary 4.5.2 still hold.*

Proof. The multipliers are invariant by affine transformation (see Section 4.6 for further details), and since the FW gap depends on the multipliers, it is also invariant under affine transformation. Also adding the multiplier related to x_0 does not change the FW gap, which is always realized in one of the vertices of the original simplex since it is the maximum of a linear function plus a constant. Therefore, the FW gap is invariant with respect to p, so that the same arguments used for Theorem 4.5.1 and Corollary 4.5.2 can still be applied to $\{x_k\} = \{p(y_k)\}$. □

Since adding a vertex alters the active set identification properties of the problem (e.g., the active set radius), we cannot apply the above results directly in the rest of this section. Instead we use some key intermediate results presented in the proof of Theorem 4.5.1.

4.5.2 A general active set identification result

We can now give a general active set identification result in the nonconvex setting. While we do not use strict complementarity when the step sizes are given by (4.2.1), without this assumption we need strict complementarity.
If $A \subseteq \mathcal{X}^*$ enjoys the SIP and if strict complementarity is satisfied for every $x \in A$, then as a direct consequence of (4.4.1) we have

$$\text{supp}(x) = [1 : n] \setminus I^c(x) = [1 : n] \setminus I_A^c \tag{4.5.14}$$

for every $x \in A$. In this case we can then define supp(A) as the (common) support of the points in A.

For the result we need an observation on connectedness which seems to be folklore in an optimization context. This property is needed, e.g. for the proof of [192, Theorem 4.1.2] and similar results are discussed in [25]. However, we are not aware of an explicit proof for this property, so for the readers' convenience we provide a short argument:

Lemma 4.5.4. *Let $\{x_k\}$ be a bounded sequence in $\mathbb{R}^n$ such that $\|x_k - x_{k+1}\| \to 0$. Then the set of limit points of $\{x_k\}$ is connected.*

Proof. Assume by contradiction that there are two open sets U_1 and U_2 separating the limit points of $\{x_k\}$. Then there must exist an infinite number of points from $\{x_k\}$ both in U_1 and U_2, and in particular a subsequence $\{x_{k(j)}\}$ of $\{x_k\}$ such that $x_{k(j)} \in U_1$ and $x_{k(j)+1} \in U_1^c$ for every $j \in \mathbb{N}_0$. By the condition $\|x_{k(j)} - x_{k(j)+1}\| \to 0$ we obtain

$$\text{dist}(x_{k(j)}, U_1^c) \to 0\,. \tag{4.5.15}$$

Since $\{x_{k(j)}\}$ is bounded by hypothesis it has a non empty set of limit points. But every limit point of $\{x_{k(j)}\}$ must be necessarily in U_1^c by (4.5.15) and also in the closure of U_1 (because $\{x_{k(j)}\} \subset U_1$) and therefore not in U_2, a contradiction. □

We proceed with the announced result.

Theorem 4.5.5. *Let $\{x_k\}$ be the sequence generated by the AFW method with step sizes satisfying $\alpha_k \geq \bar{\alpha}_k$ and* (4.5.3)*, where $\bar{\alpha}_k$ is given by* (4.2.1)*. Let $\mathcal{X}^*$ be the subset of stationary points of f. We have:*

(a) $x_k \rightarrow \mathcal{X}^*$.

(b) If $\alpha_k = \bar{\alpha}_k$ then $\{x_k\}$ converges to a connected component A of $\mathcal{X}^*$. If additionally A has the SIP then $\{x_k\}$ identifies I^c_A in finite time.

Assume now that $\mathcal{X}^* = \bigcup_{i=1}^{C} A_i$ with A_i compact for each $i \in [1:C]$, with distinct supports and such that A_i has the SIP for each $i \in [1:C]$.

(c) If $\alpha_k \geq \bar{\alpha}_k$ and if strict complementarity holds for all points in $\mathcal{X}^*$ then $\{x_k\}$ converges to A_l for some $l \in [1:C]$ and identifies $I^c_{A_l}$ in finite time.

Proof. a) By the proof of Theorem 4.5.1 and the continuity of the multiplier function we have

$$x_{k(j)} \rightarrow g^{-1}(0) = \mathcal{X}^* \ , \tag{4.5.16}$$

where $\{k(j)\}$ is the sequence of indexes corresponding to Case 1 or Case 2 steps. Let $k'(j)$ be the sequence of indexes corresponding to Case 3 steps. Since for such steps $\alpha_{k'(j)} = \bar{\alpha}_{k'(j)}$ we can apply Corollary 4.2.2 to obtain

$$\|x_{k'(j)} - x_{k'(j)+1}\| \rightarrow 0 \ . \tag{4.5.17}$$

Combining (4.5.16), (4.5.17) and the fact that there can be at most $n-1$ consecutive Case 3 steps, we get $x_k \rightarrow \mathcal{X}^*$.

b) By the boundedness of f and point 2 of Lemma 4.2.1 if $\alpha_k = \bar{\alpha}_k$ then $\|x_{k+1}-x_k\| \rightarrow 0$. Now Lemma 4.5.4 together with point a) ensures that the set of limit points must be contained in a connected component A of $\mathcal{X}^*$. By Theorem 4.4.3 it follows that if A has constant support $\{x_k\}$ identifies I^c_A in finite time.

c) Consider a disjoint family of subsets $\{U_i\}_{i=1}^{C}$ of Δ_{n-1} with $U_i = \{x \in \Delta_{n-1} \mid \text{dist}_1(x, A_i) \leq r_i\}$ where r_i is small enough to ensure some conditions that we now specify. First, we need

$$r_i < \frac{\delta_{\min}(A_i)}{2L + \delta_{\min}(A_i)}$$

so that r_i is smaller than the active set radius of every $x \in A_i$ and in particular for every $x \in U_i$ there exists $x^* \in A_i$ such that

$$\|x - x^*\|_1 < \frac{\delta_{\min}(x^*)}{2L + \delta_{\min}(x^*)}. \tag{4.5.18}$$

Second, we choose r_i small enough so that $\{U_i\}_{i=1}^C$ are disjoint and

$$\mathrm{supp}(y) \supseteq \mathrm{supp}(A_i) \ \forall y \in U_i \ , \tag{4.5.19}$$

where these conditions can be always satisfied thanks to the compactness of A_i. Assume now by contradiction that the set S of limit points of $\{x_k\}$ intersects more than one of the $\{A_i\}_{i=1}^C$. Let in particular A_l minimize $|\mathrm{supp}(A_l)|$ among the sets containing points of S. By point a) $x_k \in \cup_{i=1}^C U_i$ for $k \geq M$ large enough and we can define an infinite sequence $\{t(j)\}$ of exit times greater than M for U_l so that $x_{t(j)} \in U_l$ and $x_{t(j)+1} \in \cup_{i \in [1:C] \setminus l} U_i$. Up to considering a subsequence we can assume $x_{t(j)+1} \in U_m$ for a fixed $m \neq l$ for every $j \in \mathbb{N}_0$.

We now distinguish two cases as in the proof of Theorem 4.3.3, where by equation (4.5.18) the hypotheses of Theorem 4.3.3 are satisfied for $k = t(j)$ and some $x^* \in A_l$.

Case 1. $(x_{t(j)})_h = 0$ for every $h \in I_{A_l}^c$. In the notation of Theorem 4.3.3 this corresponds to the case $|J_{t(j)}| = 0$. Then by (4.3.10) we also have $\lambda_h(x_{t(j)}) > 0$ for every $h \in I_{A_l}^c$. Thus $(x_{t(j)+1})_h = (x_{t(j)})_h = 0$ for every $h \in I_{A_l}^c$ by Lemma 4.3.2, so that we can write

$$\mathrm{supp}(A_m) \subseteq \mathrm{supp}(x_{t(j)+1}) \subseteq [1:n] \setminus I_{A_l}^c = \mathrm{supp}(A_l), \tag{4.5.20}$$

where the first inclusion is justified by (4.5.19) for $i = m$ and the second by strict complementarity (see also (4.5.14) and the related discussion). But since by hypothesis $\mathrm{supp}(A_m) \neq \mathrm{supp}(A_l)$ the inclusion (4.5.20) is strict and so it is in contradiction with the minimality of $|\mathrm{supp}(A_l)|$.

Case 2. $|J_{t(j)}| > 0$. Then reasoning as in the proof of Theorem 4.3.3 we obtain $d_{t(j)} = x_{t(j)} - e_{\bar{h}}$ for some $\bar{h} \in J_{t(j)} \subset I_{A_l}^c$. Let $\tilde{x}^* \in A_l$, and let $\tilde{d} = \alpha_{t(j)} d_{t(j)}$. The sum of the components of $\tilde{d}$ is 0 with the only negative component being $\tilde{d}_{\bar{h}}$ and therefore

$$\tilde{d}_{\bar{h}} = - \sum_{h \in [1:n] \setminus \bar{h}} \tilde{d}_h = - \sum_{h \in [1:n] \setminus \bar{h}} |\tilde{d}_h| \tag{4.5.21}$$

We claim that $\|x_{t(j)+1} - \tilde{x}^*\|_1 \leq \|x_{t(j)} - \tilde{x}^*\|_1$. This is enough to finish because since $\tilde{x}^* \in A_l$ is arbitrary then it follows $\mathrm{dist}_1(x_{t(j)+1}, A_l) \leq \mathrm{dist}_1(x_{t(j)}, A_l)$ so that $x_{t(j)+1} \in U_l$, a contradiction.

We have

$$\begin{aligned}
&\|\tilde{x}^* - x_{t(j)+1}\|_1 = \|\tilde{x}^* - x_{t(j)} - \alpha_{t(j)} d_{t(j)}\|_1 \\
=&|\tilde{x}^*_{\bar{h}} - (x_{t(j)})_{\bar{h}} - \tilde{d}_{\bar{h}}| + \sum_{h \in [1:n] \setminus \bar{h}} |\tilde{x}^*_h - (x_{t(j)})_h - \tilde{d}_h| \\
=&|\tilde{x}^*_{\bar{h}} - (x_{t(j)})_{\bar{h}}| + \tilde{d}_{\bar{h}} + \sum_{h \in [1:n] \setminus \bar{h}} |\tilde{x}^*_h - (x_{t(j)})_h - \tilde{d}_h| \\
\leq&|\tilde{x}^*_{\bar{h}} - (x_{t(j)})_{\bar{h}}| + \tilde{d}_{\bar{h}} + \sum_{h \in [1:n] \setminus \tilde{h}} (|\tilde{x}^*_h - (x_{t(j)})_h| + |\tilde{d}_h|) \\
=&\|x_{t(j)} - \tilde{x}^*\|_1 + \tilde{d}_{\bar{h}} + \sum_{h \in [1:n] \setminus \bar{h}} |\tilde{d}_h| = \|x_{t(j)} - \tilde{x}^*\|_1
\end{aligned}$$

where in the third equality we used $0 = \tilde{x}^*_{\bar{h}} \leq -\tilde{d}_{\bar{h}} \leq (x_{t(j)})_{\bar{h}}$ and in the last equality we used (4.5.21).

Reasoning by contradiction we have proved that all the limit points of $\{x_k\}$ are in A_l for some $l \in [1, ..., C]$. The conclusion follows immediately from Theorem 4.4.3. □

4.5.3 Quantitative version of active set identification

Let $q : \mathbb{R}_{>0} \to \mathbb{N}_0$ be such that $f(x_k) - f(x_{k+1}) \leq \varepsilon$ for every $k \geq q(\varepsilon)$. In this section, we give global active set complexity bounds for non convex objectives as a function of q, which measures how long it takes for $\gamma_k = f(x_k) - f(x_{k+1})$ to fall definitely under a threshold value. We assume that the gap function $g(x)$ satisfies the Hölderian error bound condition

$$g(x) \geq \theta \; \mathrm{dist}_1(x, \mathcal{X}^*)^p \tag{4.5.22}$$

for some $\theta, p > 0$. This condition is satisfied, e.g., if $f(x)$ (and therefore $\nabla f(x)$) is a semialgebraic function. In this case then also $g(x)$ is semialgebraic because obtained by sums, products and maxima of semialgebraic functions, and (4.5.22) holds by Łojasiewicz' inequality (Corollary 2.6.7 in [35], see also [38] and references) applied to g and $\mathrm{dist}_1(x, \mathcal{X}^*)$.

In the convex case, condition (4.5.22) on the FW gap $g(x)$ is *weaker* than the more common Hölderian error bound condition on the objective, see [38, 148, 243]. This follows trivially from the fact that the FW gap $g(x)$ is always larger than the objective gap $f(x) - f^*$ for convex f. The Hölderian error bound assumption on the gap allows us to give more explicit active set complexity bounds.

Theorem 4.5.6. *Assume $\mathcal{X}^* = \bigcup_{i \in [1:C]} A_i$ where A_i is compact and with the SIP for every $i \in [1 : C]$ and $0 < d \overset{\text{def}}{=} \min_{\{i,j\} \subset [1:C]} \mathrm{dist}_1(A_i, A_j)$. Let r_* be the minimum* *active set radius of the sets* $\{A_i\}_{i=1}^{C}$*. Assume that* $g(x)$ *satisfies* (4.5.22)*. Assume that the step sizes satisfy* $\alpha_k = \bar{\alpha}_k$*, with* $\bar{\alpha}_k$ *given by* (4.2.1)*. Then the active set complexity is at most* $q(\bar{\varepsilon}) + n - 1$ *for* $\bar{\varepsilon}$ *satisfying the following conditions*

$$\bar{\varepsilon} < L\,, \quad \left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}} < r_* \quad \text{and} \quad 2\left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}} + 2n\sqrt{\frac{2\bar{\varepsilon}}{L}} \leq d\ . \tag{4.5.23}$$

The proof is essentially a quantitative version of the argument used to prove point b) of Theorem 4.5.5.

Proof. Fix $k \geq q(\bar{\varepsilon})$, so that

$$f(x_k) - f(x_{k+1}) \leq \bar{\varepsilon}\ . \tag{4.5.24}$$

We refer to Case i steps for $i \in [1:3]$ following the definitions in Theorem 4.5.1. If the step k is a Case 1 step, then by (4.5.5) with $\rho = 1/2$ we have

$$f(x_k) - f(x_{k+1}) \geq \frac{g(x_k)^2}{4L}$$

and this together with (4.5.24) implies

$$2\sqrt{L\bar{\varepsilon}} \geq 2\sqrt{L(f(x_k) - f(x_{k+1}))} \geq g(x_k)\ .$$

Analogously, if the step k is a Case 2 step, then by (4.5.6) we have

$$f(x_k) - f(x_{k+1}) \geq \frac{g(x_k)}{2}$$

so that $2\bar{\varepsilon} \geq g(x_k)$. By the leftmost condition in (4.5.23) we have $\bar{\varepsilon} < L$ so that $2\sqrt{L\bar{\varepsilon}} \geq 2\bar{\varepsilon}$, and therefore for both Case 1 and Case 2 steps we have

$$g(x_k) \leq 2\sqrt{L\bar{\varepsilon}}\ . \tag{4.5.25}$$

By inverting relation (4.2.2), we also have

$$\|x_k - x_{k+1}\| \leq \sqrt{\frac{2(f(x_k) - f(x_{k+1}))}{L}} \leq \sqrt{\frac{2\bar{\varepsilon}}{L}}\ . \tag{4.5.26}$$

Now let $\bar{k} \geq q(\bar{\varepsilon})$ be such that step $\bar{k}$ is a Case 1 or Case 2 step. By the error bound condition together with (4.5.25)

$$\text{dist}_1(x_{\bar{k}}, \mathcal{X}^*) \leq \left(\frac{g(x_{\bar{k}})}{\theta}\right)^{\frac{1}{p}} \leq \left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}} < r_*\ , \tag{4.5.27}$$

where we used (4.5.25) in the second inequality and the second condition of (4.5.23) in the third inequality. In particular there exists l such that $\mathrm{dist}_1(x_{\bar{k}}, A_l) \leq (2\sqrt{L\bar{\varepsilon}}/\theta)^{1/p}$. We claim now that $I^c_{A_l}$ is already identified at the step $\bar{k}$.

First, we claim that for every Case 1 or Case 2 step with index $\tau \geq \bar{k}$ we have $\mathrm{dist}_1(x_\tau, A_l) \leq (g(x_\tau)/\theta)^{1/p}$. We reason by induction on the sequence $\{s(k')\}$ of Case 1 or Case 2 steps following $\bar{k}$, so that in particular $s(1) = \bar{k}$ and $\mathrm{dist}_1(x_{s(1)}, A_l) \leq g(x_{s(1)})$ is true by (4.5.27). Since there can be at most $n - 1$ consecutive Case 3 steps, we have $s(k' + 1) - s(k') \leq n$ for every $k' \in \mathbb{N}_0$. Therefore

$$\begin{aligned}\|x_{s(k')} - x_{s(k'+1)}\|_1 \leq \sum_{i=s(k')}^{s(k'+1)-1} \|x_{i+1} - x_i\|_1 \leq 2 \sum_{i=s(k')}^{s(k'+1)-1} \|x_{i+1} - x_i\| \\ \leq 2[s(k' + 1) - s(k')]\sqrt{\frac{2\bar{\varepsilon}}{L}} \leq 2n\sqrt{\frac{2\bar{\varepsilon}}{L}} \ ,\end{aligned} \tag{4.5.28}$$

where in the second inequality we used part 3 of Lemma 4.2.5 to bound each of the summands of the left-hand side, and in the third inequality we used (4.5.26). Assume now by contradiction $\mathrm{dist}_1(x_{s(k'+1)}, A_l) > (g(x_{s(k'+1)})/\theta)^{1/p}$. Then by (4.5.27) applied to $s(k' + 1)$ instead of $\bar{k}$ there must exists necessarily $j \neq l$ such that $\mathrm{dist}_1(x_{s(k'+1)}, A_j) \leq (g(x_{s(k'+1)})/\theta)^{1/p}$. In particular we have

$$\begin{aligned}\|x_{s(k')} - x_{s(k'+1)}\|_1 \geq \mathrm{dist}_1(A_l, A_j) - \mathrm{dist}_1(x_{s(k'+1)}, A_j) - \mathrm{dist}_1(x_{s(k')}, A_l) \\ \geq d - \left(\frac{g(x_{s(k')})}{\theta}\right)^{\frac{1}{p}} - \left(\frac{g(x_{s(k'+1)})}{\theta}\right)^{\frac{1}{p}} \geq d - 2\left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}} \ ,\end{aligned} \tag{4.5.29}$$

where we used (4.5.25) in the last inequality. But by the second condition of (4.5.23), we have

$$d - 2\left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}} > 2n\sqrt{\frac{2\bar{\varepsilon}}{L}} \ . \tag{4.5.30}$$

Concatenating (4.5.28), (4.5.30) and (4.5.29) we get a contradiction and the claim is proved. Notice that an immediate consequence of this claim is $\mathrm{dist}_1(x_\tau, A_l) < r_*$ by (4.5.27) applied to τ instead of $\bar{k}$, where $\tau \geq \bar{k}$ is an index corresponding to a Case 1 or Case 2 step.

To finish the proof, first we have that there exists an index $\bar{k} \in [q(\bar{\varepsilon}), q(\bar{\varepsilon}) + n - 1]$ corresponding to a Case 1 or Case 2 step, since there can be at most $n-1$ consecutive Case 3 steps. Second, since by (4.5.27) we have $\mathrm{dist}_1(x_{\bar{k}}, A_l) < r_*$ and $\bar{k}$ does not correspond to a Case 3 step, by the local identification Theorem 4.3.3 necessarily $(x_{\bar{k}})_i = 0 \ \forall \ i \in I^c_{A_l}$. Moreover, by the claim every Case 1 and Case 2 step following step $\bar{k}$ happens for points inside $B_1(A_l, r_*)$ so it does not change the components corresponding to $I^c_{A_l}$ by the local identification Theorem 4.3.3. At the same time, Case 3 steps do not increase the support, so that $(x_{\bar{k}+l})_i = 0$ for every $i \in I^c_{A_l}$, $l \geq 0$. Thus active set identification happens in $\bar{k} \leq q(\bar{\varepsilon}) + n - 1$ steps. □

Remark 4.5.7. When we have an explicit expression for the convergence rate $q(\varepsilon)$, then we can get an active set complexity bound using Theorem 4.5.6. For instance, we can compare this result with the one for strongly convex objectives, assuming $C = 1, p = 2$, $\theta = u_1/2$, and $f(x_k) - f(x_{k+1}) \leq h_0 q^k$ for some $q \in (0,1)$. These conditions are always satisfied by strongly convex objectives. Applying the theorem we obtain the active set complexity bound

$$q(\bar{\varepsilon}) + n - 1 \leq \left\lceil \max\left(0, \frac{\ln(h_0) - \ln(\min(L, r_*^4 u_1^2/16L))}{\ln(1/q)}\right)\right\rceil + n \tag{4.5.31}$$

which is always larger than the bound given in (4.4.4). This is expected, given the weaker assumptions on the convergence of the objective and the weaker (at least in the convex case) error bound.

Remark 4.5.8. Assume that the set of stationary points is finite, so that $A_i = \{a_i\}$ for every $i \in [1\!:\!C]$ with $a_i \in \Delta_{n-1}$. Let

$$c_{\min} = \min_{i \in [1:C]} \min_{j:(a_i)_j \neq 0} (a_i)_j \tag{4.5.32}$$

be the minimal nonzero component of a stationary point. Then the method converges to a point a_l and identifies its support in at most $q(\bar{\varepsilon}) + |I^c(a_l)|$ iterations, where here $\bar{\varepsilon}$ has no explicit dependence on n:

$$\bar{\varepsilon} < L, \quad r(\bar{\varepsilon}) + l(\bar{\varepsilon}) < \min(r_*, c_{\min}/2) \ ,$$

where $r(\bar{\varepsilon}) = \left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}}$ and $l(\bar{\varepsilon}) = 2\sqrt{\frac{2\bar{\varepsilon}}{L}}$. We do not discuss the proof since it roughly follows the same lines of arguments leading to the proof of Theorem 4.5.6.

4.5.4 Local active set complexity bound

A key hypothesis to ensure local convergence to a strict local minimum is

$$x_k \in \operatorname{argmax}\{f(x) \mid x \in \operatorname{conv}(x_k, x_{k+1})\} \ . \tag{4.5.33}$$

which in particular holds when $\alpha_k = \bar{\alpha}_k$ as it is proved in Lemma 4.2.1. The property (4.5.33) is obviously stronger than the usual monotonicity, and it ensures that the sequence cannot escape from connected components of sublevel sets. When f is convex it is immediate to check that (4.5.33) holds if and only if $\{f(x_k)\}$ is monotone non increasing.

Let x^* be a stationary point which is also a strict local minimizer isolated from the other stationary points and $\tilde{f} = f(x^*)$. Let then β be such that there exists a connected component $V_{x^*,\beta}$ of $f^{-1}((-\infty, \beta])$ satisfying

$$V_{x^*,\beta} \cap \mathcal{X}^* = \{x^*\} = \operatorname*{arg\,min}_{x \in V_{x^*,\beta}} f(x) \, .$$

Theorem 4.5.9. *Let $x_0 \in V_{x^*,\beta}$, and let $\{x_k\}$ be the sequence generated by the AFW with step size $\alpha_k = \bar{\alpha}_k$. Let*

$$r_* = \frac{\delta_{\min}(x^*)}{2L + \delta_{\min}(x^*)} \, .$$

Then $x_k \to x^*$ and the sequence identifies the support in at most

$$\left\lceil \max\left(\frac{4(f(x_0) - f(x^*))}{\tau}, \frac{8L(f(x_0) - f(x^*))}{\tau^2} \right) \right\rceil + n$$

steps with

$$\tau = \min\{g(x) \mid x \in f^{-1}([m, +\infty)) \cap V_{x^*,\beta}\} \, ,$$

where

$$m = \min\{f(x) \mid x \in V_{x^*,\beta} \setminus B_{r_*}(x^*)\} \, .$$

Proof. As in the proof of Corollary 4.5.2, the assumptions of Theorem 4.5.1 are satisfied with $\rho = \frac{1}{2}$. By point 1 of Lemma 4.2.1, the condition $\alpha_k = \bar{\alpha}_k$ on the step sizes implies that $\{x_k\}$ satisfies (4.5.33). In particular, $\{x_k\}$ can not leave connected components of level sets so that $\{x_k\} \subset V_{x^*,\beta}$ and

$$\lim_{k \to \infty} f(x_k) \geq f(x^*) \, .$$

By (4.5.7) and (4.5.9) it follows

$$f(x_0) - f(x^*) \geq [n_1(T) + n_2(T)] \min\left(\frac{(g_T^*)^2}{4L}, \frac{g_T^*}{2} \right) . \tag{4.5.34}$$

Moreover applying (4.5.8) we obtain

$$n_1(T) + n_2(T) \geq \frac{T + |S_T| - |S_0|}{2} \geq \frac{T - n + 1}{2} \tag{4.5.35}$$

where the second inequality follows from $|S_T|-|S_0| \geq -n+1$. Concatenating (4.5.34) and (4.5.35) we get

$$f(x_0) - f(x^*) \geq \frac{T-n+1}{2} \min\left(\frac{(g_T^*)^2}{4L}, \frac{g_T^*}{2}\right) \tag{4.5.36}$$

from which we have the following bound on g_T^*:

$$g_T^* \leq \max\left(\sqrt{\frac{8L(f(x_0) - f(x^*))}{T-n+1}}, \frac{4(f(x_0) - f(x^*))}{T-n+1}\right) \tag{4.5.37}$$

for $T \geq n$. It is now straightforward to check that if

$$\bar{h} = \left\lceil \max\left(\frac{4(f(x_0) - f^*)}{\tau}, \frac{8L(f(x_0) - f^*)}{\tau^2}\right) \right\rceil + n\,,$$

then

$$g_{\bar{h}}^* < \tau\,.$$

Since (4.5.34) is derived considering the gap g only in case 1 and case 2 indexes, we have that there exists $\tilde{h} \leq \bar{h}$ case 1 or case 2 index such that $g(x_{\tilde{h}}) < \tau$. Therefore, by the definition of τ, we get $f(x_{\tilde{h}}) < m$. We claim that $x_h \in B_{r_*}(x^*)$ for every $h \geq \tilde{h}$. Indeed, since $f(x_{\tilde{h}}) < m$ and $\{x_k\}$ can not leave connected components of level sets we have for every $h \geq \tilde{h}$

$$x_h \in V_{x^*,\beta} \cap f^{-1}((-\infty, m)) \subset B_{r_*}(x^*)\,,$$

where the inclusion follows directly from the definition of m. Since the index $\tilde{h}$ corresponds to a case 1 or a case 2 step done in the active set region $B_{r_*}(x^*)$ by the local identification Theorem 4.3.3 the method must have already done all the case 3 steps needed to identify $I^c(x^*)$. Then we obtain the active set complexity bound

$$\tilde{h} \leq \bar{h} = \left\lceil \max\left(\frac{4(f(x_0) - f^*)}{\tau}, \frac{8L(f(x_0) - f^*)}{\tau^2}\right) \right\rceil + n\,, \tag{4.5.38}$$

as desired. □

4.6 AFW complexity for generic polytopes

It is well known as anticipated in the introduction that every application of the AFW to a polytope can be seen as an application of the AFW to the probability simplex. Even though rewriting an optimization problem on the simplex can lead to a dramatic increase in complexity, this equivalence is still useful because it allows us to extend the properties we proved on the simplex to generic polytopes. Furthermore, in practice the AFW only needs a linear minimization oracle and the points appearing in the convex combination giving the current iterate [157], while knowledge of the whole transformation between the polytope and the simplex is not needed.

In this section we show the connection between the active set and the face of the polytope exposed by $-\nabla f(y^*)$, where y^* is a stationary point for f. We then proceed to show with a couple of examples how the results proved for the probability simplex can be adapted to general polytopes. In particular we generalize Theorem 4.4.3, thus proving that under a convergence assumption the AFW identifies the face exposed by the gradients of some stationary points. An analogous result is already well known for the gradient projection algorithm, and was first proved in [61] building on [60] which used an additional strict complementarity assumption but worked in a more general setting than polytopes, that of convex compact sets with a polyhedral optimal face.

Before stating the generalized theorem we need to introduce additional notation and prove a few properties mostly concerning the generalization of the simplex multiplier function λ to polytopes.

Let P be a polytope and $f : P \rightarrow \mathbb{R}^n$ be a function with gradient having Lipschitz constant L.

To define the AFW algorithm we need a finite set of atoms $\mathcal{A}$ such that $\mathrm{conv}(\mathcal{A}) = P$. As for the probability simplex we can then define for every $a \in \mathcal{A}$ the multiplier function $\lambda_a : P \rightarrow \mathbb{R}$ by

$$\lambda_a(y) = \nabla f(y)^\top (a - y) \ .$$

Finally, let A be a matrix having as columns the atoms in $\mathcal{A}$, so that A is also a linear transformation mapping $\Delta_{|\mathcal{A}|-1}$ in P with $Ae_i = A^i \in \mathcal{A}$ (but the same results hold with the same proofs if we have an affine transformation $e_i \rightarrow Ae_i + b$).

In order to apply Theorem 4.3.3 we need to check that the transformed problem

$$\min\{f(Ax) \mid x \in \Delta_{|\mathcal{A}|-1}\}$$

still has all the necessary properties under the assumptions we made on f.

Let $\tilde{f}(x) = f(Ax)$. First, it is easy to see that the gradient of $\tilde{f}$ is still Lipschitz. Also λ is invariant under affine transformation, meaning that $\lambda_{A^i}(Ax) = \lambda_i(x)$ for every $i \in [1 : |\mathcal{A}|]$, $x \in \Delta_{|\mathcal{A}|-1}$. Indeed,

$$\lambda_{A^i}(Ax) = \nabla f(Ax)^\top (A^i - Ax) = \nabla f(Ax)^\top A(e_i - x) = \nabla \tilde{f}(x)^\top (e_i - x) = \lambda_i(x) \ .$$

Let Y^* be the set of stationary points for f on P, so that by invariance of multipliers $\mathcal{X}^* = A^{-1}(Y^*)$ is the set of stationary points for $\tilde{f}$. The invariance of the identification property follows immediately from the invariance of λ: if the support of the multiplier functions for f restricted to B is $\{A^i\}_{i \in I^c}$, then the support of the multiplier functions for $\tilde{f}$ restricted to $A^{-1}(B)$ is I^c.

We now show the connection between the face exposed by $-\nabla f$ and the support of the multiplier function. Let $y^* = Ax^* \in Y^*$ and let

$$\begin{aligned} P^*(y^*) &= \{y \in P \mid \nabla f(y^*)^\top y = \nabla f(y^*)^\top y^*\} \\ &= \mathrm{argmax}\{-\nabla f(y^*)^\top y \mid y \in P\} = \mathcal{F}_e(-\nabla f(y^*)) \end{aligned}$$

be the face of the polytope P exposed by $-\nabla f(y^*)$. The complementarity conditions for the generalized multiplier function λ can be stated very simply in terms of inclusion in $P^*(y^*)$: since $y^* \in P^*(y^*)$ we have $\lambda_a(y^*) = 0$ for every $a \in P^*(y^*)$, $\lambda_a(y^*) > 0$ for every $a \notin P^*(y^*)$. But P is the convex hull of the set of atoms in $\mathcal{A}$ so that the previous relations mean that the face $P^*(y^*)$ is the convex hull of the set of atoms for which $\lambda_a(y^*) = 0$:

$$P^*(y^*) = \mathrm{conv}\{a \in \mathcal{A} \mid \lambda_a(y^*) = 0\}$$

or in other words since $\lambda_{A^i}(y^*) = 0$ if and only if $i \in I(x^*) = \{i \in [1:n] \mid \lambda_i(x^*) = 0\}$:

$$P^*(y^*) = \mathrm{conv}\{a \in \mathcal{A} \mid a = A^i,\ i \in I(x^*)\}\ . \tag{4.6.1}$$

A consequence of (4.6.1) is that given any subset B of P with the SIP, we necessarily get $P^*(w) = P^*(z)$ for every $w, z \in B$, since $I(w) = I(z)$. For such a subset B we can then define

$$P^*(B) = P^*(y^*) \text{ for any } y^* \in B$$

where the definition does not depend on the specific $y^* \in B$ considered. We can now restate Theorem 4.4.3 in slightly different terms:

Theorem 4.6.1. *Let $\{y_k\}$ be a sequence generated by the AFW on P and let $\{x_k\}$ be the corresponding sequence of weights in $\Delta_{|\mathcal{A}|-1}$ such that $\{y_k\} = \{Ax_k\}$. Assume that the step sizes satisfy $\alpha_k \geq \bar{\alpha}_k$ (using $\tilde{f}$ instead of f in* (4.2.1)*). If there exists a compact subset B of Y^* with the SIP such that $y_k \to B$, then there exists M such that*

$$y_k \in P^*(B) \textit{ for every } k \geq M.$$

Proof. Follows from Theorem 4.4.3 and the affine invariance properties discussed above. □

In Theorem 4.6.1, in order to compute $\bar{\alpha}_k$ the Lipschitz constant L of $\nabla\tilde{f}$ (defined on the simplex) is necessary. When optimizing on a general polytope, the calculation of an accurate estimate of L for $\tilde{f}$ may be problematic. However, by Lemma 4.2.1 if the AFW uses exact line search, the step size $\bar{\alpha}_k$ (and in particular the constant L) is not needed because the inequality $\alpha_k \geq \bar{\alpha}_k$ is automatically satisfied.
We now generalize the analysis of the strongly convex case. The technical problem here is that strong convexity, which is used in Corollary 4.4.5, is not maintained by affine transformations, so that instead we have to use a weaker error bound condition. As a possible alternative, in [157] linear convergence of the AFW is proved with dependence only on affine invariant parameters, so that any version of Theorem 4.3.3 and Corollary 4.4.5 depending on those parameters instead of u_1, L would not need this additional analysis.
Let $P = \{y \in \mathbb{R}^n \mid Cy \leq b\}$, y^* be the unique minimizer of f on P and $u > 0$ be such that

$$f(y) \geq f(y^*) + \frac{u}{2}\|y - y^*\|^2 \ .$$

The function $\tilde{f}$ inherits the error bound condition necessary for Corollary 4.4.5 from the strong convexity of f: for every $x \in \Delta_{|\mathcal{A}|-1}$ by [27], Lemma 2.2 we have

$$\operatorname{dist}(x, \mathcal{X}^*) \leq \theta\|Ax - y^*\|$$

where θ is the Hoffman constant related to $[C^T, [I; e; -e]^T]^T$. As a consequence if $\tilde{f}^*$ is the minimum of $\tilde{f}$

$$\tilde{f}(x) - \tilde{f}^* = f(Ax) - f(y^*) \geq \frac{u}{2}\|Ax - y^*\|^2 \geq \frac{u}{2\theta^2}\operatorname{dist}(x, \mathcal{X}^*)^2$$

and using that $n\|\cdot\|^2 \geq \|\cdot\|_1^2$ we can finally retrieve an error bound condition with respect to $\|\cdot\|_1$:

$$\tilde{f}(x) - \tilde{f}^* \geq \frac{u}{2n\theta^2}\operatorname{dist}_1(x, \mathcal{X}^*)^2. \tag{4.6.2}$$

Having proved this error bound condition for $\tilde{f}$ we can now generalize (4.3.5):

Corollary 4.6.2. *The sequence $\{y_k\}$ generated by the AFW is in $P^*(y^*)$ for*

$$k \geq \max\left(0, \frac{\ln(h_0) - \ln(u_P r_*^2/2)}{\ln(1/q)}\right) + |I^c|$$

where $q \in (0,1)$, is the constant related to the linear convergence rate $f(y_k) - f(y^*) \leq q^k(f(y_0) - f(y^*))$, $u_P = \frac{u}{2n\theta^2}$, $r_* = \frac{\delta_{\min}}{2L+\delta_{\min}}$ with $\delta_{\min} = \min\{\lambda_a(y^*) \mid \lambda_a(y^*) > 0\}$.

Proof. Let $I = \{i \in [1 : |\mathcal{A}|] \mid \lambda_{A^i}(y^*) = 0\}$, $P^* = P^*(y^*)$. Since $P^* = \mathrm{conv}(\mathcal{A} \cap P^*)$ and by (4.6.1) $\mathrm{conv}(\mathcal{A} \cap P^*) = \mathrm{conv}\{A^i \mid i \in I\}$ the theorem is equivalent to prove that for every k larger than the bound, we have $y_k \in \mathrm{conv}\{A^i \mid i \in I\}$. Let $\{x_k\}$ be the sequence generate by the AFW on the probability simplex, so that $y_k = Ax_k$. We need to prove that, for every k larger than the bound, we have

$$x_k \in \mathrm{conv}\ \{e_i \mid i \in I\}\ ,$$

or in other words $(x_k)_i = 0$ for every $i \in I^c$.
Reasoning as in Corollary 4.4.5 we get that $\mathrm{dist}_1(x_k, \mathcal{X}^*) < r_*$ for every

$$k \geq \frac{\ln(h_0) - \ln(u_P r_*^2/2)}{\ln(1/q)}\ . \tag{4.6.3}$$

Let $\bar{k}$ be the minimum index such that (4.6.3) holds. For every $k \geq \bar{k}$ there exists $x^* \in \mathcal{X}^*$ with $\|x_k - x^*\|_1 < r_*$. But $\lambda_i(x) = \lambda_{A^i}(y^*)$ for every $x \in \mathcal{X}^*$ by the invariance of λ, so that we can apply Theorem 4.3.3 with fixed point x^* and obtain that if $J_k = \{i \in I^c \mid (x_k)_i > 0\}$ then $J_{k+1} \leq \max(0, J_k - 1)$. The conclusion follows exactly as in Corollary 4.4.5. □

Chapter 5

Fast Cluster Detection in Networks with a FW variant

Cluster detection plays a fundamental role in the analysis of data. In this chapter, we focus on the use of s-defective clique models for network-based cluster detection and propose a nonlinear optimization approach that efficiently handles those models in practice. In particular, we introduce an equivalent continuous formulation for the problem under analysis, and we analyze some tailored variants of the FW algorithm that enable us to quickly find maximal s-defective cliques. The good practical behavior of those algorithmic tools, which is closely connected to their support identification properties, makes them very appealing in practical applications. The reported numerical results clearly show the effectiveness of the proposed approach. [1]

5.1 A continuous optimization approach for maximum s-defective clique

In the context of network analysis the clique model, dating back at least to the work of Luce and Perry [177] about social networks, refers to subsets with every two elements in a direct relationship. The problem of finding maximal cliques has numerous applications in domains including telecommunication networks, biochemistry,

[1] This chapter is based on the article "Fast Cluster Detection in Networks by First Order Optimization" in *SIAM Journal on Mathematics of Data Science, vol. 4, iss. 1, pp. 285-305, 2022* [49].

financial networks, and scheduling ([43], [241]). From an optimization perspective, this problem has been the subject of extensive studies stimulating new research directions in both continuous and discrete optimization (see, e.g., [41], [43], [45], [217]). The Motzkin-Straus quadratic formulation [188] in particular has motivated several algorithmic approaches (see [40], [133] and references therein) to the maximum clique problem, beside being of independent interest for its connection with Turán's theorem [7].

Since the strict requirement that every two elements have a direct connection is often not satisfied in practice, many relaxations of the clique model have been proposed (see, e.g., [199] for a survey). We are here interested in s-defective cliques ([76], [224], [247]), allowing up to s links to be missing, and introduced in [247] for the analysis of protein interaction networks obtained with large scale techniques subject to experimental errors.

In this chapter, we first define a regularized version of a cubic continuous formulation for the maximum s-defective clique problem proposed in [217], and then apply variants of the classic FW method [101] to this formulation.

The support identification properties of FW variants are especially suited for our maximal s-defective clique formulation, since in this case the optimization process can stop as soon as the support of a solution is identified.

5.1.1 Problem formulation

For a vector $r \in \mathbb{R}^d$, the d-dimensional Euclidean space, and a set $A \subset [1:d]$, we denote with r_A the components of r with indexes in A. Let $\mathcal{G} = (V, E)$ be a graph with vertices V and and edges E, $n = |V|$, $A_{\mathcal{G}}$ the adjacency matrix of $\mathcal{G}$, and let $\bar{\mathcal{G}} = (V, \bar{E})$ the complementary graph. Recall that the Motzkin-Strauss formulation for the maximum clique problem is

$$\max\{x^\top A_{\mathcal{G}} x \mid x \in \Delta_{n-1}\}\,. \tag{MS}$$

We now introduce the cubic continuous formulation for the $s-$defective clique problem given in [217]. For $s \in \mathbb{N}$ with $s \leq |\bar{E}|$ we define

$$D_s(\mathcal{G}) = \{y \in \{0,1\}^{\bar{E}} \mid e^\intercal y \leq s\}\,,$$

representing the set of "fake edges" to be added to the graph in order to complete an s-defective clique, and its continuous relaxation as

$$D'_s(\mathcal{G}) = \{y \in [0,1]^{\bar{E}} \mid e^\intercal y \leq s\}\,.$$

For $y \in D'_s(\mathcal{G})$ we define the induced adjacency matrix $A(y) \in \mathbb{R}^{n \times n}$ as

$$A(y)_{ij} = \begin{cases} y_{ij} & \text{if } \{i, j\} \in \bar{E}\,, \\ 0 & \text{if } \{i, j\} \notin E\,. \end{cases}$$

For $y \in D_s(\mathcal{G})$ in particular we define $\mathcal{G}(y)$ as the graph with adjacency matrix $A_{\mathcal{G}} + A(y)$, that is the graph where we add to $\mathcal{G}$ the edge $\{i, j\}$ whenever $y_{ij} = 1$. We also define $E(i)$ and $E^y(i)$ as the neighbors of i in $\mathcal{G}$ and $\mathcal{G}(y)$ respectively.
Let $\mathcal{P}_s = \Delta_{n-1} \times D'_s(\mathcal{G})$. The objective of the s-defective clique relaxation defined in [217] is

$$f_{\mathcal{G}}(z) = f_{\mathcal{G}}(x, y) := x^\intercal [A_{\mathcal{G}} + A(y)]x\,, \quad z = (x, y) \in \mathcal{P}_s \tag{5.1.1}$$

so that when $A(y) = 0$ one retrieves Motzkin-Straus quadratic objective. The corresponding formulation for the maximum $s-$defective clique problem is then

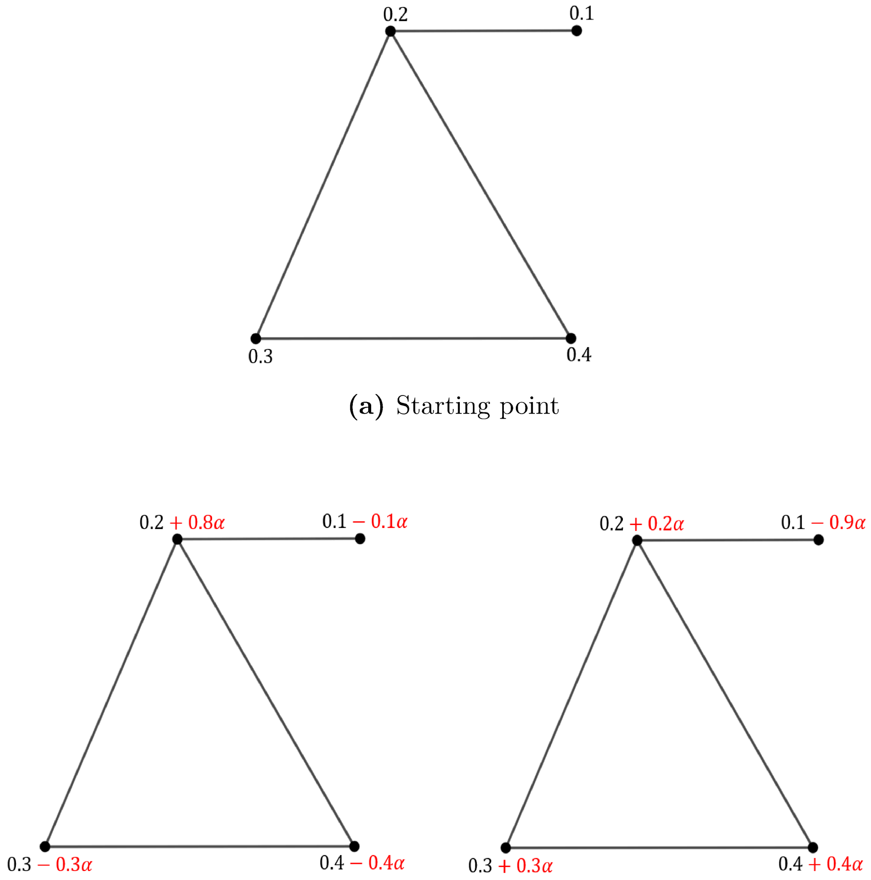
$$\max\{f_{\mathcal{G}}(z) \mid z \in \mathcal{P}_s\}\,. \tag{S}$$

5.1.2 Contributions

Our contributions can be summarized as follows:

- We solve the spurious solution problem for the maximum s-defective clique formulation proposed in [217] by introducing a regularized version, for which we prove equivalence between local maximizers and maximal s-defective cliques. In particular, no postprocessing algorithms are needed to derive the desired structure from a local solution. Our work develops along the lines of analogous results proved for regularized versions of the Motzkin - Straus quadratic formulation ([43], [133]).
- We prove that the FDFW applied to our formulation identifies the support of a maximal s-defective clique in a finite number of iterations.
- We propose a tailored Frank-Wolfe variant for the s-defective clique formulation at hand exploiting its product domain structure. This method retains the identification properties of the FDFW while significantly outperforming it in numerical tests.

The codes of the methods described in the chapter, together with the tested instances, are available at the following link: `https://github.com/DamianoZeffiro/s_defective_fw`.



(a) Starting point

(b) FW step

(c) Away step

Figure 5.1: FDFW for an instance of problem (MS)

5.2 A regularized maximum s-defective clique formulation

Here we consider the problem

$$\max\{h_{\mathcal{G}}(z) \mid z \in \mathcal{P}_s\}\,, \tag{P}$$

where $h_{\mathcal{G}} : \mathcal{P}_s \longrightarrow \mathbb{R}_{>0}$ is a regularized version of $f_{\mathcal{G}}$:

$$h_{\mathcal{G}}(z) = h_{\mathcal{G}}(x, y) := f_{\mathcal{G}}(x, y) + \frac{\alpha}{2}\|x\|^2 + \frac{\beta}{2}\|y\|^2$$

for some $\alpha \in (0, 2)$ and $\beta > 0$. In particular, when $y = 0$ the objective $h_{\mathcal{G}}$ corresponds to the quadratic regularized maximal clique formulation introduced in [40]. As we shall see in Proposition 5.2.1, the main advantage of the regularized objective $h_{\mathcal{G}}$ is that, in sharp contrast to $f_{\mathcal{G}}$, it does not admit any spurious local solutions, i.e., the support of the x component of *every* local maximizer $p = (x, y)$ of $h_{\mathcal{G}}$ (i.e., a maximizer in a neighborhood $U \subseteq \mathcal{P}_s$ of p) is a maximal s-defective clique.

For non-empty $C \subseteq V$ let $x^{(C)} = \frac{1}{|C|} \sum_{i \in C} e_i$ be the characteristic vector in Δ_{n-1} of the clique C, and

$$\Delta^{(C)} = \{x \in \Delta_{n-1} \mid x_i = 0 \text{ for all } i \in V \setminus C\}$$

be the minimal face of Δ_{n-1} containing $x^{(C)}$ in its relative interior.

For $p \in \mathcal{P}_s$ we define as $T_{\mathcal{P}_s}(p) = \{v - p : v \in \mathcal{P}_s\}$ as the cone of feasible directions at p in $\mathcal{P}_s$, while for $r \in \mathbb{R}^{|V|+|\bar{E}|}$ we define $T^0_{\mathcal{P}_s}(p, r)$ as the intersection between $T_{\mathcal{P}_s}(p)$ and the plane orthogonal to r:

$$T^0_{\mathcal{P}_s}(p, r) = \{d \in T_{\mathcal{P}_s}(p) \mid d^\top r = 0\} .$$

We now prove that (i) every local maximizer of $h_{\mathcal{G}}$ is strict and that (ii) there is a one-to-one correspondence between (strict) local maximizers of $h_{\mathcal{G}}$ and s-defective cliques coupled together with s fake edges including the one missing on the clique. Recall that in our polytope-constrained setting, (second order) sufficient conditions for the local maximality of $p \in \mathcal{P}_s$ are (see, e.g., [30])

$$\nabla h_{\mathcal{G}}(p)^\top d \leq 0 \text{ for all } d \in T_{\mathcal{P}_s}(p) \tag{5.2.1}$$

and

$$d^\intercal D^2 h_{\mathcal{G}}(p) d < 0 \text{ for all } d \in T^0_{\mathcal{P}_s}(p, \nabla h_{\mathcal{G}}(p)) . \tag{5.2.2}$$

In the rest of the chapter we use $\mathcal{M}_s(\mathcal{G})$ to denote the set of strict local maximizers of $h_{\mathcal{G}}$.

Proposition 5.2.1 (characterization of local maxima for $h_{\mathcal{G}}$)**.** *The following are equivalent:*

(i) $p \in \mathcal{P}_s$ *is a local maximizer for* $h_{\mathcal{G}}(x, y)$*;*

(ii) $p \in \mathcal{M}_s(\mathcal{G})$*;*

(iii) $p = (x^{(C)}, y^{(p)})$ *where* $s = e^\intercal y^{(p)} \in \mathbb{N}$*, with* C *an* s*-defective clique in* $\mathcal{G}$ *which is also a maximal clique in* $\mathcal{G}(y^{(p)})$*, and* $y^{(p)} \in D_s(\mathcal{G})$ *such that* $y^{(p)}_{ij} = 1$ *for every* $\{i, j\} \in \binom{C}{2} \cap \bar{E}$*.*

In either of these equivalent cases, we have

$$h_{\mathcal{G}}(p) = 1 - \frac{2-\alpha}{2|C|} + s\frac{\beta}{2}\,. \tag{5.2.3}$$

Proof. Let $p = (x^{(p)}, y^{(p)}) \in \mathcal{P}_s$, $g = \nabla h_{\mathcal{G}}(p), H = D^2 h_{\mathcal{G}}(p)$.
(ii) $\Rightarrow$ (i) is trivial.
(i) $\Rightarrow$ (iii). If $s := e^\intercal y^{(p)}$ were fractional, then for some $\{i,j\} \in \bar{E}$ we would have $y^{(p)}_{ij} < 1$. Furthermore

$$\frac{\partial h_{\mathcal{G}}(p)}{\partial y_{ij}} = 2x^{(p)}_i x^{(p)}_j + \beta y^{(p)}_{ij} \geq 0, \quad \frac{\partial h_{\mathcal{G}}(p)}{\partial^2 y_{ij}} = \beta > 0\,. \tag{5.2.4}$$

Thus for $\varepsilon > 0$ small enough we have $h_{\mathcal{G}}(p+\varepsilon e_{ij}) > h_{\mathcal{G}}(p)$ with $p+\varepsilon e_{ij} \in \mathcal{P}_s$, which means that p is not a local maximizer. Hence $s \in \mathbb{N}$ and obviously $s \leq |\bar{E}|$ as well as $y^{(p)} \in D'_s(\mathcal{G})$.
Assume now by contradiction that p is a local maximizer but $y^{(p)} \notin D_s(\mathcal{G})$. Then for two distinct edges $\{i,j\}$, $\{l,m\} \in \bar{E}$ we must have $y^{(p)}_{ij}, y^{(p)}_{lm} \in (0,1)$. Let $d = (0, e_{ij} - e_{lm})$. Since $\pm d$ are both feasible directions and p is a local maximizer, necessarily $g^\top d = 0$. But we also have

$$d^\intercal H d = \frac{\partial h_{\mathcal{G}}(p)}{\partial^2 y_{ij}} + \frac{\partial h_{\mathcal{G}}(p)}{\partial^2 y_{lm}} - 2\frac{\partial h_{\mathcal{G}}(p)}{\partial y_{ij}\partial y_{lm}} = 2\beta > 0\,. \tag{5.2.5}$$

so that again for $\varepsilon > 0$ small enough $h_{\mathcal{G}}(p+\varepsilon d) > h_{\mathcal{G}}(p)$ with $p + \varepsilon d \in \mathcal{P}_s$, a contradiction.
We proved that if p is a local maximizer, then $s = e^\intercal y^{(p)} \in \mathbb{N}$ and $y^{(p)} \in D_s(\mathcal{G})$. But $x^{(p)}$ must be a local maximizer for the function $x \mapsto h_{\mathcal{G}}(x, y^{(p)})$, which is (up to a constant) a regularized maximal clique relaxation for the augmented graph $\mathcal{G}(y^{(p)})$. By the characterization of local maximizers for this function given in [133, Proposition 2.2] (see also [40, Theorem 9]) we must have $x = x^{(C)}$ with C a maximal clique in $\mathcal{G}(y^{(p)})$. In particular, since $\mathcal{G}(y^{(p)})$ is defined by adding s edges to $\mathcal{G}$, C must be an s-defective clique in $\mathcal{G}$.
(iii) $\Rightarrow$ (ii). For a fixed $p = (x^{(C)}, y^{(p)})$ with $C, y^{(p)}$ satisfying the conditions of point (iii) let $\bar{C} = V \setminus C$, $S = \mathrm{supp}(y^{(p)})$ and $\bar{S} = \bar{E} \setminus S$. We abbreviate $E^{(p)}(i) = E^y(i)$ with $y = y^{(p)}$. For every $i \in V$ we have

$$g_i = \alpha x^{(C)}_i + \sum_{j \in E^{(p)}(i)} 2x^{(C)}_j \tag{5.2.6}$$

In particular for $i \in C$

$$g_i = \frac{\alpha}{|C|} + \sum_{j \in C\setminus\{i\}} 2x^{(C)}_j = \frac{1}{|C|}(\alpha + 2|C| - 2) \tag{5.2.7}$$

and for every $i \in \bar{C}$

$$g_i = \sum_{j \in E^{(p)}(i) \cap C} 2x_j^{(C)} \leq \frac{2|C| - 2}{|C|} \tag{5.2.8}$$

where we used $x_j^{(C)} = 1/|C|$ for every $j \in C$, $x_j^{(C)} = 0$ otherwise.
For $\{i, j\} \in \bar{E}$ we have

$$g_{ij} = \beta y_{ij}^{(p)} + 2x_i^{(C)} x_j^{(C)} \tag{5.2.9}$$

and in particular $g_{ij} = 0$ for $\{i, j\} \in \bar{S}$, while for $\{i, j\} \in S$

$$g_{ij} = \beta + 2x_i^{(C)} x_j^{(C)} \geq \beta > 0\,, \tag{5.2.10}$$

where we used $y_{ij}^{(p)} = 1$ for $\{i, j\} \in S$, 0 otherwise, and $x_i^{(C)} x_j^{(C)} = 0$ for $\{i, j\} \in \bar{S} \subseteq \bar{E}$. Let d be a feasible direction from p, so that $d = v - p$ with $v \in \mathcal{P}_s$. Let $\sigma_S = \sum_{\{i,j\} \in S} g_{ij}$, $\sigma_C = \sum_{i \in C} v_i$
$= 1 - \sum_{i \in \bar{C}} v_i \in [0, 1]$, $m_{\bar{C}} = \max_{i \in \bar{C}} g_i$, so that by (5.2.8) we have $m_{\bar{C}} \leq \frac{2|C|-2}{|C|}$. Then

$$g^\top p = \sum_{i \in \bar{C}} x_i^{(C)} g_i + \sum_{i \in C} x_i^{(C)} g_i + \sum_{(i,j) \in S} y_{ij}^{(p)} g_{ij} = \frac{1}{|C|} \sum_{i \in C} g_i + \sum_{\{i,j\} \in S} g_{ij} = \frac{1}{|C|}(\alpha + 2|C| - 2) + \sigma_S \tag{5.2.11}$$

where we used (5.2.7) in the last equality. We also have

$$g_V^\top v_V = g_C^\top v_C + g_{\bar{C}}^\top v_{\bar{C}} \leq \frac{\alpha + 2|C| - 2}{|C|} \sigma_C + (1 - \sigma_C) m_{\bar{C}} \leq \frac{\alpha + 2|C| - 2}{|C|} \tag{5.2.12}$$

where we used (5.2.7) together with the Hölder inequality in the first inequality, $m_{\bar{C}} \leq \frac{2|C|-2}{|C|}$ in the second inequality and $\sigma_C \leq 1$. Finally,

$$g_{\bar{E}}^\top v_{\bar{E}} = g_S^\top v_S + g_{\bar{S}}^\top v_{\bar{S}} = g_S^\top v_S \leq \sigma_S \tag{5.2.13}$$

where we used $g_{\bar{S}} = 0$ in the second equality, and $v_i \leq 1$ for every $i \in \bar{E}$ in the inequality. We can conclude

$$g^\top d = g_V^\top v_V + g_{\bar{E}}^\top v_{\bar{E}} - g^\top p \leq 0 \tag{5.2.14}$$

where we used (5.2.13), (5.2.11) and (5.2.12) in the inequality. We have equality iff there is equality both in (5.2.12) and (5.2.13), and thus iff $v = (x^{(v)}, y^{(v)})$ with $\text{supp}(x^{(v)}) = C$ and $y^{(v)} = y^{(p)}$. In particular p is a first order stationary point with

$$T_{\mathcal{P}_s}^0(p, g) = \{d \in T_{\mathcal{P}_s}(p) \mid d = v - p, v_{\bar{C}} = 0, v_{\bar{E}} = p_{\bar{E}}\} = \{d \in T_{\mathcal{P}_s}(p) \mid d_{\bar{C}} = d_{\bar{E}} = 0\}\,. \tag{5.2.15}$$

Let H_C be the submatrix of H with indices in C. We have, for $(i, j) \in C^2$ with $i \neq j$, $H_{ij} = 1$ since C is a clique in the augmented graph $\mathcal{G}(y_p)$, while $H_{ii} = \alpha$ for every $i \in V$ and in particular for every $i \in C$. This proves

$$H_C = 2ee^\intercal + (\alpha - 2)\mathbb{I}\,. \tag{5.2.16}$$

Now if $T^0_{\mathcal{P}_s}(p, g) \ni d \neq 0$ we have

$$d^\intercal Hd = d_C^\intercal H_C d_C = d_C^\intercal(2ee^\intercal + (\alpha - 2)\mathbb{I})d_C = (\alpha - 2)\|d_C\|^2 < 0 \tag{5.2.17}$$

where we used $d_{\bar{C}} = d_{\bar{E}} = 0$ in the first equality, $e^\intercal d_C = e^\intercal(v_V - p_V) = 1 - 1 = 0$ in the third one. This proves the claim, since we have sufficient conditions for local optimality thanks to (5.2.14) and (5.2.17). □

As a corollary, the global optimum of $h_{\mathcal{G}}$ is achieved on maximum s-defective cliques.

Corollary 5.2.2. *The global maximizers of $h_{\mathcal{G}}(z)$ are all the points p of the form $p = (x^{C^*}, y^{(p)})$ where C^* is an s-defective clique of maximum cardinality, and $y^{(p)} \in D_s(\mathcal{G})$ such that $e^\intercal y^{(p)} = s$.*

Proof. Let $p = (x^{(C)}, y^{(p)})$ a local maximizer for $h_{\mathcal{G}}(z)$. Then its objective value is, by (5.2.3), $h_{\mathcal{G}}(p) = 1 - \frac{2-\alpha}{2|C|} + s\,\frac{\beta}{2}$, which is (globally) maximized when $|C|$ is as large as possible, because $2 - \alpha > 0$ by assumption. □

Thanks to Proposition 5.2.1, for every $p \in \mathcal{M}_s(\mathcal{G})$ we can define $y^{(p)} \in D_s(\mathcal{G})$ and a maximal clique C of $\mathcal{G}(y^{(p)})$ such that $p = (x^{(C)}, y^{(p)})$.
We now prove that the face of $\mathcal{P}_s$ exposed by the gradient in $p \in \mathcal{M}_s(\mathcal{G})$ is simply the product between $\Delta^{(C)}$ and the singleton $\{y^{(p)}\}$. This property, sometimes referred to as strict complementarity, is of key importance to prove identification results for Frank-Wolfe variants (see [46], [47], [107], and the discussion of external regularity in [42, Section 5.3]). We use it to prove a local identification and convergence result for the FDFW (see Theorem 5.3.1).

Lemma 5.2.3. *Let $p = (x^{(C)}, y^{(p)}) \in \mathcal{M}_s(\mathcal{G})$. Then the face exposed by $\nabla h_{\mathcal{G}}(p)$ coincides with the minimal face $\mathcal{F}(p)$ of $\mathcal{P}_s$ containing p:*

$$\mathcal{F}_e(\nabla h_{\mathcal{G}}(p)) = \mathcal{F}(p) = \Delta^{(C)}_{n-1} \times \{y^{(p)}\}\,. \tag{5.2.18}$$

Proof. To start with, the second equality follows from the fact that $y^{(p)}$ is a vertex of $D'_s(\mathcal{G})$ and that $\Delta^{(C)}_{n-1}$ is the minimal face of Δ_{n-1} containing $x^{(C)}$. The first equality is then equivalent to proving that for every vertex $a = (a_x, a_y)$ of $\mathcal{P}_s$ with $a \in \mathcal{P}_s \backslash \mathcal{F}(p)$ we have $\lambda_a(p) < 0$. Given that stationarity conditions must hold in Δ_{n-1} and $D'_s(\mathcal{G})$ separately, $\lambda_a(p) < 0$ if and only if

$$\lambda_a^x(p) := \nabla_x h_{\mathcal{G}}(p)^\top (a_x - x^{(C)}) \leq 0 \,, \tag{5.2.19a}$$

$$\lambda_a^y(p) := \nabla_y h_{\mathcal{G}}(p)^\top (a_y - y^{(p)}) \leq 0 \,, \tag{5.2.19b}$$

and at least one of these relations must be strict. Since a is a vertex of $\mathcal{P}_s$, $a_x = e_l$ with $l \in [1:n]$ and $a_y \in D_s(\mathcal{G})$, while $a \notin \mathcal{F}(p)$ implies $l \notin C$ or $a_y \neq y^{(p)}$. If $l \in C$ then $\lambda_a^x(p) = 0$ by stationarity conditions, otherwise

$$\nabla_x h_{\mathcal{G}}(p)^\top x^{(C)} = 2(x^{(C)})^\intercal [A + A(y^{(p)})] x^{(C)} + \alpha \|x^{(C)}\|^2 = 2 - \frac{2-\alpha}{|C|} \tag{5.2.20}$$

and

$$\nabla_x h_{\mathcal{G}}(p)^\top a_x = \frac{\partial}{\partial x_l} h_{\mathcal{G}}(p) = \alpha x_l + \sum_{j \in C \cap E^{(p)}(l)} 2x_j = 2\frac{|C \cap E^{(p)}(l)|}{|C|} \leq 2 - \frac{2}{|C|} \,, \tag{5.2.21}$$

where we used $a_x = e_l$ in the first equality, $l \notin C$ together with $x_j = 1/|C|$ for every $j \in C$ in the third equality, and the maximality of the clique C in the augmented graph $\mathcal{G}(y^{(p)})$ in the inequality. Combining (5.2.20) and (5.2.21), we obtain

$$\nabla_x h_{\mathcal{G}}(p)^\top (a_x - x^{(C)}) \leq -\frac{\alpha}{|C|} < 0 \,, \tag{5.2.22}$$

which proves that (5.2.19a) holds with strict inequality if $l \notin C$, or else with equality if $l \in C$.

In a similar vain we proceed with (5.2.19b). If $a_y = y^{(p)}$ then (5.2.19b) holds with equality but then $l \in V \setminus C$ and we are done. So suppose $a_y \neq y^{(p)}$, and consider the supports $S_y = \{\{i,j\} \in \bar{E} \mid (a_y)_{ij} = 1\}$ and $S_p = \{\{i,j\} \in \bar{E} \mid y_{ij}^{(p)} = 1\}$. Since $a_y \in D_s(\mathcal{G})$, we have $|S_y| \leq s$ and on the other hand, by Proposition 5.2.1(iii), $|S_p| = s$. As S_y and S_p must be distinct, we conclude $S_y \setminus S_p \neq \emptyset$. Furthermore, by (5.2.4) for every $\{i,j\}$ in A_p we have

$$\frac{\partial}{\partial y_{ij}} h_{\mathcal{G}}(p) \geq \beta y_{ij}^{(p)} = \beta > 0 \,, \tag{5.2.23}$$

while for every $\{i,j\}$ in $A_y \setminus A_p$ we have

$$\frac{\partial}{\partial y_{ij}} h_{\mathcal{G}}(p) = 0 \tag{5.2.24}$$

because $y_{ij}^{(p)} = 0$ by definition of A_p and $x_i^{(C)} x_j^{(C)} = 0$ since, again invoking Propositoin 5.2.1(iii), $\{i, j\} \in \bar{E} \setminus \binom{C}{2}$. So we can finally prove (5.2.19b) by observing

$$\nabla_y h_{\mathcal{G}}(p)^\top (a_y - y^{(p)}) = \sum_{\{i,j\}\in A_y} \frac{\partial}{\partial y_{ij}} h_{\mathcal{G}}(p) - \sum_{\{i,j\}\in A_p} \frac{\partial}{\partial y_{ij}} h_{\mathcal{G}}(p)$$

$$= \sum_{\{i,j\}\in A_y \setminus A_p} \frac{\partial}{\partial y_{ij}} h_{\mathcal{G}}(p) - \sum_{\{i,j\}\in A_p \setminus A_y} \frac{\partial}{\partial y_{ij}} h_{\mathcal{G}}(p) = - \sum_{\{i,j\}\in A_p \setminus A_y} \frac{\partial}{\partial y_{ij}} h_{\mathcal{G}}(p) < 0 \tag{5.2.25}$$

where we used (5.2.24) in the third equality and (5.2.23) together with $A_p \setminus A_y \neq \emptyset$ in the inequality. □

5.3 Frank-Wolfe method with in face directions

Let $Q = \text{conv}(A) \subset \mathbb{R}^n$ with $|A| < +\infty$. In this section, we consider the FDFW for the solution of the smooth constrained optimization problem

$$\max\{f(w) \mid w \in Q\} .$$

In particular, $\{w_k\}$ is always a sequence generated by the FDFW applied to the polytope Q with objective f. For $w \in Q$ we denote with $\mathcal{F}(w)$ the minimal face of Q containing w. The FDFW at every iteration chooses between the classic FW direction $d_k^{\mathcal{FW}}$ calculated at Step 2 and the in face direction $d_k^{\mathcal{FD}}$ calculated at Step 10 with the criterion in Step 12. When $f = h_{\mathcal{G}}$ and $Q = \mathcal{P}_s$, it is not difficult to see that the main cost to compute v_k is finding the smallest s components of a vector with size at most $|\bar{E}|$. After the algorithm performs an in face step, we have that the minimal face containing the current iterate either stays the same or its dimension drops by one. The latter case occurs when the method performs a maximal feasible in face step (i.e., a step with $\alpha_k = \alpha_k^{\max}$ and $d_k = d_k^{\mathcal{FD}}$), generating a point on the boundary of the current minimal face. As we prove formally in Proposition 5.3.3, this drop in dimension is what allows the method to quickly identify low dimensional faces containing solutions.

We often require the following lower bound on the stepsizes:

$$\alpha_k \geq \bar{\alpha}_k := \min(\alpha_k^{\max}, c\frac{\nabla f(w_k)^\top d_k}{\|d_k\|^2}) \tag{S1}$$

for some $c > 0$. Furthermore, for some convergence results we need the following sufficient increase condition for some constant $\rho > 0$:

$$f(w_k + \alpha_k d_k) - f(w_k) \geq \rho \bar{\alpha}_k \nabla f(w_k)^\top d_k . \tag{S2}$$

Algorithm 7 Frank-Wolfe method with in face directions (FDFW) on a polytope

```
Initialize $w_0 \in Q$, $k := 0$
Let $s_k \in \arg\max_{y \in Q} \nabla f(w_k)^\top y$ and $d_k^{\mathcal{FW}} := s_k - w_k$.
if $w_k$ is stationary then
    STOP
end if
Let $v_k \in \arg\min_{y \in \mathcal{F}(w_k)} \nabla f(w_k)^\top y$ and $d_k^{\mathcal{FD}} := w_k - v_k$.
if $\nabla f(w_k)^\top d_k^{\mathcal{FW}} \geq \nabla f(w_k)^\top d_k^{\mathcal{FD}}$ then
    $d_k := d_k^{\mathcal{FW}}$
else
    $d_k := d_k^{\mathcal{FD}}$
end if
Choose the stepsize $\alpha_k \in (0, \alpha_k^{\max}]$ with a suitable criterion
Update: $w_{k+1} := w_k + \alpha_k d_k$
Set $k := k + 1$. Go to step 2.
```

These conditions generalize properties of exact and Armijo line search, as a corollary of the results in Section 4.2.2.

We now state a local convergence and identification result for the FDFW applied to our maximal s-defective clique formulation (P).

Theorem 5.3.1 (FDFW local identification and convergence)**.** *Let $p = (x^{(C)}, y^{(p)}) \in \mathcal{M}_s(\mathcal{G})$, let $\{z_k\}$ be a sequence generated by the FDFW. Then under* (S1) *there exists a neighborhood $U(p)$ of p such that if $\bar{k} := \min\{k \in \mathbb{N}_0 \mid z_k \in U(p)\}$ we have the following properties:*

(a) if $h_{\mathcal{G}}(z_k)$ is monotonically increasing, then $\mathrm{supp}(z_k) = C$ *and $y_k = y^{(p)}$ for every $k \geq \bar{k} + \dim \mathcal{F}(w_{\bar{k}})$;*

(b) if (S2) *also holds then $z_k \to p$.*

Before presenting the proof of Theorem 5.3.1, it is convenient to prove some generic convergence results for the FDFW. To start with, it is useful to define the multiplier functions λ_a for $a \in A$, $w \in \mathbb{R}^n$ as

$$\lambda_a(w) = \nabla f(w)^\top (a - w)\,. \tag{5.3.1}$$

We adapt FW gap to the maximization case, thus obtaining the following measure of stationarity

$$G(w) := \max_{y \in Q} \nabla f(w)^\top (w - y) = \max_{a \in A} \nabla f(w)^\top (w - a) = \max_{a \in A} -\lambda_a(w)\,, \tag{5.3.2}$$

as well as an "in face" gap

$$G_{\mathcal{F}}(w) = \max(G(w), \max_{b \in \mathcal{F}(w) \cap A} \lambda_b(w)) \,. \tag{5.3.3}$$

Using these definitions, we have

$$\begin{aligned} \nabla f(w_k)^\top d_k &= \max(\nabla f(w_k)^\top d_k^{\mathcal{FW}}, \nabla f(w_k)^\top d_k^{\mathcal{FD}}) \\ &= \max(G(w_k), \max_{y \in \mathcal{F}(w_k)} \nabla f(w_k)^\top (w_k - y)) = G_{\mathcal{F}}(w_k) \,, \end{aligned} \tag{5.3.4}$$

where in the second equality we used

$$\nabla f(w_k)^\top d_k^{\mathcal{FW}} = \max_{y \in Q} \nabla f(w_k)^\top (y - w_k) \tag{5.3.5}$$

and in the third equality

$$\nabla f(w_k)^\top d_k^{\mathcal{FD}} = \max_{b \in \mathcal{F}(w_k)} \nabla f(w_k)^\top (w_k - b) = \max_{b \in \mathcal{F}(w_k) \cap A} -\lambda_b(w_k) \,. \tag{5.3.6}$$

From the definitions it also immediately follows

$$G_{\mathcal{F}}(w) \geq G(w) \geq 0 \tag{5.3.7}$$

with equality iff w is a stationary point.
In order to obtain a local identification result, we need to prove that under certain conditions the method does consecutive maximal in face steps, thus identifying a low dimensional face containing a minimizer. First, in the following lemma we give an upper bound for the maximal feasible stepsize.

Lemma 5.3.2. *If w_k is not stationary, then $\alpha_k \leq G(w_k)/G_{\mathcal{F}}(w_k)$.*

Proof. Notice that since w_k is not stationary we have $G(w_k) > 0$ and therefore also $G_{\mathcal{F}}(w_k) > 0$. Now

$$\begin{aligned} &\nabla f(w_k)^\top (w_k + \alpha_k d_k) \leq \max_{y \in Q} \nabla f(w_k)^\top y = \nabla f(w_k)^\top (w_k + d_k^{\mathcal{FW}}) \\ &= \nabla f(w_k)^\top w_k + G(w_k) \,, \end{aligned}$$

where in the inequality we used $w_k + \alpha_k d_k \in Q$. Subtracting $\nabla f(w_k)^\top w_k$ on both sides we obtain

$$\alpha_k \nabla f(w_k)^\top d_k \leq G(w_k) \,. \tag{5.3.8}$$

and the thesis follows by applying (5.3.4) to the LHS. □

We can now prove a local identification result.

Proposition 5.3.3. *Let p be a stationary point for f defined on Q and assume that* (S1) *holds. We have the following properties:*

(a) there exists $r^*(p) > 0$ such that if $w_k \in B_{r^*(p)}(p) \cap \mathcal{F}_e(\nabla f(p))$ then $w_{k+1} \in \mathcal{F}_e(\nabla f(p))$;

(b) for any $\delta > 0$ there exists $r(\delta, p) \leq \delta$ such that if $w_k \in B_{r(\delta,p)}(p)$ then $w_{k+j} \in \mathcal{F}_e(\nabla f(p)) \cap B_\delta(p)$ for some $j \leq \dim(\mathcal{F}(w_k))$.

Proof. (a) Notice that by definition of exposed face and stationarity conditions

$$\lambda_a(p) \leq 0 \tag{5.3.9}$$

for every $a \in A$, with equality iff $a \in \mathcal{F}_e(\nabla f(p))$. Then by continuity we can take $r^*(p)$ small enough so that $\lambda_a(w) < 0$ for every $a \in A \setminus (A \cap \mathcal{F}_e(\nabla f(p)))$. Under this condition, if $w_k \in B_{r^*(p)}(p)$ then the method cannot select a FW direction pointing toward an atom outside the exposed face $\mathcal{F}_e(\nabla f(p))$, because all the atoms maximizing the RHS of (5.3.2) must necessarily be in $\mathcal{F}_e(\nabla f(p))$. In particular if $w_k \in B_{r^*(p)}(p) \cap \mathcal{F}_e(\nabla f(p))$ then the method selects either an in face direction or a FW direction pointing toward a vertex in $\mathcal{F}_e(\nabla f(p))$. In both cases, $w_{k+1} \in \mathcal{F}_e(\nabla f(p))$.

(b) Let D be the diameter of Q. We now consider $r^{(0)}(\delta, p) \leq \min(\delta, r^*(p))$ such that, for every $w \in B_{r^{(0)}(\delta,p)}(p)$

$$\max_{a \in A} \lambda_a(w) < \min_{b \in A \setminus \mathcal{F}_e(\nabla f(p))} \min(-\lambda_b(w), \frac{c}{D^2} \lambda_b(w)^2) . \tag{5.3.10}$$

As we will see in the rest of the proof this upper bound together with Lemma 5.3.2 ensures in particular that the FDFW performs maximal in face steps in $B_{r^{(0)}(\delta,p)}(p) \setminus \mathcal{F}_e(\nabla f(p))$. Furthermore, (5.3.10) can always be satisfied thanks to (5.3.9) and by the continuity of multipliers. We then define recursively for $1 \leq l \leq n$ a sequence $r^{(l)}(\delta, p) \leq r^{(l-1)}(\delta, p)$ of radii small enough so that, for

$$M_l = \sup_{w \in B_{(l)}(p) \setminus \mathcal{F}_e(\nabla f(p))} G(w)/G_{\mathcal{F}}(w) , \tag{5.3.11}$$

with $B_{(l)}(p) := B_{r^{(l)}(\delta,p)}(p)$ we have

$$r^{(l)}(\delta, p) + DM_l < r^{(l-1)}(\delta, p) . \tag{5.3.12}$$

Again this sequence can always be defined thanks to the continuity of multipliers. Finally, we define $r(\delta, p) = r^{(n)}(\delta, p)$.

Given these definitions, when $w_k \in B_{(l)}(p) \subset B_{(0)}(p)$ and w_k is not in $\mathcal{F}_e(\nabla f(p))$ an in face direction is selected, because

$$\begin{aligned} &\nabla f(w_k)^\top d_k^{\mathcal{FW}} = \max_{a \in A} \lambda_a(w) < \min_{b \in A \setminus \mathcal{F}_e(\nabla f(p))} -\lambda_b(w) \\ &\leq \max_{b \in \mathcal{F}(w_k) \cap A} -\lambda_b(w_k) = \nabla f(x_k)^\top d_k^{\mathcal{FD}} , \end{aligned} \tag{5.3.13}$$

where we used (5.3.10) in the first inequality, $w_k \notin \mathcal{F}_e(p)$ in the second, and (5.3.6) in the second equality. We now want to prove that in this case α_k is maximal reasoning by contradiction. On the one hand, we have

$$\alpha_k \geq c \frac{\nabla f(x_k)^\top d_k}{\|d_k\|^2} \geq \frac{c}{D^2} \nabla f(x_k)^\top d_k = \frac{c}{D^2} G_{\mathcal{F}}(w_k) \tag{5.3.14}$$

where we used the assumption (S1) in the first inequality, $\|d_k\| \leq D$ in the second and $G_{\mathcal{F}}(w_k) = \nabla f(x_k)^\top d_k^{\mathcal{FD}}$ together with $d_k = d_k^{\mathcal{FD}}$ in the last one.

On the other hand,

$$\begin{aligned} G(w_k) = \max_{a \in A} \lambda_a(w_k) <& \frac{c}{D^2} \min_{b \in A \setminus \mathcal{F}_e(\nabla f(p))} \lambda_b(w)^2 \leq \frac{c}{D^2} \max_{b \in \mathcal{F}(w_k)} \lambda_b(w)^2 \\ =& \frac{c}{D^2} (\nabla f(w_k)^\top d_k)^2 = \frac{c}{D^2} G_{\mathcal{F}}(w_k)^2 \end{aligned} \tag{5.3.15}$$

where we used (5.3.10) in the first inequality, $w_k \notin \mathcal{F}_e(\nabla f(p))$ in the second, (5.3.6) together with $d_k = d_k^{\mathcal{FD}}$ in the second equality, and (5.3.4) in the third equality.

The inequality (5.3.15) leads us to a contradiction with the lower bound on α_k given by (5.3.14), since it implies

$$\alpha_k \leq \frac{G(w_k)}{G_{\mathcal{F}}(w_k)} < \frac{c}{D^2} G_{\mathcal{F}}(w_k) , \tag{5.3.16}$$

where we applied Lemma 5.3.2 in the first inequality and (5.3.15) in the second.

Assume now $w_k \in B_{(n)}(p)$. We prove by induction that, for every $j \in [-1 : \dim(\mathcal{F}(w_k)) - 1]$, if $\{w_{k+i}\}_{0 \leq i \leq j} \cap \mathcal{F}_e(\nabla f(p)) = \emptyset$ then $w_{k+j+1} \in B_{(n-j-1)}(p)$. For $j = -1$ we have $w_k \in B_{(n)}(p)$ by assumption. Now if $\{w_{k+i}\}_{0 \leq i \leq j} \cap \mathcal{F}_e(\nabla f(p)) = \emptyset$ we have

$$\begin{aligned} &\|w_{k+j+1} - p\| \leq \|w_{k+j} - p\| + \|w_{k+j+1} - w_{k+j}\| < r^{(n-j)}(\delta, p) + \|w_{k+j+1} - w_{k+j}\| \\ =& r^{(n-j)}(\delta, p) + \alpha_k \|d_k\| \leq r^{(n-j)}(\delta, p) + D \frac{G(w_k)}{G_{\mathcal{F}}(w_k)} \\ &\leq r^{(n-j)}(\delta, p) + D M_{n-j} < r^{(n-j-1)}(\delta, p) , \end{aligned} \tag{5.3.17}$$

where we used the inductive hypothesis $w_{k+j} \in B_{(n-j)}(p)$ in the second inequality, Lemma 5.3.2 in the third inequality, (5.3.11) in the fourth inequality and the assumption (5.3.12) in the last one. In particular $w_{k+j+1} \in B_{(n-j-1)}(p)$ and the induction is completed.
Since $B_{(n-j)}(p) \subset B_{(0)}(p)$, if $w_{k+j} \in (B_{(n-j)}(p) \setminus \mathcal{F}_e(\nabla f(p))$ then α_{k+j} must be maximal and therefore $\dim(\mathcal{F}(w_{k+j+1})) < \dim(\mathcal{F}(w_{k+j}))$. But starting from the index k the dimension of the current face can decrease at most $\dim(\mathcal{F}(w_k)) < n$ times in consecutive steps, so there must exists $j \in [0, \dim(\mathcal{F}(w_k))]$ such that $w_{k+j} \in \mathcal{F}_e(\nabla f(p))$. Taking the minimum j satisfying this condition we also obtain $w_{k+j} \in B_{(0)}(p) \subset B_\delta(p)$. □

A straightforward adaptation of results from [47] implies convergence to the set of stationary points for the FDFW.

Proposition 5.3.4. *If* (S1) *and* (S2) *hold, then all the limit points of the FDFW are contained in the set of stationary points of f.*

Proof. The proof presented in the special case of the simplex in [47], where the FDFW coincides with the away-step Frank-Wolfe, extends to generic polytopes in a straightforward way. □

In the next lemma we improve the FDFW local identification result given in Proposition 5.3.3 under an additional strong concavity assumption for the face containing the solution, satisfied in particular by $h_{\mathcal{G}}$.

Lemma 5.3.5. *Let p be a local maximizer for f restricted to $\mathcal{Q}$. Assume that* (S1) *holds and that f is strongly concave*[2] *in $\mathcal{F}_e(\nabla f(p))$. Then, for a neighborhood $U(p)$ of p, if $w_0 \in U(p)$:*

(a) if $\{f(w_k)\}$ is increasing, there exists $k \in [0 : \dim(\mathcal{F}(w_0))]$ such that $w_{k+i} \in \mathcal{F}_e(\nabla f(p))$ for every $i \geq 0$;

(b) if in addition also (S2) *holds, then $\{w_{k+i}\}_{i\geq 0}$ converges to p.*

Proof. (a) Let μ be the strong concavity constant of f restricted to $\mathcal{F}_e(\nabla f(p))$, so that

$$f(w) \leq f(p) - \frac{\mu}{2}\|w - p\|^2 \tag{5.3.18}$$

for every $w \in \mathcal{F}_e(\nabla f(p))$. For $\varepsilon = \frac{\mu r^*(p)^2}{2}$, let $\mathcal{L}_\varepsilon$ be the superlevel of f for $f(p) - \varepsilon$:

$$\mathcal{L}_\varepsilon = \{y \in \mathcal{Q} \mid f(y) > f(p) - \varepsilon\}\,. \tag{5.3.19}$$

[2] in fact, we only need strict concavity of f here.

Let now $\bar{r} = r(\delta, p)$ defined as in Proposition 5.3.3, with $\delta = r^*(p)$. By (5.3.18) it follows $\mathcal{L}_\varepsilon \cap \mathcal{F}_e(\nabla f(p)) \subset B_{r^*(p)}(p)$. Assume now $w_0 \in U(p)$ with $U(p) = B_{\bar{r}}(p) \cap \mathcal{L}_\varepsilon$. By applying Proposition 5.3.3 we obtain that there exists $k \in [0 : \dim(\mathcal{F}(w_0))]$ such that w_k is in $\mathcal{F}_e(\nabla f(p)) \cap B_{r^*(p)}(p)$. But since $f(w_k) \geq f(w_0) > f(p) - \varepsilon$ we have the stronger condition $w_k \in \mathcal{L}_\varepsilon \cap \mathcal{F}_e(\nabla f(p))$.To conclude, notice that the sequence cannot escape from this set, because for $i \geq 0$ $w_{k+i} \in \mathcal{L}_\varepsilon$ implies that also w_{k+i+1} is in $\mathcal{L}_\varepsilon$, and $w_{k+i} \in \mathcal{L}_\varepsilon \cap \mathcal{F}_e(\nabla f(p)) \subset B_{r^*(p)}(p) \cap \mathcal{F}_e(\nabla f(p))$ implies that also w_{k+i+1} is in $\mathcal{F}_e(\nabla f(p))$.
(b) By point (a) $\{w_{k+i}\}_{i \geq 0}$ is contained in $\mathcal{F}_e(\nabla f(p))$. But by assumption f is strongly concave in $\mathcal{F}_e(\nabla f(p))$ with p global maximum and the only stationary point. To conclude it suffices to apply Proposition 5.3.4. □

Corollary 5.3.6. *Let $\{w_k\}$ be a sequence generated by the FDFW algorithm. Assume that there are no saddle points in the limit set of $\{w_k\}$, and that for every local maximizer $\tilde{p}$ the objective f is strongly concave in $\mathcal{F}_e(\nabla f(\tilde{p}))$. Then under the conditions* (S1) *and* (S2) *on the stepsize, we have $w_k \to p$ with p a local maximizer satisfying $w_k \in \mathcal{F}_e(\nabla f(p))$ for k large enough.*

Proof. Follows from (5.3.5) by observing that the sequence must be ultimately contained in $U(p)$ for some local maximizer p. □

Proof of Theorem 5.3.1. Let $A(p) = A_{\mathcal{G}} + A(y^{(p)})$. Then for $x \in \Delta^{(C)}$

$$\begin{aligned} x^\intercal A(p) x &= \sum_{(i,j) \in V^2} x_i A(p)_{ij} x_j = \sum_{i \in C} x_i (\sum_{j \in C} A(p)_{ij} x_j) = \sum_{i \in C} x_i (\sum_{j \in C \setminus \{i\}} x_j) \\ &= \sum_{i \in C} (x_i \sum_{j \in C} x_j - x_i^2) = (\sum_{i \in C} x_i)^2 - \sum_{i \in C} x_i^2 , \end{aligned} \tag{5.3.20}$$

where in the first equality we used $\operatorname{supp}(x) = C$, in the second that C is a clique n $G(y^{(p)})$, and $\sum_{i \in C} x_i = \sum_{i \in V} x_i = 1$.
Observe now that the function $x \mapsto h_{\mathcal{G}}(x, y^{(p)})$ is strongly concave in $\Delta^{(C)}$. Indeed for $x \in \Delta^{(C)}$

$$\begin{aligned} h_{\mathcal{G}}(x, y^{(p)}) &= x^\intercal A(p) x + \frac{\alpha}{2} \|x\|^2 + \frac{\beta}{2} \|y^{(p)}\|^2 \\ &= (\sum_{i \in C} x_i)^2 - \sum_{i \in C} x_i^2 + \frac{\alpha}{2} \|x\|^2 + \frac{\beta}{2} \|y^{(p)}\|^2 = 1 - (1 - \frac{\alpha}{2}) \sum_{i \in C} x_i^2 + \frac{\beta}{2} \|y^{(p)}\|^2 , \end{aligned} \tag{5.3.21}$$

where in the second equality we used (5.3.20). The RHS of (5.3.21) is strongly concave in x since $\alpha \in (0, 2)$ so that $-(1 - \alpha/2) \in (-1, 0)$. This together with Lemma 5.2.3 gives us the necessary assumptions to apply (5.3.5). □

As a corollary, we have the following global convergence result under the mild assumption that the set of limit points contains no saddle points.

Corollary 5.3.7 (FDFW global convergence)**.** *Let $\{z_k\}$ be a sequence generated by the FDFW and assume that there are no saddle points in the limit set of $\{z_k\}$. Then under the conditions* (S1) *and* (S2) *on the stepsize we have $z_k \to p = (x^{(C)}, y^{(p)}) \in \mathcal{M}_s(\mathcal{G})$, with* $\text{supp}(x_k) \subset C$ *and $y_k = y_p$ for k large enough.*

Proof. Follows from Corollary 5.3.6, where all the necessary assumptions are satisfied as for Proposition 5.3.1. □

5.4 FWdc: A Frank-Wolfe variant for s-defective clique

As can be seen from numerical results, one drawback of the standard FDFW applied to the s-defective clique formulation (P) is the slow convergence of the high dimensional y component. Since this component is "tied" to the x component, it is not possible to speed up the convergence by changing the regularization term without compromising the quality of the solution. Motivated by this challenge, we introduce a tailored Frank-Wolfe variant, namely FWdc, for the maximum s-defective clique formulation (P), which exploits the product domain structure of the problem at hand by employing separate updating rules for the two blocks.

In particular, at every iteration the method alternates a FDFW step on the x variables (Step 5) with a full FW step on the y variable (Step 6), so that y_k is always chosen in the set of vertices $D_s(\mathcal{G})$ of $D'_s(\mathcal{G})$. Furthermore, as stated in the next proposition, $\{y_k\}$ is ultimately constant. This allows us to obtain convergence

Algorithm 8 FWdc: Frank-Wolfe variant for s-defective clique

```
1: Initialize $z_0 := (x_0, y_0) \in \mathcal{P}_s$, $k := 0$
2: if $z_k$ is stationary then
3:    STOP
4: end if
5: Compute $x_{k+1}$ applying one iterate of Algorithm 7 with $w_0 = x_k$ and $f(w) = h_{\mathcal{G}}(w, y_k)$.
6: Let $y_{k+1} \in \arg\max_{y \in D'_s(\mathcal{G})} \nabla_y h_{\mathcal{G}}(x_{k+1}, y_k)^\top y$.
7: Set $k := k + 1$. Go to step 2.
```

results by applying the general properties of the FDFW proved in the previous section to the x component.

Proposition 5.4.1. *In Algorithm 8, if $h_{\mathcal{G}}(x_{k+1}, y_k) \geq h_{\mathcal{G}}(x_k, y_k)$ for every $k \in \mathbb{N}_0$, then $\{y_k\}$ can change at most $\frac{2}{\beta} - \frac{2-\alpha}{\beta|C^*|} + s$ times, with C^* s-defective clique of maximal cardinality.*

Proof. Assume that y_k and y_{k+1} are distinct vertices of $D'_s(\mathcal{G})$, and let $z_k^+ = (x_{k+1}, y_k)$. Then

$$\begin{aligned} h_{\mathcal{G}}(z_{k+1}) - h_{\mathcal{G}}(z_k^+) &\geq \nabla h_{\mathcal{G}}(z_k^+)^\top (z_{k+1} - z_k^+) + \frac{\beta}{2}\|z_{k+1} - z_k^+\|^2 \\ =&\nabla_y h_{\mathcal{G}}(z_k^+)^\top (y_{k+1} - y_k) + \frac{\beta}{2}\|y_{k+1} - y_k\|^2 \geq \frac{\beta}{2} > 0 \end{aligned} \tag{5.4.1}$$

where we used the $\beta-$strong convexity of $y \mapsto h_{\mathcal{G}}(x, y)$ in the first inequality, $z_k^+ - z_k = (0, y_k - y_{k+1})$ in the equality, $y_{k+1} \in \arg\max_{y \in \mathcal{P}_s} \nabla_y h_{\mathcal{G}}(z_k^+)^\top y$ and the fact that the distance between vertices of $D'_s(\mathcal{G})$ is at least 1 in the second inequality.

Therefore y_k can change at most

$$\max_{z \in \mathcal{P}_s} \frac{2(h_{\mathcal{G}}(z) - h_{\mathcal{G}}(z_0))}{\beta} \leq \max_{z \in \mathcal{P}_s} \frac{2h_{\mathcal{G}}(z)}{\beta} = \frac{1 - 1/|C^*| + \alpha/2|C^*| + s\beta/2}{\beta/2} = \frac{2}{\beta} + \frac{\alpha - 2}{\beta|C^*|} + s$$

times, where we used $h_{\mathcal{G}} \geq 0$ in the first inequality, and Corollary (5.2.2) in the second inequality. □

Theorem 5.4.2. *Let $\{z_k\}$ be a sequence generated by Algorithm 8, with regularization coefficient $\alpha = 1$. If conditions* (S1) *and* (S2) *hold on the stepsizes, then $\{z_k\}$ converges to a stationary point and identifies its support in finite time.*

Proof. As a corollary of Proposition 5.4.1, an application of Algorithm 8 reduces, after a finite number of changes for the variable y, to an application of the FDFW on the simplex for the optimization of the quadratic objective

$$\tilde{f}(x) = x^\top A(\bar{y})x + \frac{\alpha}{2}\|x\|^2 + \frac{\beta}{2}\|\bar{y}\|^2 = x^\top \hat{A}_{\mathcal{G}(\bar{y})}x + \frac{\beta}{2}\|\bar{y}\|^2\,, \tag{5.4.2}$$

for a fixed $\bar{y} \in D_s(\mathcal{G})$ and $\hat{A}_{\mathcal{G}(\bar{y})} = A(\bar{y}) + \frac{\alpha}{2}\mathbb{I}$.

This is, up to a constant, a regularized Motzkin-Straus quadratic formulation for the maximal clique problem associated to the graph $\mathcal{G}(\bar{y})$. For $\alpha = 1$, by the proof of [50, Theorem 12] we have that all principal minors of $\hat{A}_{\mathcal{G}(\bar{y})}$ do not vanish, and consequently by [50, Theorem 8] there can be at most one stationary point in the

relative interior of any face of the domain $\Delta_{|V|-1}$. Furthermore, by [182, Theorem 2.5], strict complementarity conditions hold in every stationary point.
By the above reasoning in particular we have that there is a finite number of stationary points, all satisfying strict complementarity conditions and with distinct supports. After noticing that on the simplex the FDFW coincides with the AFW, we have all the assumptions to conclude by [47, Theorem 4.5]. □

Remark 5.4.3. While all local solutions correspond to cliques by Proposition 5.2.1, both Corollary 5.3.7 and Theorem 5.4.2 do not rule out convergence to saddle points. However, this is not an issue in practice. First, in our numerical tests the methods always converged to a local solution, in line with studies showing that many first order methods avoid saddle points with probability one (see, e.g., [50], [165]). Second, while local solutions are attractive as proved in Theorem 5.3.1, a saddle point by definition can never be attractive for any strictly monotone method. Lastly, for our specific problem there are cheap strategies to escape saddle points even when the starting point is "unlucky" (e.g. a saddle point itself). We now describe one such strategy for Algorithm 8, to be applied e.g. if the FW gap (5.3.2) is below a certain threshold and $\text{supp}(x_k)$ is not yet a clique in $\mathcal{G}(y_k)$. The first step is to select $\{i, j\} \subset \text{supp}(x_k) \setminus E$, an operation which requires checking at most $\binom{\text{supp}(x_k)}{2}$ entries of the adjacency matrix. The second step, assuming without loss of generality $\frac{\partial}{\partial x_i} h_{\mathcal{G}}(x_k, y_k) \leq \frac{\partial}{\partial x_j} h_{\mathcal{G}}(x_k, y_k)$, is to replace $(x_k)_i$ and $(x_k)_j$ with $(1-\epsilon)(x_k)_i$ and $(x_k)_j + \epsilon(x_k)_i$ respectively, for some fixed $\epsilon \in (0, 1]$. The resulting point can then be used as a new starting point for Algorithm 8. It is not difficult to prove that if (x_k, y_k) is close enough to a saddle point p, then the algorithm escapes from p after restarting.

For a clique C of $\mathcal{G}(y)$ different from $\mathcal{G}$ we define $m(C, \mathcal{G}(y))$ as

$$\min_{v \in V \setminus C} |C| - |E^y(v) \cap C| , \tag{5.4.3}$$

that is the minimum number of edges needed to increase by 1 the size of the clique. We now give an explicit bound on how close the sequence $\{x_k\}$ generated by Algorithm 8 must be to $x^{(C)}$ for the identification to happen.

Proposition 5.4.4. *Let $\{z_k\}$ be a sequence generated by Algorithm 8, $\bar{y} \in D^s(\mathcal{G})$, C be a clique in $\mathcal{G}(\bar{y})$, let $\delta_{\max}$ be the maximum eigenvalue of the adjacency matrix $\bar{A} := A_{\mathcal{G}} + A(\bar{y})$. Let $\bar{k}$ be a fixed index in $\mathbb{N}_0$, I^c the components of* $\text{supp}(x_{\bar{k}})$ *with index not in C and let $L := 2\delta_{\max} + \alpha$. Assume that $y_{\bar{k}+j} = \bar{y}$ is constant for*

$0 \leq j \leq |I^c|$, *that* (S1) *holds for* $c = 1/L$, *and that*

$$\|x_{\bar{k}} - x^{(C)}\|_1 \leq \frac{m_\alpha(C, \mathcal{G}(y_{\bar{k}}))}{m_\alpha(C, \mathcal{G}(y_{\bar{k}})) + 2|C|\delta_{\max} + |C|\alpha} \tag{5.4.4}$$

for $m_\alpha(C, \mathcal{G}(y_{\bar{k}})) = m(C, \mathcal{G}(y_{\bar{k}})) - 1 + \alpha/2$. *Then* $\operatorname{supp}(x_{\bar{k}+|I^c|}) = C$.

Proof of Proposition 5.4.4. Since y_k does not change for $k \in [\bar{k} : \bar{k} + |I^c|]$, Algorithm 8 corresponds to an application of the AFW to the simplex Δ_{n-1} on the variable x. For $1 \leq i \leq n$ let $\lambda_i(x) = \frac{\partial}{\partial x_i} h_{\mathcal{G}}(x, y_{\bar{k}})$ be the multiplier functions associated to the vertices of the simplex, and let

$$\lambda_{\min} = \min_{i \in V \setminus C} -\lambda_i(x^{(C)})\,, \tag{5.4.5}$$

be the smallest negative multiplier with corresponding index not in C. Let L' be a Lipschitz constant for $\nabla_x h_{\mathcal{G}}(x, y)$ with respect to the variable x. By [47, Theorem 3.3] if

$$\|x_{\bar{k}} - x^{(C)}\|_1 < \frac{\lambda_{\min}}{\lambda_{\min} + 2L'} \tag{5.4.6}$$

we have the desired identification result.

We now prove that we can take L' equal to L in the following way:

$$\|\nabla_x h_{\mathcal{G}}(x', y_{\bar{k}}) - \nabla_x h_{\mathcal{G}}(x, y_{\bar{k}})\| = \|2\bar{A}(x' - x) + \alpha(x' - x)\| \leq (2\delta_{\max} + \alpha)\|x' - x\|\,, \tag{5.4.7}$$

where we used $\nabla_x h_{\mathcal{G}}(x, y) = 2\bar{A}x + \alpha x$ in the equality. As for the multipliers, for $i \in V \setminus C$ we have the lower bound

$$\begin{aligned} -\lambda_i(x^{(C)}) = \nabla_x h_{\mathcal{G}}(x^{(C)}, y_{\bar{k}})^\top (x^{(C)} - e_i) &= \frac{-2|C \cap E^{y_{\bar{k}}}(i)| + 2|C| - 2 + \alpha}{|C|} \\ &\geq \frac{2m_\alpha(C, \mathcal{G}(y_{\bar{k}}))}{|C|} \end{aligned} \tag{5.4.8}$$

by combining (5.2.20) and (5.2.21) in the second equation. We can now bound $\lambda_{\min}$ from below:

$$\lambda_{\min} = \min_{i \in V \setminus C} -\lambda_i(x^{(C)}) \geq \min_{i \in V \setminus C} \frac{2|C| - 2|C \cap E^y_{\bar{k}}(i)| - 2 + \alpha}{|C|} \geq \frac{2m_\alpha(C, \mathcal{G}(y_{\bar{k}}))}{|C|}\,, \tag{5.4.9}$$

where we applied (5.4.8) in the inequality. Finally, we have

$$\frac{\lambda_{\min}}{\lambda_{\min} + 2L} \leq \frac{m_\alpha(C, \mathcal{G}(y_{\bar{k}}))}{m_\alpha(C, \mathcal{G}(y_{\bar{k}})) + 2|C|\delta_{\max} + |C|\alpha} \tag{5.4.10}$$

where we applied (5.4.8) together with (5.4.9) in the inequality. The thesis follows applying (5.4.10) to the RHS of (5.4.6). □

Remark 5.4.5. It is a well known result that for any graph the maximal eigenvalue $\delta_{\max}$ of the adjacency matrix is less than or equal to $d_{\max}$, the maximum degree of a node (see, e.g., [87]). Then condition (5.4.4) can be replaced by

$$\|x_{\bar{k}} - x^{(C)}\|_1 \leq \frac{m_\alpha(C, \mathcal{G}(y_{\bar{k}}))}{m_\alpha(C, \mathcal{G}(y_{\bar{k}})) + 2|C|d_{\max} + |C|\alpha} \,. \tag{5.4.11}$$

5.5 Numerical results

In this section we report on a numerical comparison of the methods. We remark that, even though these methods only find maximal s-defective cliques, they can still be applied as a heuristic to derive lower bounds on the maximum s-defective clique within a global optimization scheme. With our tests, we aim to achieve the followings:

- empirically verify the active set identification property of the proposed methods;
- prove that the proposed FW variant is faster than the FDFW on these regularized problems, while mantaining the same solution quality;
- show that the proposed FW variant give better performances than a given black-box solver (i.e., CONOPT) on these regularized problems both in terms of CPU time and solution found;
- show that our approach, which is based on solving the regularized problem (P) via the FWdc algorithm, finds solutions as good as the ones found by the method described in [217], which consists in solving the Motzkin-Straus problem
$$\max\{f_{\mathcal{G}}(z) \mid z \in \mathcal{P}_s\} \,, \tag{MS}$$
using the CONOPT solver combined with a tailored post processing routine.

In the tests, the regularization parameters were set to $\alpha = 1$ and $\beta = 2/n^2$. An intuitive motivation for this choice of β can be given by imposing that the missing edges for an identified s-defective clique are always included in the support of the FW vertex. Formally, if $x_k = x^{(C)}$ with C an s-defective clique and $(y_k)_{ij} = 0$ with $\{i, j\} \in \binom{C}{2}$ we want to ensure that the FW vertex $s_k = (x^{(s_k)}, y^{(s_k)})$ is such that $y_{ij}^{(s_k)} = 1$. Now for $\{l, m\} \notin \binom{C}{2}$ and assuming $|C| < n$ (otherwise $C = V$ and the problem is trivial) we have

$$\frac{\partial}{\partial y_{ij}} h_{\mathcal{G}}(x_k, y_k) = \frac{2}{|C|^2} > \frac{2}{n^2} = \beta \geq \frac{\partial}{\partial y_{lm}} h_{\mathcal{G}}(x_k, y_k) \tag{5.5.1}$$

where the first equality and the last inequality easily follow from (5.2.4). From (5.5.1) it is then immediate to conclude that $\{i, j\}$ must be in the support of $y^{(s_k)}$. We used the stepsize $\alpha_k = \bar{\alpha}_k$ with $\bar{\alpha}_k$ given by (S1) for $c = 1$, corresponding to an estimate of 1 for the Lipschitz constant L of $\nabla h_{\mathcal{G}}$. The SSC was used to improve the performance of the methods (see Chapter 3 for details). The code was written in MATLAB and the tests were performed on an Intel Core i7-10750H CPU 2.60GHz, 16GB RAM.
The 50 graph instances we used in the tests are taken from the Second DIMACS Implementation Challenge [140]. These graphs are a common benchmark to assess the performance of algorithms for maximum (defective) clique problems (see references in [217]), and the particular instances we selected coincide with the ones employed in [217] in order to ensure a fair comparison at least for the quality of the solutions. Following the rule adopted in [217], for each triple $(\mathcal{G}, s, \mathcal{A})$ with $\mathcal{G}$ a graph from the 50 instances considered, $s \in [1:4]$, $\mathcal{A}$ the FDFW, the FWdc or the CONOPT solver, we set a global time limit of 600 seconds and employed a simple restarting scheme with up to 100 random starting points. For all the algorithms the x component of the starting point was generated with MATLAB's function rand and then normalized dividing it by its sum. An analogous rule was applied to generate the y component for the starting point of the FDFW and the CONOPT solver, while for the FWdc algorithm the y component was simply initialized to 0. To improve the performance of the FDFW, we exploit the quick reduction in the dimension of the minimal face containing the current iterate for the y variable. This improvement is possible using that the SSC with the FDFW method always operates on the minimal face containing the iterate given as input, until at least the last step (which can be a FW step and move the iterate away from the starting face). For the stopping criterion of the FDFW and the FWdc, two conditions are required: the current support of the x components coincides with an s-defective clique, and the FW gap is less than or equal to $\varepsilon := 2 * 10^{-3}$. For the CONOPT solver there are no identification guarantees, so the default stopping criterion was used. In the experiments, both the FDFW and the FWdc always terminated having identified an s-defective clique, thus providing an empirical verification of the results we proved in this chapter.

In the boxplots, each series consists of 50 values corresponding to aggregate data

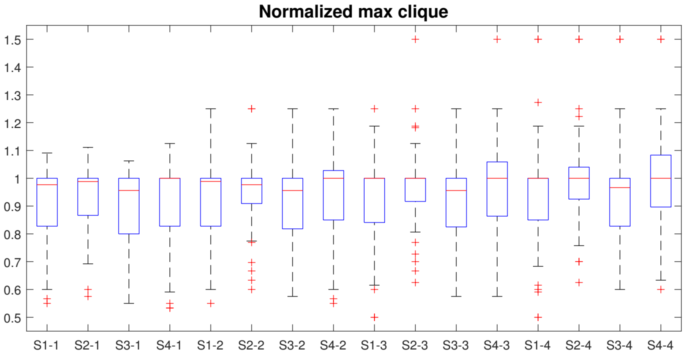


Figure 5.2: *Si-j* is the box plot related to the maximum clique found for the instance by strategy i for $s = j$, divided by the clique number/maximum clique cardinality known of the instance.

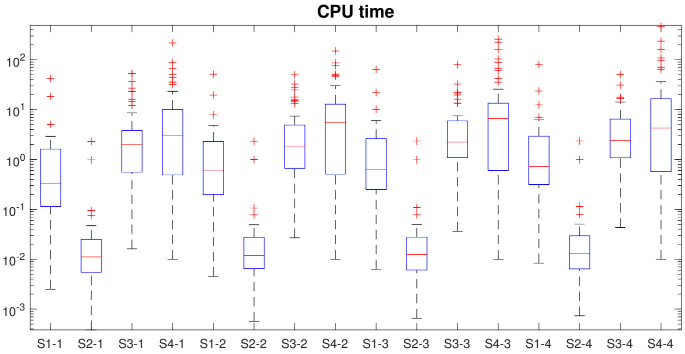


Figure 5.3: *Si-j* is the box plot related to the average running time for strategy i with $s = j$.

for the runs performed on the 50 instances. Here we list the strategies considered in our experiments:

- **Strategy 1** and **Strategy 2** (abbreviated **S1** and **S2**) consist in solving the regularized problem (P) using, respectively, the FDFW and the FWdc algorithm with the parameters reported above.

- **Strategy 3** (abbreviated **S3**) consists in solving the regularized problem (P) by means of the CONOPT solver.

- **Strategy 4** (abbreviated **S4**) consists in solving the Motzkin-Straus problem (MS) by means of the CONOPT solver combined with a post processing routine.

The numerical results related to Strategy 4 are taken from [217], while the results for Strategy 3 were replicated on our machine using the CONOPT/MATLAB integration provided by TOMLAB. We highlight that the results reported for Strategy 4 are meant to give the reader a baseline for the quality of the solutions found by our method. The red lines represent the median of the values in each series, and the boxes extend from the 25th percentile q_1 of the observed data to the 75th percentile q_3. The whiskers cover all the other values in a range of $[q_1 - w(q_3 - q_1), q_3 + w(q_3 - q_1)]$, with the coefficient w equal to 2.7 times the standard deviation of the values.

In Figure 5.2, the box plot *Si-j* represents the distribution of the maximum cardinality of the s-defective clique found by strategy i with $s = j$, divided by the maximum clique cardinality known of the instance. Notice that some data points are greater than 1, as expected since for $s > 0$ the cardinality of an s-defective clique can exceed the maximum clique cardinality. The solutions obtained using both FWdc anf FDFW on the regularized problem (Strategy 1 and 2) are generally better than the ones obtained using the CONOPT solver on the same problem (Strategy 3). Furthermore, while the variance is higher for the solutions found by Strategy 4, no significant difference can be seen in the median quality of the solutions found by Strategy 1, Strategy 2 and Strategy 4.

In Figure 5.3, *Si-j* represents the distribution of average running times in seconds (on a logarithmic scale, explaining the asymmetry of the box plots) of strategy i for $s = j$. Here we can see that FWdc outperforms both FDFW and the CONOPT on the regularized problem (MS). Indeed, FWdc is more efficient (as it requires a much smaller median execution time) and more robust (as the variance of the CPU time is remarkably smaller). Furthermore, we notice that the CPU times reported for Strategy 2 are good if compared with the ones obtained by Strategy 4 in [217]. The results hence indicate that the proposed strategy is a viable alternative when searching for s-defective maximal cliques. We refer the reader to the supplementary material for detailed numerical results.

Chapter 6

Direct search methods

While there is no unique definition of direct search methods, these can be characterized as derivative free methods that do not build, implicitly or explicitly, a model of the gradient. Starting mostly as intuitive and easy to implement heuristics in the '50, they have now become a diverse set of algorithms with rigorous convergence analyses, global and local convergence guarantees, and a wide range of applications. In this chapter, we review some classic direct search methods and properties relevant for the algorithms studied in Chapters 7 and 8.

6.1 A short history

Direct search methods are first of all zeroth order (or derivative free) methods, requiring a black box oracle only for the objective value. However, beside this elementary property there is no formal definition of what makes a method "direct search". This term, in reference to a class optimization algorithms, was first used in [127], for iterative methods with a strategy to select new trial points based on previous function evaluations and in particular on the best solution obtained up to that time. Today "direct search" is used more broadly, with M. Wright's [239] application to any method that "does not in its heart develop an approximate gradient" widely accepted (see, e.g., [151,163]). With respect to model based derivative free methods, direct search approaches arguably require weaker assumptions on the objective [16], being easily adaptable even to problems with discontinuities [53]. Moreover, many of these methods were originally developed as heuristics, and in spite of an increasing number of works proving rigorous convergence properties with

classic analysis arguments (see, e.g., [81, 151]), direct search algorithms still include steps whose effectiveness cannot be easily quantified (see also Section 6.3).
In their history of direct search methods in [169], the authors distinguish three classes: simplex, pattern search, and adaptive sets of search directions methods. Pattern search methods choose tentative points in a rational lattice, with a parameter to define the resolution of the lattice updated at every iteration based on the function value of new trial points. The exploration strategies of these methods are devised to visit enough points in a neighborhood of the current tentative solution to guarantee stationarity at the limit. This class includes coordinate search, widely recognized as the oldest direct search method and first described by E. Fermi and N. Metropolis in [99], generalized pattern search (GPS, [222]), integrating heuristics in between local exploration steps, and mesh adaptive direct search (MADS, [18]), a further development considering a dense set of directions for local exploration steps in order to deal with non smoothness and constraints. Simplex methods maintain a simplex with the respective function values of the vertices, and modify this simplex at every iteration in a way to adapt it to the features of the objective function. The first instance of a simplex method appeared in [216], and was based on the single operation of reflecting a vertex with respect to the baricenter of the opposite face. The most popular algorithm in this class is instead the Nelder-Mead method [190], relying on other operations called contractions and expansions beside reflections. Finally, adaptive set of search directions methods at every iteration change the set of poll directions, possibly to adapt it to information obtained about the objective. The first method in this class was proposed by Rosenbrock in [212], motivated by the inefficiency of coordinate search on certain objectives with minimizers in narrow valleys like the so called Rosenbrock's "banana function". The main idea of Rosenbrock's algorithm is to rotate the set of search directions at certain steps, ensuring the inclusion of a direction derived chaining several previous steps. The most known adaptive search method is Powell's method [204], adapting conjugate gradient to the derivative free case. More recent developments that can be included in this class are variants with line search extrapolation (see, e.g., [179]), increasing the stepsize along a poll direction until a decrease condition is no longer satisfied, and randomized direct search variants (see, e.g., [113]), relying on a random set of search directions and able to achieve optimal iteration complexity for smooth nonconvex objectives.
In this chapter, we focus on some pattern search and adaptive sets of search directions methods for unconstrained optimization. These will provide some context for the extensions to the Riemannian and stochastic setting in chapter 7 and 8 respec-

tively. A thorough description of the history and modern developments of direct search methods is beyond the scope of this chapter. We refer the reader to [169] for a detailed history until the '90; to [151] for a survey focusing on the theoretical convergence properties of direct search methods both in the constrained and the unconstrained case; to [16] for a survey including recent techniques and applications to real world optimization problems; to [15, 81, 163] for books presenting direct search methods within the context of derivative free optimization.

6.2 Clarke directional derivative and cosine measure

We consider the following global optimization problem:

$$\min_{x\in\mathbb{R}^n} f(x) \tag{6.2.1}$$

where f is locally Lipschitz continuous. While there are plenty of works that deal with the constrained case, in this survey we focus only on the unconstrained case, given its relevance for Chapters 7 and 8.

We now introduce two important preliminaries. The first one is the *Clarke directional derivative* of f at x in the direction $v \in \mathbb{R}^n$, defined as

$$f^\circ(x, v) = \lim_{t\to 0} \sup_{y\to x} \frac{f(y + tv) - f(y)}{t}\,. \tag{6.2.2}$$

A point x^* is said to be Clarke stationary if all its directional derivatives are nonnegative: $f(x^*, v) \geq 0$ for every $v \in \mathbb{R}^n$. It is a well known result (see, e.g., [15, Theorem 6.9]) that if x^* is a local minimizer then it is Clarke stationary, while as in the differentiable case, the converse is not true.

The second important preliminary is the *cosine measure*. As we will see in Section 6.3, many direct search methods require sets of search directions with the special property of being positive spanning sets, where a set $\boldsymbol{D} \subset \mathbb{R}^n$ is a positive spanning set iff every element in $\mathbb{R}^n$ can be written as a linear combination with nonnegative coefficients of elements in $\boldsymbol{D}$. This concept is strictly related to that of cosine measure. For a finite subset $\boldsymbol{D} \subset \mathbb{R}^n$ (with nonzero vectors) the cosine measure related to a vector $r \in \mathbb{R}^n \setminus \{0\}$ is defined as the maximum cosine between a direction in $\boldsymbol{D}$ and r:

$$\mathrm{cm}(\boldsymbol{D}, r) = \max_{d\in D} \frac{r^\top d}{\|d\|\|r\|}\,. \tag{6.2.3}$$

The cosine measure of D itself can then be defined as the minimum cosine measure related to a vector r varying in $\mathbb{R}^n \setminus \{0\}$:

$$\mathrm{cm}(D) = \min_{r \in \mathbb{R}^n \setminus \{0\}} \mathrm{cm}(D, r)\,. \tag{6.2.4}$$

It is a well known property (see, e.g., [81, Section 2.2]) that D is a positive spanning set iff $\mathrm{cm}(D) > 0$.

6.3 Directional direct search methods

We will focus on a class of methods roughly following the basic scheme presented in Algorithm 10, which is a slight adaptation of [163, Algorithm 2]. As in [163], we will call the algorithm "directional direct search method". It relies on the `testdescent` subroutine (Algorithm 9), looking for some points in a set of tentative points satisfying a predetermined decrease condition. At every iteration, it performs a search step and a poll step. In the search step, a finite set of search points is chosen and ordered to update x_k using the `testdescent` subroutine. This step is driven by heuristics and not crucial for convergence purposes. In the poll step, another set of tentative points is generated by moving with stepsize α_k along each directions in the poll set D_k. With respect to [163, Algorithm 2], we do not impose a specific rule for the decrease or increase of α_k, since there can be strategies different than the linear one considered in [163, Algorithm 2], as we will se for MADS in Section 6.3.2. Furthermore, we note that line search variants do not strictly adhere to this scheme, since the stepsize α_k can depend from the tentative direction $d \in D_k$ (see Section 6.3.5). We finally remark that the scheme in Algorithm 10 covers pattern search and adaptive set of search directions methods but does not cover simplex methods.

Algorithm 9 `testdescent`(f, x, P)

```
Set $x^+ = x$
for $p \in P$ do
    Evaluate $f(p)$
    if $f(p) - f(x)$ acceptable then
        $x^+ = p$
        optional break
    end if
end for
```

Algorithm 10 `Directional direct search method`

```
Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 > 0$
for $k = 0, \dots$ do
    Choose and order a finite set $Y_k \subset \mathbb{R}^n$
    Set $x_k^+ = \texttt{testdescent}(f, x_k, Y_k)$                {search step}
    if $x_k^+ = x_k$ then
        Choose and order poll directions $D_k \subset \mathbb{R}^n$
        Set $x_k^+ = \texttt{testdescent}(f, x_k, \{x_k + \alpha_k d_i : d_i \in D_k\})$    {poll step}
    end if
    if $x_k^+ = x_k$ then
        decrease $\alpha_k$
    else
        increase $\alpha_k$
    end if
    $x_{k+1} = x_k^+$
end for
```

The most used acceptance tests for the decrease of f are the *simple decrease* condition

$$f(p) < f(x)\,, \tag{6.3.1}$$

and the *sufficient decrease* condition

$$f(p) < f(x) - \rho(\alpha)\,, \tag{6.3.2}$$

with α stepsize and some $\rho : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ non decreasing and such that

$$\lim_{t \to 0} \frac{\rho(t)}{t} = 0\,. \tag{6.3.3}$$

6.3.1 Coordinate search

For coordinate search, Y_k is empty (there is no search step), and $D_k = D$ is the set of coordinate directions:

$$D = \{\pm e_i \mid i \in \{1, \dots, n\}\}\,. \tag{6.3.4}$$

The stepsize α_k is always increased or decreased by a fixed rational constant τ.

6.3.2 Mesh based methods

Before illustrating the next two methods, that is GPS and MADS, we report the definition of mesh as given in [15, Part 3]. For a positive spanning set $\boldsymbol{D}$, a center x and a mesh size parameter $\delta > 0$, the related mesh is defined as

$$M = \{x + \delta D y \mid y \in \mathbb{N}^p\}, \tag{6.3.5}$$

where with a slight abuse of notation we use $\boldsymbol{D}$ also for the matrix $\boldsymbol{D} \in \mathbb{R}^{n \times p}$ with columns corresponding to the elements of $\boldsymbol{D}$.

Generalized pattern search

Given $\boldsymbol{G} \in \mathbb{R}^{n \times n}$ invertible and $\boldsymbol{Z} \in \mathbb{Z}^{n \times p}$ with columns forming a positive spanning set, GPS uses the mesh $\boldsymbol{M_k}$ with size parameter α_k, positive spanning set given by the columns of $\boldsymbol{D} = \boldsymbol{GZ}$ and center x. The method then follows the scheme presented in Algorithm 10 with search set $Y_k \subset \boldsymbol{M_k}$, and poll set $\boldsymbol{D_k}$ positive spanning subset of the columns of $\boldsymbol{D}$. In order for the method to show some convergence properties, the stepsize must always be increased or decreased by a predetermined constant $\tau \in \mathbb{Q}$. Finally, the decrease condition used by GPS is simple decrease. We have the following convergence property (see, e.g., [15, Theorem 7.7]).

Theorem 6.3.1. *If the level subsets of f are bounded, then there exists a subsequence $\{x_k\}_{k \in K}$ of $\{x_k\}$ convergent to a point x^* and such that:*

(i) if d appears infinitely often in $\{\boldsymbol{D_k}\}_{k \in K}$, then $f^\circ(x^*, d) \geq 0$.

(ii) if $f \in C^1$, then $\nabla f(x^*) = 0$

Mesh adaptive direct search

One key issue with GPS is that the set of poll directions is finite. Hence, as for coordinate search, even when the generated sequence converges there is no guarantee that the limit point is Clarke stationary (see [151] for a counterexample). Moreover, for constrained optimization problems GPS gets stuck in points where the cone of feasible descent directions does not include elements of $\boldsymbol{D}$. This motivated the introduction of MADS. Beside the mesh $\boldsymbol{M_k}$ defined exactly as for GPS, MADS makes use of the frame $\boldsymbol{F_k}$ of extent determined by the frame size parameter Δ_k, defined as

$$F_k = \{x \in M_k \mid \|x - x_k\|_\infty \leq \Delta_k b\}, \tag{6.3.6}$$

for $b = \max\{\|d\|_\infty \mid d \in D\}$. A popular rule relating the frame size parameter with the mesh size parameter is $\alpha_k = \min(\Delta_k, \Delta_k^2)$. In MADS, the search set is a finite subset of M_k like for GPS, while the poll set D_k must be a positive spanning set such that $x_k + \alpha_k D_k \subset F_k \cap M_k$. The acceptance criterion is still simple descent. For MADS, we have the following convergence property (see [15, Chapter 8] for a reference and more convergence results).

Theorem 6.3.2. *If the level subsets of f are bounded, then there exists a subsequence $\{x_k\}_{k\in K}$ of $\{x_k\}$ convergent to a point x^* and such that:*

(i) if d is a limit point of $\{d_k\}_{k\in K}$ with $d_k \in D_k$ for every $k \in \mathbb{N}_0$, then $f^\circ(x^*, d) \geq 0$.

(ii) if $f \in C^1$ and $\mathrm{cm}(D_k) \geq \kappa_{\min}$ for every $k \in \mathbb{N}_0$ and for a constant $\kappa_{\min} > 0$ independent from k, then $\nabla f(x^*) = 0$.

A result analogous to Theorem 6.3.1 holds for MADS, with $f^\circ(x^*, d) \geq 0$ for any d limit of a sequence of directions used in the poll steps of the convergence subsequence with index set K. A lower bound on the cosine measure of D_k is needed to ensure $\nabla f(x^*) = 0$

6.3.3 Generating set search

The generating set search approach (GSS) is another variant of Algorithm 10. For this method, there is no search step and no mesh. The only conditions on the set of poll directions D_k is that it must contain a positive spanning set G_k with $\mathrm{cm}(G_k) \geq \kappa_{\min}$ for some constant $\kappa_{\min} > 0$, and elements with uniformly lower and upper bounded norm. The following convergence result holds (see, e.g., [151, Theorem 3.11]) when the method uses the sufficient decrease condition (6.3.2).

Theorem 6.3.3. *Assume that f is differentiable with ∇f Lipschitz continuous, and that $[f \leq f(x_0)]$ is compact. Then*

$$\liminf_{k\to+\infty} \|\nabla f(x_k)\| = 0. \tag{6.3.7}$$

Notice therefore how we have a result analogous to point (ii) of Theorem 6.3.1, replacing the use of a mesh with the sufficient decrease condition.
Another important result for this method is the $O(\frac{n^2}{\epsilon^2})$ function evaluation complexity proved in [228] in the case where $\rho(\alpha) = \gamma\alpha^2$ for some $\gamma > 0$.

6.3.4 Direct search based on probabilistic descent

The use of random directions in the poll set has been a popular choice for several direct search methods including MADS. A suitable choice of random directions in MADS implies in fact convergence to Clarke stationary points (see [15, Chapter 8]). Direct search methods based on probabilistic descent take this idea one step further, relaxing the requirement that D_k must be a positive spanning set to a probabilistic assumption. More precisely, the version introduced in [113] assumes that all the directions in D_k are in the unit sphere and that with some probability $p > 0$, for a constant $\kappa_{\min} > 0$:

$$\mathbb{P}(\mathrm{cm}(D_k, -\nabla f(x_k)) > \kappa_{\min} \mid D_0, ..., D_{k-1}) > p \,. \tag{6.3.8}$$

In other words, D_k must have positive cosine measure with respect to $-\nabla f(x_k)$ with positive probability and in a uniform way.
The two main features of direct search based on probabilistic descent are the following:

- the condition (6.3.8) can be achieved by sampling any number of directions uniformly at random in the unit sphere;
- for continuously differentiable functions it has a function evaluation complexity of $O(\frac{mn}{\epsilon^2})$, for m number of directions sampled at every iteration, thus improving on the $O(\frac{n^2}{\epsilon^2})$ GSS complexity and achieving for m constant the same complexity of zeroth order methods (see, e.g., [111]), which is state of the art for smooth non convex problems.

6.3.5 Direct search methods with line search extrapolation

Direct search methods with line search aim to combine the benefits of line search, which exploits knowledge of good descent directions, together with those of pattern search, which obtains local information about the objective. We report here a special case of [179, Algorithm 2], one of the first algorithms proposed with this approach, rewriting it in a way that underlines its resemblance to the general scheme 10, without altering its main properties. The main innovations with respect to Algorithm 10 consists in the subroutine 12 and in the introduction of a tailored stepsize for each direction. Instead of testing all the directions in the poll set with a fixed stepsize, the method increases the tailored stepsize related to a direction $p \in P$ (Step 3 of Algorithm 12) until a sufficient decrease condition is no longer satisfied (Step 2 of Algorithm 12).

Algorithm 11 Direct search method with LS

```
Input: $x_0 \in \mathbb{R}^n$, $(\alpha_0^j)_{j\in[1:K]} \in \mathbb{R}^K_{>0}$, $\gamma > 0$, $\theta \in (0,1)$, positive spanning set $P = \{p^j\}_{j\in[1:K]}$.
Set $j_0 = 0$
for $k = 0, 1, \ldots$ do
    for $i = 1, \ldots, K$ do
        Set $x_{k+1}, \alpha^i_{k+1}$ = testacceptanceLS$(x_k, \alpha^i_k, p^i, \theta, \gamma)$
    end for
end for
```

Algorithm 12 `testacceptanceLS`$(x, \alpha, p, \theta, \gamma)$

```
if $f(x + \alpha p) \le f(x) - \gamma\alpha^2$ then
    while $f(x + \alpha p) \le f(x) - \gamma\alpha^2$ do
        Set $\alpha = \alpha/\theta$
    end while
    Set $x = x + \theta\alpha p$
end if
Set $\alpha = \theta\alpha$
Return $(x, \alpha)$
```

We have the following convergence result, which can be proved along the lines of [179, Proposition 5.2].

Proposition 6.3.4. *If* $[f \le f(x_0)]$ *is compact, and* f *is continuously differentiable,*

$$\lim_{k\to\infty} \nabla f(x_k) = 0\,. \tag{6.3.9}$$

6.4 Applications

While it is well understood that direct search methods are a poor choice for optimization problems where the gradient is available (see, e.g. [81]), there are a number of cases where these methods should be considered. First, direct search methods can be a good choice when the gradient of the objective is discontinuous at a solution, or when the gradient has many discontinuities with no special structure (see [151, Section 6]). Second, they can be useful for simulation based optimization problems where applying automatic differentiation is not possible because of a proprietary or legacy code too expensive to rewrite. Third, they can be useful when objective evaluations are costly, so that computing gradient estimates is too expensive, or when objective evaluations are noisy, so that computing an accurate estimate of the gradient might not be possible at all.

We now report practical examples, some taken from [81, Section 1] or [15, Section 6], together with more recent ones. We refer the reader to those works for a more detailed description as well as for a more extensive list. Several examples use NOMAD, the open source implementations of MADS (see, e.g., [1]).

The first example is hyperparameter tuning, i.e. finding the choice of parameters optimizing the performance of an algorithm (see, e.g., [132] for a survey on the subject). Examples of applications of direct search include [19], where NOMAD was used to optimize the performance of trust region methods on a standard set of problems; [227], where direct search methods were used to fine tune regression parameters for data streams; [159], where NOMAD was used to tune both learning and structural parameters of a deep neural network; [245], where MADS was used to tune some parameters in a generative adversarial network for text-based CAPTCHAs.

The second and perhaps most known example is engineering design. In [51] for instance direct search methods were used to optimize the design of an helicopter rotor blaze with respect to the vibration trasmitted to the hub. In [63] a computer aided material selection tool to support design of aircraft structure was developed using the Direct multi-search (DMS) solver from [86]. In [145] aerodynamic optimization of airfoils was performed with MADS.

The third example is molecular design, where computer aided simulation is a key tool to obtain structures with desirable properties, partly replacing inefficient trial-and-error experiments. Applications of direct search methods to these problems can be found in [8, 183, 219].

Lastly, direct search methods can be used in drug design and testing. When a mathematical model of the impact of a certain drug is available, optimization methods can be used to tune several parameters. As an example, in [70, 71] MADS was used to optimize drug distribution in a nanoparticle-mediated drug delivery treatment for cancer. In [141] NOMAD was used in the study of a key antimalarian substance.

Chapter 7

Retraction based Direct Search Methods for Riemannian Optimization

In this chapter, we explore the application of direct search methods to Riemannian optimization, wherein minimization is to be performed with respect to variables restricted to lie on a manifold. More specifically, we consider classic and line search extrapolation variants of direct search, and, by making use of retractions, we devise tailored strategies for the minimization of both smooth and nonsmooth functions. As such we analyze, for the first time in the literature, a class of direct search algorithms for minimizing nonsmooth objectives on a Riemannian manifold without having access to (sub)derivatives. Along with convergence guarantees we provide a set of numerical performance illustrations on a standard set of problems.

7.1 Derivative free optimization on Riemannian manifolds

Riemannian optimization, or solving minimization problems wherein the decision variable is constrained to lie on a Riemannian manifold, is an active area of research considering the numerous problems in data science, robotics, and other settings wherein there is an important geometric structure characterizing the allowable inputs.

To the best of our knowledge, thorough studies of derivative free optimization (DFO) on Riemannian manifolds have only been carried out recently in the literature. In [171], the authors focus on a model based method using a two point function approximation for the gradient. The paper [244] presents a specialized Polak-Ribiéere-Polyak procedure for finding a zero of a tangent vector field on a Riemannian manifold. In [92], it is claimed that the convergence analysis of MADS for unconstrained objectives can be extended to the case of embedded Riemannian manifolds using the exponential map. In the subsequent work [93], the author focuses on a specific class of manifolds (reductive homogeneous spaces, including several matrix manifolds), discussing more in detail how thanks to the properties of exponential maps, a straightforward extension of MADS is possible at least for that class. Some DFO methods and nonsmooth problems on Riemannian manifolds without convergence analysis can be found in [130] and references therein.

7.1.1 Contributions

This chapter presents the introduction of a classic set of direct search algorithms to the case of Riemannian optimization, as well as the first analysis of retraction based direct search strategies on Riemannian manifolds. In particular, we first adapt, thanks to the use of retractions, a classic direct search scheme (see, e.g., [81,151]) and a line search based scheme (see, e.g., [85,174,178,179] for further details on this class of methods) to deal with the minimization of a given smooth function over a manifold. Then, inspired by the ideas in [98], we extend the two proposed strategies to the nonsmooth case. The introduction of the manifold constraint presents significant challenges: namely the stable structure of the Euclidean vector space makes it natural for a fixed set of coordinate-like directions to consistently approximate desired directions by spanning the space in a uniform way. The fact that this geometric structure can change necessitates that we carefully adjust the poll directions corresponding to the change in this structure, and do so with minimal computational expense. The associated convergence theory presents some novel results that could be of independent interest.

The codes relevant to the numerical tests are available at the following link: `https://github.com/DamianoZeffiro/riemannian-ds`.

7.2 Preliminaries

We now introduce some notation for the formalism we use in this chapter. We refer the reader to, e.g., [3, 55, 56] for an overview of the relevant background.
Let $\mathcal{M}$ be a smooth manifold. We are interested here in the problem

$$\min_{x \in \mathcal{M}} f(x) \tag{7.2.1}$$

with f continuous and bounded below. We consider both the case of $f(x)$ being continuously differentiable, as well as a more general nonsmooth case. For $x \in \mathcal{M}$, let $T_x\mathcal{M}$ be the tangent vector space at x and $T\mathcal{M}$ be the tangent bundle $\cup_{x \in \mathcal{M}} T_x\mathcal{M}$. We assume that $\mathcal{M}$ is a compact and connected Riemannian manifold, but all our results can be extended to geodesically complete Riemannian manifolds in a straightforward way. For x in $\mathcal{M}$, we have a scalar product $\langle \cdot, \cdot \rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$ and a norm $\| \cdot \|_x$ on $T_x\mathcal{M}$ smoothly depending on x. Let $\operatorname{dist}(\cdot, \cdot)$ be the distance induced by the scalar product, so that for $x, y \in \mathcal{M}$ we have that $\operatorname{dist}(x, y)$ is the length of the shortest geodesic connecting x and y. Furthermore, let $\nabla_{\mathcal{M}}$ be the Levi-Civita connection for $\mathcal{M}$ (see [55, Theorem 5.5] for a precise definition), and $\Gamma : T\mathcal{M} \times \mathcal{M} \to T\mathcal{M}$ be a parallel transport with respect to $\nabla_{\mathcal{M}}$ along distance minimizing geodesics, with $\Gamma_x^y(v) \in T_y\mathcal{M}$ transport of the vector $v \in T_x\mathcal{M}$ to one in $T_y\mathcal{M}$ along a distance minimizing geodesic connecting x and y. Any nonuniqueness in the definition of Γ is either explicitly accounted for or inconsequential without loss of generality in the context.

When $\mathcal{M}$ is embedded in $\mathbb{R}^n$, we define P_x as the orthogonal projection from $\mathbb{R}^n$ to $T_x\mathcal{M}$, and $S(x, r) \subset \mathbb{R}^n$ as the sphere centered at x and with radius r.
We write $\{a_k\}$ as a shorthand for $\{a_k\}_{k \in I}$ when the index set I is clear from the context. We also use the shorthand notations $T_k\mathcal{M}, \mathsf{P}_k, \langle \cdot, \cdot \rangle_k, \| \cdot \|_k$, Γ_i^j for $T_{x_k}\mathcal{M}, \mathsf{P}_{x_k}, \langle \cdot, \cdot \rangle_{x_k}, \| \cdot \|_{x_k}$ and $\Gamma_{x_i}^{x_j}$. When there is no ambiguity on the value of x, we use simply $\| \cdot \|$ instead of $\| \cdot \|_x$.
We define the distance dist^* between vectors in different tangent spaces in a standard way using parallel transport (see for instance [20]): for $x, y \in \mathcal{M}$, $v \in T_x\mathcal{M}$ and $w \in T_yM$,

$$\operatorname{dist}^*(v, w) = \|v - \Gamma_y^x w\| = \|w - \Gamma_x^y v\|, \tag{7.2.2}$$

and for a sequence $\{(y_k, v_k)\}$ in $T\mathcal{M}$ we write $v_k \to v$ if $y_k \to y$ in $\mathcal{M}$ and $\operatorname{dist}^*(v_k, v) \to 0$. By compactness, for $\operatorname{dist}(x, y)$ small enough the minimizing geodesic is uniquely defined, and consequently the parallel transport Γ and the distance dist^* also are, as we will use in several proofs. Furthermore, by compactness and connectedness, a geodesic connecting x and y always exists and dist^* is

always well defined.
As it is common in the Riemannian optimization literature (see, e.g., [4]), to define our tentative descent directions we use a retraction $R : T\mathcal{M} \to \mathcal{M}$. We assume $R \in C^1(T\mathcal{M}, \mathcal{M})$, with

$$\mathrm{dist}(R(x, d), x) \leq L_r \|d\|\,, \tag{7.2.3}$$

(true in any compact subset of $T\mathcal{M}$ given the C^1 regularity of R, without any further assumptions)

For a scalar-valued function $f : \mathcal{M} \to \mathbb{R}$, let the gradient $\mathrm{grad} f(x)$ be defined as the unique element of $T_x\mathcal{M}$ such that for all $v \in \mathcal{M}$, it holds that,

$$Df(x)[v] = \langle v, \mathrm{grad} f(x)\rangle_x\,.$$

When $\mathcal{M}$ is embedded in $\mathbb{R}^n$, the (Riemannian) gradient is a simple projection onto $T_x\mathcal{M}$, i.e., $\mathrm{grad} f(x) = \mathsf{P}_x(\nabla f(x))$.

7.3 Smooth optimization problems

In this section, we consider solving (7.2.1) with the objective satisfying $f \in C^1(\mathcal{M})$, indicating that the gradient $\mathrm{grad} f(x)$ is continuous on $\mathcal{M}$ as a function of x. We now formally present the Lipschitz continuous gradient assumption.

Assumption 7.1. There exists $L_f > 0$ such that for all $x, y \in \mathcal{M}$

$$\mathrm{dist}^*(\mathrm{grad} f(x), \mathrm{grad} f(y)) = \|\Gamma_x^y \mathrm{grad} f(x) - \mathrm{grad} f(y)\| \leq L_f\, \mathrm{dist}(x, y)\,. \tag{7.3.1}$$

Consider this descent Lemma type decrease property,

$$f(R(x, d)) \leq f(x) + \langle \mathrm{grad} f(x), d\rangle + \frac{L}{2}\|d\|^2\,. \tag{7.3.2}$$

Like in the unconstrained case, the Lipschitz gradient property implies the standard descent property.

Proposition 7.3.1. *Assume that $\mathcal{M}$ is compact and R is a C^2 retraction. If condition* (7.3.1) *holds, then the decrease property* (7.3.2) *holds for some constant $L > 0$.*

Proof. Let (φ) be a chart defined in a neighborhood U of $x \in \mathcal{M}$. We can take the neighborhood small enough so that for y, z varying in U the parallel transport Γ_y^z depends smoothly on y, z and is uniquely defined. We use the notation $(\tilde{x}, \tilde{d}) = (\varphi(x), d\varphi(x)d)$ for $(x, d) \in T\mathcal{M}$. We pushforward the manifold and the related structure with the chart φ, i.e. for $\bar{\varphi} = \varphi^{-1}$ we define $\tilde{f} = f \circ \bar{\varphi}$, $\tilde{U} = \varphi(U)$, $\tilde{R}(\tilde{y}, \tilde{d}) = R(y, d)$, for $d, q \in T_x\mathcal{M}$ we define $g(\tilde{d}, \tilde{q}) = \langle d, q\rangle_x$, $\|\tilde{d}-\tilde{q}\|_{\tilde{x}} = \|d-q\|_x$, and $\tilde{\Gamma}_{\tilde{x}}^{\tilde{y}}(\tilde{d}) = \Gamma_x^y(d)$. With slight abuse of notation we use $\mathrm{dist}(\tilde{x}, \tilde{y})$ to denote $\mathrm{dist}(x, y)$. We also define as $\mathrm{grad}\tilde{f}$ the gradient of $\tilde{f}$ with respect to the scalar product g, so that $g(\mathrm{grad}\tilde{f}(\tilde{x}), \tilde{d}) = \langle\nabla\tilde{f}(x), d\rangle$ for any $\tilde{d} \in \mathbb{R}^m$. Importantly, by the equivalence of norms in $\mathbb{R}^m$ we can use $O(\|\tilde{d}\|_x)$ and $O(\|\tilde{d}\|)$ interchangeably.

We first prove (7.3.2) in x for some constant $L > 0$ and any d with $\|d\| \leq B$ for some $B > 0$. Equivalently, we want to prove

$$\tilde{f}(\tilde{R}(\tilde{x}, \tilde{d})) \leq \tilde{f}(\tilde{x}) + g(\mathrm{grad}\tilde{f}(\tilde{x}), \tilde{d}) + \frac{L}{2}\|\tilde{d}\|_{\tilde{x}}^2 . \tag{7.3.3}$$

for $\tilde{d}$ s.t. $\|\tilde{d}\| \leq B$.

By compactness we can choose (φ, U) and $B > 0$ in such a way that, for every $\tilde{y} \in \tilde{U}_1 \subset \tilde{U}$ and $\tilde{d}$ with $\|\tilde{d}\|_{\tilde{y}} \leq B$ we have $\tilde{R}(\tilde{y}, \tilde{d}) \in \tilde{U}_2 \subset \tilde{U}$, with $\tilde{U}_2$ compact and $B > 0$ independent from $\tilde{x}, \tilde{y}, \tilde{d}$.

First, since $\tilde{R}$ is in particular C^1 regular

$$\tilde{R}(\tilde{x}, \tilde{d}) = \tilde{x} + O(\|\tilde{d}\|_{\tilde{x}}) , \tag{7.3.4}$$

and by smoothness of the parallel transport

$$\tilde{\Gamma}_{\tilde{x}}^{\tilde{y}}\tilde{q} = \tilde{q} + O(\|\tilde{x} - \tilde{y}\|) . \tag{7.3.5}$$

Furthermore,

$$\mathrm{grad}\tilde{f}(\tilde{x} + \tilde{q}) = \tilde{\Gamma}_{\tilde{x}}^{\tilde{x}+\tilde{q}}\mathrm{grad}\tilde{f}(\tilde{x}) + O(\mathrm{dist}(\tilde{x}, \tilde{x} + \tilde{q})) , \tag{7.3.6}$$

by the Lipschitz continuity assumption (7.3.1), and consequently

$$\begin{aligned} &\mathrm{grad}\tilde{f}(\tilde{R}(\tilde{x}, \tilde{q})) = \tilde{\Gamma}_{\tilde{x}}^{\tilde{R}(\tilde{x},\tilde{q})}\mathrm{grad}\tilde{f}(\tilde{x}) + O(\mathrm{dist}(\tilde{x}, \tilde{R}(\tilde{x}, \tilde{q}))) \\ =&\tilde{\Gamma}_{\tilde{x}}^{\tilde{R}(\tilde{x},\tilde{q})}\mathrm{grad}\tilde{f}(\tilde{x}) + O(\|\tilde{q}\|) , \end{aligned} \tag{7.3.7}$$

where we used (7.2.3) in the last equality.

Finally, since, $\frac{d}{dt}\tilde{R}(\tilde{x}, t\tilde{d})$ is C^1 regular, we also have

$$\begin{aligned} &\frac{d}{dt}\tilde{R}(\tilde{x}, t\tilde{q})|_{t=h} = \frac{d}{dt}\tilde{R}(\tilde{x}, t\tilde{q})|_{t=0} + O(\|h\tilde{q}\|) \\ =&\tilde{q} + O(h\|\tilde{q}\|) = \tilde{\Gamma}_{\tilde{x}}^{R(\tilde{x},h\tilde{q})}\tilde{q} + O(\|R(\tilde{x}, h\tilde{q}) - \tilde{x}\|) + O(h\|\tilde{q}\|) = \tilde{\Gamma}_{\tilde{x}}^{R(\tilde{x},h\tilde{q})}\tilde{q} + O(h\|\tilde{q}\|) , \end{aligned} \tag{7.3.8}$$

where we used (7.3.5) in the third equality, and (7.2.3) in the last one. Again by compactness, for $\tilde{y} \in \tilde{U}_1$, $t \leq 1$, $\|\tilde{q}\|, \|\tilde{d}\| \leq B$ the implicit constants can be taken with no dependence from the variables.
Now for $\tilde{d}$ s.t. $\tilde{d} \leq B$ define $\tilde{q} = B\tilde{d}/\|\tilde{d}\|$, so that $\tilde{d} = \bar{t}\tilde{q}$ for $\bar{t} = \|\tilde{d}\|/B$. Then we obtain (7.3.3) reasoning as follows:

$$\begin{aligned}
&\tilde{f}(\tilde{R}(\tilde{x}, \tilde{d})) - \tilde{f}(\tilde{R}(\tilde{x}, 0)) = \tilde{f}(\tilde{R}(\tilde{x}, \bar{t}q)) - \tilde{f}(\tilde{R}(\tilde{x}, 0)) \\
=&\int_0^{\bar{t}} \frac{d}{dt}\tilde{f}(\tilde{R}(\tilde{x} + t\tilde{q}))dt = \int_0^{\bar{t}} g(\mathrm{grad} f(\tilde{R}(\tilde{x}, t\tilde{q})), \frac{d}{dt}\tilde{R}(\tilde{x}, t\tilde{d}))dt \\
=&\int_0^{\bar{t}} g(\tilde{\Gamma}_{\tilde{x}}^{\tilde{R}(\tilde{x},t\tilde{q})}\mathrm{grad}\tilde{f}(\tilde{x}) + O(t\|\tilde{q}\|), \tilde{\Gamma}_{\tilde{x}}^{\tilde{R}(\tilde{x},t\tilde{d})}\tilde{d} + O(t\|\tilde{q}\|))dt \qquad (7.3.9) \\
=&\int_0^{\bar{t}} \left(g(\tilde{\Gamma}_{\tilde{x}}^{\tilde{R}(\tilde{x},t\tilde{q})}\mathrm{grad}\tilde{f}(\tilde{x}), \tilde{\Gamma}_{\tilde{x}}^{\tilde{R}(\tilde{x},t\tilde{d})}\tilde{d}) + O(t\|\tilde{q}\|)\right) dt \\
=&g(\mathrm{grad} f(\tilde{x}), \tilde{d}) + O(\bar{t}^2\|\tilde{q}\|) = g(\mathrm{grad} f(\tilde{x}), \tilde{d}) + O(\|\tilde{d}\|^2)\,,
\end{aligned}$$

where we used (7.3.7) and (7.3.8) in the fourth inequality. To conclude, notice that the above argument does not depend from the choice of $\tilde{x} \in \tilde{U}_1$, so that it can be extended to every $\tilde{y} \in \tilde{U}_1$ and then by compactness to every $y \in \mathcal{M}$. □

We remark that Proposition 7.3.1 is a key tool to extend convergence properties from the unconstrained case to the Riemannian case. To the best of our knowledge, this result is new to the literature. Under the stronger assumption that f has Lipschitz gradient as a function in $\mathbb{R}^n$, the standard descent property was proved for retractions in [56]. For f twice differentiable, a local version of (7.3.2) was proved in [55, Lemma 10.58].
Another assumption we make in this context is that the gradient norm is globally bounded.

Assumption 7.2. There exists $M_f > 0$ such that

$$\|\mathrm{grad} f(x)\| \leq M_f, \qquad (7.3.10)$$

for every $x \in \mathcal{M}$.

For each of the algorithms in this section, we further assume that, at each iteration k, we have a positive spanning basis $\{p_k^j\}_{j\in[1:K]}$ of the tangent space $T_{x_k}M$ of the iterate x_k (further details on how to get a positive spanning basis can be found, e.g., in [81]). More specifically, we assume that the basis stays bounded and does not become degenerate during the algorithm, that is,

Assumption 7.3. There exists $B > 0$ such that

$$\max_{j \in [1:K]} \|p_k^j\| \leq B, \tag{7.3.11}$$

for every $k \in \mathbb{N}$. Furthermore there is a constant $\tau > 0$ such that

$$\max_{j \in [1:K]} \langle r, p_k^j \rangle \geq \tau \|r\|, \tag{7.3.12}$$

for every $k \in \mathbb{N}$ and $r \in T_{x_k} M$.

Notice how given the boundedness of the basis vectors (7.3.12) is equivalent to imposing that the cosine measure of $\{p_k^j\}$ as a positive spanning basis of $T_k\mathcal{M}$ is uniformly lower bounded for $k \in \mathbb{N}$.

7.3.1 Direct search algorithm

We present here our Riemannian Direct Search method based on Spanning Bases (RDS-SB) for smooth objectives as Algorithm 13.

Algorithm 13 RDS-SB

```
Input: $x_0 \in \mathcal{M}$, $\gamma_1 \in (0,1)$, $\gamma_2 \geq 1$, $\alpha_0 > 0$, $\rho > 0$
for $k = 0, 1, \ldots$ do
    Compute a positive spanning basis $\{p_k^j\}_{j=1:K}$ of $T_k\mathcal{M}$
    for $j = 1, \ldots, K$ do
        Let $x_k^j = R(x_k, \alpha_k p_k^j)$
        if $f(x_k^j) \leq f(x_k) - \rho\alpha_k^2$ then
            $\alpha_{k+1} = \gamma_2\alpha_k, x_{k+1} = x_k^j$
            Declare the step $k$ successful
            Break
        end if
    end for
    if $f(x_k^j) > f(x_k) - \rho\alpha_k^2$ for $j \in [1:K]$ then
        $\alpha_{k+1} = \gamma_1\alpha_k$, $x_{k+1} = x_k$
        Declare the step $k$ unsuccessful
    end if
end for
```

This procedure resembles the standard direct search algorithm for unconstrained derivative free optimization (see, e.g., [81, 151]) with two significant modifications. First, at every iteration a positive spanning basis is computed for the current tangent vector space $T_k\mathcal{M}$. As this space is expected to change at every iteration, it is not possible to use the same standard positive spanning sets appearing in the classic algorithms. Second, the candidate point x_k^j is computed by retracting the step $\alpha_k p_k^j$ from the current tangent space $T_{x_k^j}\mathcal{M}$ to the manifold, ensuring satisfaction of the geometric constraint.

7.3.2 Convergence analysis

Now we show convergence of the method, under the assumption that $\mathcal{M}$ is compact. We will first prove that the gradient, in unsuccessful iterates, must be bounded by a constant proportional to the stepsize (Lemma 7.3.3). This is a well known bound in the unconstrained case (see, e.g. [228, Theorem 1]), and we are able to extend it to the Riemannian case thanks to Proposition 7.3.1. Given that the stepsize converges to zero, the bound implies that the gradient converges to zero for unsuccessful steps. We then prove, using the Lipschitz continuity of the gradient, that the gradient converges to zero for successful steps as well. This is a novel result also for the unconstrained case, where only subsequential convergence guarantees are typically given for the gradient norm (see, e.g., [228] for some complexity bounds).

The first lemma states a bound on the scalar product between the gradient and the descent direction for an unsuccessful iteration.

Lemma 7.3.2. *If $f(R(x_k, \alpha_k p_k^j)) > f(x_k) - \rho\alpha_k^2$, then*

$$\alpha_k(LB^2 + \rho) > -\langle \mathrm{grad} f(x_k), p_k^j \rangle \,. \tag{7.3.13}$$

Proof. To start with, we have

$$\begin{aligned} f(x_k) - \rho\alpha_k^2 < f(R(x, \alpha_k p_k^j)) \leq f(x_k) + \alpha_k \langle \mathrm{grad} f(x_k), p_k^j \rangle + L\alpha_k^2 \|p_k^j\|^2 \\ \leq & f(x_k) + \alpha_k \langle \mathrm{grad} f(x_k), p_k^j \rangle + L\alpha_k^2 B^2 \,, \end{aligned} \tag{7.3.14}$$

where we used (7.3.2) in the second inequality, and (7.3.11) in the third one. The above inequality can be rewritten as

$$\alpha_k \langle \mathrm{grad} f(x_k), p_k^j \rangle + \alpha_k^2 (LB^2 + \rho) > 0. \tag{7.3.15}$$

Given that $\alpha_k > 0$, the above is true iff

$$\alpha_k > -\frac{\langle \mathrm{grad} f(x_k), p_k^j \rangle}{(LB^2 + \rho)} \,, \tag{7.3.16}$$

which rearranged gives the thesis. □

From this we can infer a bound on the gradient with respect to the stepsize.

Lemma 7.3.3. *If iteration k is unsuccessful, then*

$$\|\text{grad}f(x_k)\| \leq \frac{\alpha_k(2LB^2+\rho)}{\tau}\,. \tag{7.3.17}$$

Proof. If iteration k is unsuccessful, equation (7.3.13) must hold for every $j \in [1:K]$. We obtain the thesis by applying the positive spanning property (7.3.12) in the RHS:

$$\alpha_k(LB^2+\rho) > \max_{j\in[1:K]} -\langle \text{grad}f(x_k), p_k^j\rangle \geq \tau\|\text{grad}f(x_k)\|\,. \tag{7.3.18}$$

□

Finally, we are able to show convergence of the gradient norm using the lemmas above and appropriate arguments regarding the step sizes.

Theorem 7.3.4. *For the sequence $\{x_k\}$ generated by Algorithm 13 we have*

$$\lim_{k\to\infty} \|\text{grad}f(x_k)\| = 0\,. \tag{7.3.19}$$

Proof. To start with, it holds that $\alpha_k \to 0$ since the objective is bounded below, $\{f(x_k)\}$ is non increasing with $f(x_{k+1}) \leq f(x_k) - \rho\alpha_k^2$ if the step k is successful, and so there can be a finite number of successful steps with $\alpha_k \geq \varepsilon$ for any $\varepsilon > 0$.
For a fixed $\varepsilon > 0$, let $\bar{k}$ such that $\alpha_k \leq \varepsilon$ for every $k \geq \bar{k}$. We now show that, for every $\varepsilon > 0$ and $k \geq \bar{k}$ large enough, we have

$$\|\text{grad}f(x_k)\| \leq \varepsilon\left(\frac{(2LB^2+\rho)}{\tau} + L_f L_r B\frac{\gamma_2}{\gamma_2 - 1}\right)\,, \tag{7.3.20}$$

which implies the thesis given that ε is arbitrary.
First, (7.3.20) is satisfied for $k \geq \bar{k}$ if the step k is unsuccessful by Lemma 7.3.3:

$$\|\text{grad}f(x_k)\| \leq \frac{\alpha_k(2LB^2+\rho)}{\tau} \leq \frac{\varepsilon(2LB^2+\rho)}{\tau}\,, \tag{7.3.21}$$

using $\alpha_k \leq \varepsilon$ in the second inequality.
If the step k is successful, then let j be the minimum positive index such that the

step $k+j$ is unsuccessful. We have that $\alpha_{k+i} = \alpha_k \gamma_2^i$ for $i \in [0 : j-1]$, and since $\alpha_{k+j-1} \leq \varepsilon$ by induction we get $\alpha_{k+i} \leq \varepsilon \gamma_2^{i-j+1}$. Therefore

$$\sum_{i=0}^{j-1} \alpha_{k+i} \leq \sum_{i=0}^{j-1} \varepsilon \gamma_2^{i-j+1} \leq \varepsilon \sum_{h=0}^{\infty} \gamma_2^{-h} = \varepsilon \frac{\gamma_2}{\gamma_2 - 1} . \tag{7.3.22}$$

Then

$$\begin{aligned} \operatorname{dist}(x_k, x_{k+j}) &\leq \sum_{i=0}^{j-1} \operatorname{dist}(x_{k+i}, x_{k+i+1}) = \sum_{i=0}^{j-1} \operatorname{dist}(x_{k+i}, R(x_{k+i}, \alpha_{k+i} p_{k+i}^{j(k+i)})) \\ &\leq \sum_{i=0}^{j-1} L_r \alpha_{k+i} B \leq L_r B \varepsilon \frac{\gamma_2}{\gamma_2 - 1} . \end{aligned} \tag{7.3.23}$$

where we used (7.2.3) together with (7.3.11) in the second inequality, and (7.3.22) in the third one.
In turn,

$$\begin{aligned} &\|\operatorname{grad} f(x_k)\| \leq \operatorname{dist}^*(\operatorname{grad} f(x_k), \operatorname{grad} f(x_{k+j})) + \|\operatorname{grad} f(x_{k+j})\| \\ \leq & L_f \operatorname{dist}(x_k, x_{k+j}) + \frac{\varepsilon(2LB^2 + \rho)}{\tau} \leq \varepsilon \left(\frac{2LB^2 + \rho}{\tau} + L_f L_r B \frac{\gamma_2}{\gamma_2 - 1} \right) , \end{aligned} \tag{7.3.24}$$

where we used (7.3.1) and (7.3.21) with $k+j$ instead of k for the first and second summand respectively in the second inequality, and (7.3.23) in the last one. □

7.3.3 Incorporating line search extrapolation

The works [178, 179] (see also Section 6.3.5) introduced the use of an extrapolating line search that tests the objective on variable inputs farther away from the current iterate than the tentative point obtained by direct search on a given direction (i.e., an element of the positive spanning set). Such a thorough exploration of the search directions ultimately yields better performances in practice. We found that the same technique can be applied in the Riemannian setting to good effect. We present here our Riemannian Direct Search with Extrapolation method based on Spanning Bases (RDSE-SB) for smooth objectives. The scheme is presented in detail as Algorithm 14, which can be viewed as a Riemannian version of [179, Algorithm 2]. As we can easily see, the method uses a specific stepsize for each direction in the positive spanning basis, so that instead of α_k we have a set of stepsizes $\{\alpha_k^j\}_{j \in [1:K]}$ for every $k \in \mathbb{N}_0$. Furthermore a retraction based line search procedure (see Algorithm 15) is used to better explore a given direction in case a sufficient decrease of the objective is obtained.

When analyzing the RDSE-SB method, due to the changes in the tangent space, we cannot keep the same basis for different iterates as is done in the unconstrained case (see [179, Algorithm 2, Step 2 and 3]). We therefore introduce, using the distance dist* to compare vectors in different tangent spaces, a novel condition ensuring some continuity in the choice of the basis.

Assumption 7.4. For every $l, m \in \mathbb{N}$, $j \in [1:K]$, there exists a constant $L_\Gamma > 0$ such that

$$\operatorname{dist}^*(p_l^j, p_m^j) \leq L_\Gamma \operatorname{dist}(x_l, x_m)\,. \tag{7.3.25}$$

By compactness, condition (7.3.25) always holds globally if it holds when $\operatorname{dist}(x_l, x_m)$ is small enough. In turn, when $\mathcal{M}$ is embedded in $\mathbb{R}^n$ it is easy to see that this is true if $\{p_k^j\}_{j \in [1:K]}$ is the projection of a spanning basis of $\mathbb{R}^n$ (independent from k) into $T_k\mathcal{M}$, using that $T_x\mathcal{M}$ varies smoothly with x.

Algorithm 14 RDSE-SB

```
**Input:** $x_0 \in \mathbb{R}^n$, $\{\alpha_0^j\}_{j\in[1:K]}$, $\gamma > 0, \gamma_1 \in (0,1), \gamma_2 \geq 1$.
**for** $k = 0, 1, \ldots$ **do**
    Compute a positive spanning basis $\{p_k^j\}_{j\in[1:K]}$ of $T_k\mathcal{M}$
    Set $j(k) =$ mod $(k, n)$, $\alpha_k^i = \tilde{\alpha}_k^i$ and $\tilde{\alpha}_{k+1}^i = \tilde{\alpha}_k^i$ for $i \in [1:K] \setminus \{j(k)\}$.
    Compute $\alpha_k^{j(k)}, \tilde{\alpha}_{k+1}^{j(k)}$ with **Linesearchprocedure**($\tilde{\alpha}_k^{j(k)}, x_k, p_k^{j(k)}, \gamma, \gamma_1, \gamma_2$)
    Set $x_{k+1} = R(x_k, \alpha_k^{j(k)} p_k^{j(k)})$
**end for**
```

Algorithm 15 Linesearchprocedure($x, \alpha, d, \gamma, \gamma_1, \gamma_2$)

```
**if** $f(R(x_k, \alpha d)) > f(x) - \gamma\alpha^2$ **then**
    **Return** $(0, \gamma_1\alpha)$
**end if**
**while** $f(R(x_k, \alpha d)) < f(x) - \gamma\alpha^2$ **do**
    Set $\alpha = \gamma_2\alpha$
**end while**
**Return** $(\alpha/\gamma_2, \alpha/\gamma_2)$
```

We now proceed to prove convergence of this method.

Lemma 7.3.5. *We have, at every iteration k, that the following inequality holds:*

$$-\langle \mathrm{grad} f(x_k), p_k^{j(k)} \rangle < \tilde{\alpha}_{k+1}^{j(k)} \frac{\gamma_2}{\gamma_1} (2LB^2 + \gamma). \tag{7.3.26}$$

Proof. It is immediate to check that we must always have

$$f(R(x_k, \Delta_k p_k^{j(k)})) > f(x_k) - \gamma \Delta_k^2, \tag{7.3.27}$$

for $\Delta_k = \frac{1}{\gamma_1} \tilde{\alpha}_{k+1}^{j(k)}$ if the Linesearchprocedure terminates at the second line, and $\Delta_k = \gamma_2 \tilde{\alpha}_{k+1}^{j(k)}$ if the Linesearchprocedure terminates in the last line. Then in both cases

$$-\langle \mathrm{grad} f(x_k), p_k^{j(k)} \rangle < \Delta_k (2LB^2 + \gamma) \leq \tilde{\alpha}_{k+1}^{j(k)} \frac{\gamma_2}{\gamma 1} (2LB^2 + \gamma)\,, \tag{7.3.28}$$

where we used Lemma 7.3.2 in the first inequality. □

The assumption 7.4 allows us to extend [179, Proposition 5.2] to the Riemannian case.

Theorem 7.3.6. *For $\{x_k\}$ generated by Algorithm 14, we have*

$$\lim_{k \to \infty} \| \mathrm{grad} f(x_k) \| \to 0\,. \tag{7.3.29}$$

Proof. Let $\bar{\alpha}_k = \max_{j \in [1:K]} \tilde{\alpha}_{k+1}^{j(k)}$, so that $\bar{\alpha}_k \to 0$ since $\tilde{\alpha}_k^{j(k)} \to 0$, reasoning as in the proof of Theorem 7.3.4. As a consequence of Lemma 7.3.5 we have

$$-\langle \mathrm{grad} f(x_k), p_k^{j(k)} \rangle < \bar{\alpha}_k c_1\,, \tag{7.3.30}$$

for the constant $c_1 = \frac{\gamma_2}{\gamma_1}(2LB^2 + \gamma)$ independent from $j(k)$.
It remains to bound $\langle \mathrm{grad} f(x_k), p_k^i \rangle$ for $i \neq j$. To start with, we have the following bound:

$$\begin{aligned} &-\langle \mathrm{grad} f(x_k), p_k^i \rangle \leq -\langle \mathrm{grad} f(x_{k+h}), p_{k+h}^i \rangle + |\langle \mathrm{grad} f(x_{k+h}), p_{k+h}^i \rangle - \langle \mathrm{grad} f(x_k), p_k^i \rangle| \\ &\leq c_1 \bar{\alpha}_{k+h} + |\langle \mathrm{grad} f(x_{k+h}), p_{k+h}^i \rangle - \langle \mathrm{grad} f(x_k), p_k^i \rangle|\,, \end{aligned} \tag{7.3.31}$$

for $h \leq K$ such that $k+h = j(i)$, and where in the second inequality we used (7.3.30) with $k+h$ instead of k. For the second summand appearing in the RHS of (7.3.31), we can write the following bound

$$\begin{aligned}
&|\langle \mathrm{grad} f(x_{k+h}), p^i_{k+h}\rangle - \langle \mathrm{grad} f(x_k), p^i_k\rangle| \\
&= |\langle \mathrm{grad} f(x_{k+h}), p^i_{k+h}\rangle - \langle \Gamma_k^{k+h}\mathrm{grad} f(x_k), \Gamma_k^{k+h} p^i_k\rangle| \\
\leq& |\langle \mathrm{grad} f(x_{k+h}) - \Gamma_k^{k+h}\mathrm{grad} f(x_k), p^i_{k+h}\rangle| + |\langle \Gamma_k^{k+h}\mathrm{grad} f(x_k), p^i_{k+h} - \Gamma_k^{k+h} p^i_k\rangle| \\
+& |\langle \mathrm{grad} f(x_{k+h}) - \Gamma_k^{k+h}\mathrm{grad} f(x_k), p^i_{k+h} - \Gamma_k^{k+h} p^i_k\rangle| \\
\leq& L_f \operatorname{dist}(x_k, x_{k+h}) \|p^i_{k+h}\| + L_\Gamma \|\mathrm{grad} f(x_k)\| \operatorname{dist}(x_{k+h}, x_k) + L_f L_\Gamma \operatorname{dist}(x_k, x_{k+h})^2 \\
\leq& (L_f B + L_\Gamma M_f + L_f L_\Gamma \operatorname{dist}(x_{k+h}, x_k)) \operatorname{dist}(x_{k+h}, x_k)\,,
\end{aligned} \tag{7.3.32}$$

where in the second inequality we used the Cauchy-Schwartz inequality together with the Assumptions on the Lipschitz property of the iterates (7.3.1) and (7.3.25), while in the third inequality we used conditions (7.3.11) and (7.3.10).
We can now bound $\operatorname{dist}(x_k, x_{k+h})$ as follows

$$\begin{aligned}
&\operatorname{dist}(x_{k+h}, x_k) \leq \sum_{l=0}^{h-1} \operatorname{dist}(x_{k+l+1}, x_{k+l}) \\
=& \sum_{l=0}^{h-1} \operatorname{dist}(x_{k+l}, R(x_{k+l}, \bar{\alpha}_{k+l} p^{j(k+l)}_{k+l})) \leq \sum_{l=0}^{h-1} L_r \bar{\alpha}_{k+l} \|p^{j(k+l)}_{k+l}\| \\
\leq& B L_r \sum_{l=0}^{h-1} \bar{\alpha}_{k+l} \leq h B L_r \max_{l\in[0:h-1]} \bar{\alpha}_{k+l} \\
\leq& K B L_r \max_{l\in[0:K]} \bar{\alpha}_{k+l}\,,
\end{aligned} \tag{7.3.33}$$

where we used (7.2.3) in the second inequality, (7.3.11) in the third one, and $h \leq K$ in the last one.
Now let $\Delta_k = \max_{l\in[0:K]} \bar{\alpha}_{k+l}$, so that in particular $\Delta_k \to 0$. We apply (7.3.33) to the RHS of (7.3.32) and obtain

$$|\langle \mathrm{grad} f(x_{k+h}), p^i_{k+h}\rangle - \langle \mathrm{grad} f(x_k), p^i_k\rangle| \leq (L_f B + L_\Gamma M_f + L_f L_\Gamma c_2 \Delta_k) c_2 \Delta_k \to 0\,, \tag{7.3.34}$$

for $k \to \infty$ and $c_2 = KBL_r$. Finally, for every $i \in [1:K]$

$$-\langle \mathrm{grad} f(x_k), p^i_k\rangle \leq c_1 \bar{\alpha}_{k+h} + (L_f B + L_\Gamma M_f + L_f L_\Gamma c_2 \Delta_k) c_2 \Delta_k \to 0\,, \tag{7.3.35}$$

and the thesis follows after observing that, by (7.3.12),

$$\|\mathrm{grad} f(x_k)\| \leq \frac{1}{\tau} \max_{i\in[1:K]} -\langle \mathrm{grad} f(x_k), p^i_k\rangle \to 0\,, \tag{7.3.36}$$

where the convergence of the gradient norm to zero is a consequence of (7.3.35). □

7.4 Nonsmooth objectives

Now we proceed to present and study direct search methods in the context where f is Lipschitz continuous, and bounded from below, but not necessarily continuously differentiable. The algorithms we devise are built around the ideas given in [98], where the authors consider direct search methods for nonsmooth objectives in Euclidean space.

7.4.1 Clarke stationarity for nonsmooth functions on Riemannian manifolds

In order to perform our analysis, we first need to define the Clarke directional derivative for a point $x \in \mathcal{M}$. The standard approach is to write the function in coordinate charts and take the standard Clarke derivative in an Euclidean space (see, e.g., [129] and [131]). Formally, given a chart (φ, U) at $x \in \mathcal{M}$ and $v \in T_x\mathcal{M}$, we define

$$f^\circ(x, v) = \tilde{f}^\circ(\varphi(x), d\varphi(x)v)\,, \tag{7.4.1}$$

for $\tilde{f}(y) = f(\varphi^{-1}(y))$. The following lemma shows the relationship between definition (7.4.1) and a directional derivative like object defined with retractions. This nontrivial result is the key tool allowing us to extend the analysis of direct search methods on $\mathbb{R}^n$ to the Riemannian setting.

Lemma 7.4.1. *If $(y_k, q_k) \to (x, d)$ and $t_k \to 0$,*

$$f^\circ(x, d) \geq \limsup_{k\to\infty} \frac{f(R(y_k, t_k q_k)) - f(y_k)}{t_k}\,. \tag{7.4.2}$$

In order to prove the above result we first need the following lemma.

Lemma 7.4.2. *For a Lipschitz continuous function $h : \mathbb{R}^m \to \mathbb{R}$, $\tilde{y}, \tilde{v} \in \mathbb{R}^m$, if $\tilde{y}_k \to \tilde{y}$, $\tilde{v}_k \to \tilde{v}$ and $t_k \to 0$ then*

$$h^\circ(\tilde{y}, \tilde{v}) \geq \limsup_{k\to\infty} \frac{h(\tilde{y}_k + t_k\tilde{v}_k) - h(\tilde{y}_k)}{t_k}\,. \tag{7.4.3}$$

Proof. We have

$$|h(\tilde{y}_k + t_k\tilde{v}_k) - h(\tilde{y}_k + t_k\tilde{v})| \leq t_k L_h \|\tilde{v} - \tilde{v}_k\| = o(t_k)\,, \tag{7.4.4}$$

with L_h the Lipschitz constant of h. Then

$$\begin{aligned}
&\limsup_{k\to\infty}\frac{h(\tilde{y}_k+t_k\tilde{v}_k)-h(\tilde{y}_k)}{t_k}=\limsup_{k\to\infty}\frac{h(\tilde{y}_k+t_k\tilde{v})+o(t_k)-h(\tilde{y}_k)}{t_k}\\
=&\limsup_{k\to\infty}\frac{h(\tilde{y}_k+t_k\tilde{v})-h(\tilde{y}_k)}{t_k}\leq h^\circ(\tilde{y},\tilde{v})\,,
\end{aligned} \tag{7.4.5}$$

where we used (7.4.4) in the first equality, and with the inequality true by definition of the Clarke derivative. □

Proof of Lemma 7.4.1. With the notation introduced in the proof of Proposition 7.3.1, without loss of generality we assume that U is bounded and that φ can be extended to a neighborhood containing the closure of U.

First, since pushforward $\tilde{R}$ of a C^2 retraction on $\mathbb{R}$ is a C^2 retraction itself of $T\mathbb{R}^m$ on $\mathbb{R}^m$, we have the Taylor expansion

$$\tilde{R}(\tilde{y},\tilde{v})=\tilde{y}+\tilde{v}+O(\|\tilde{v}\|^2)\,, \tag{7.4.6}$$

with the implicit constant uniform for $\tilde{y}$ varying in $\tilde{U}$ and $\tilde{v}$ chosen in $\mathbb{R}^m$.

Second, for any fixed constant $B>0$, by continuity we have

$$\|\tilde{\Gamma}_{\tilde{x}}^{\tilde{x}_k}\tilde{q}-\tilde{q}\|\leq O\left(\|\tilde{x}-\tilde{x}_k\|\right)\,, \tag{7.4.7}$$

for $k\longrightarrow\infty$, $\tilde{q}\in\mathbb{R}^m$ with $\|\tilde{q}\|\leq B$, and with a uniform implicit constant.

Therefore

$$\begin{aligned}
&\|\tilde{d}_k-\tilde{d}\|\leq\|\tilde{d}_k-\tilde{\Gamma}_{\tilde{x}}^{\tilde{x}_k}\tilde{d}\|+\|\tilde{\Gamma}_{\tilde{x}}^{\tilde{x}_k}\tilde{d}-\tilde{d}\|\leq O\left(\|\tilde{d}_k-\tilde{\Gamma}_{\tilde{x}}^{\tilde{x}_k}(\tilde{d})\|_{\tilde{x}}\right)+O\left(\|\tilde{x}-\tilde{x}_k\|\right)\\
&=O\left(\|d_k-\Gamma_x^{x_k}(d)\|_x\right)+O\left(\|\tilde{x}-\tilde{x}_k\|\right)=o(1)\,,
\end{aligned} \tag{7.4.8}$$

where in the second inequality we used (7.4.7), and in the last equality we used $d_k\longrightarrow d$ together with $\tilde{x}_k\longrightarrow\tilde{x}$.

Let now $\tilde{v}_k=(\tilde{R}(\tilde{x}_k,t_k\tilde{d}_k)-\tilde{x}_k)/t_k$. Then

$$\begin{aligned}
&\|\tilde{v}_k-\tilde{d}\|=\frac{1}{t_k}\|\tilde{R}(\tilde{x}_k,t_k\tilde{d}_k)-\tilde{x}_k-t_k\tilde{d}\|\leq\frac{1}{t_k}(\|R(\tilde{x}_k,t_k\tilde{d}_k)-\tilde{x}_k-t_k\tilde{d}_k\|+t_k\|d_k-\tilde{d}_k\|)\\
=&\frac{1}{t_k}(O(t_k^2\|\tilde{d}_k\|^2)+t_ko(1))=o(1)\,,
\end{aligned} \tag{7.4.9}$$

where we used (7.4.6) and (7.4.8) for the first and the second summand in the second equality. In other words, $\tilde{v}_k\longrightarrow\tilde{d}$. To conclude,

$$\begin{aligned}
&\limsup_{k\to\infty}\frac{f(R(y_k,t_kd_k))-f(y_k)}{t_k}=\limsup_{k\to\infty}\frac{\tilde{f}(\tilde{R}(\tilde{y}_k,t_k\tilde{d}_k))-\tilde{f}(\tilde{y}_k)}{t_k}\\
=&\limsup_{k\to\infty}\frac{\tilde{f}(\tilde{y}_k+t_k\tilde{v}_k)-\tilde{f}(\tilde{y}_k)}{t_k}\geq\tilde{f}^\circ(\tilde{x},\tilde{d})=f^\circ(x,d)\,,
\end{aligned} \tag{7.4.10}$$

where in the inequality we were able to apply (7.4.2) because $\tilde{v}_k \to \tilde{d}$ by (7.4.9). □

7.4.2 Refining subsequences

We now adapt the definition of refining subsequence used in the analysis of direct search methods (see, e.g., [17, 98]) to the Riemannian setting. Let (x_k, d_k) be a sequence in $T\mathcal{M}$.

Definition 7.4.3. We say that the subsequence $\{x_{i(k)}\}$ is refining if $x_{i(k)} \to x$, and if for every $d \in T_x\mathcal{M}$ with $\|d\|_x = 1$ there is a further subsequence $\{j(i(k))\}$ such that

$$\lim_{k\to\infty} \operatorname{dist}^*(d_{j(i(k))}, d) = 0\,. \tag{7.4.11}$$

We now give a sufficient condition for a sequence to be refining, assuming that the manifold is embedded in $\mathbb{R}^n$ and that the directions are obtained projecting from the unit sphere to the tangent spaces.

Proposition 7.4.4. *If $x_{i(k)} \to x^*$, $\bar{d}_{i(k)}$ is dense in the unit sphere, and $d_{i(k)} = \mathsf{P}_k(\bar{d}_{i(k)})/\|\mathsf{P}_k(\bar{d}_{i(k)})\|_k$ for $\mathsf{P}_k(\bar{d}_{i(k)}) \neq 0$ and $d_{i(k)} = 0$ otherwise, then it holds that the subsequence $\{x_{i(k)}\}$ is refining.*

Proof. Fix $d \in T_{x^*}\mathcal{M}$, with $\|d\|_{x^*} = 1$, and let $\bar{d} = d/\|d\|$. By density, we have that $\bar{d}_{j(i(k))} \to \bar{d}$ for a proper choice of the subsequence $\{j(i(k))\}$. Then

$$\lim_{k\to\infty} d_{j(i(k))} = \lim_{k\to\infty} \frac{\mathsf{P}_k(\bar{d}_{j(i(k))})}{\|\mathsf{P}_k(\bar{d}_{j(i(k))})\|_k} = \frac{\mathsf{P}_{x^*}(\bar{d})}{\|\mathsf{P}_{x^*}(\bar{d})\|_{x^*}} = \frac{\bar{d}}{\|\bar{d}\|_{x^*}} = d\,, \tag{7.4.12}$$

where in the second equality we used the continuity of P_x and of the norm $\|\cdot\|_x$, and in the third equality we used $\mathsf{P}_{x^*}(\bar{d}) = \bar{d}$ since $\bar{d} \in T_{x^*}\mathcal{M}$ by construction. □

7.4.3 Direct search for nonsmooth objectives

We present here our Riemannian Direct Search method based on Dense Directions (RDS-DD) for nonsmooth objectives. The scheme is presented in detail as Algorithm 16. The algorithm performs three simple steps at an iteration k. First, a search direction is selected randomly in the current tangent space. Then a tentative point is generated by retracting the step $\alpha_k d_k$ from the tangent space to the manifold. Such a point is then eventually accepted as the new iterate if a sufficient decrease condition of the objective function is satisfied (and the stepsize is expanded), otherwise the iterate stays the same (and the stepsize is reduced).

Algorithm 16 RDS-DD

```
Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 > 0$, $\gamma > 0, \gamma_1 \in (0,1), \gamma_2 \geq 1$
for  $k = 0, 1, \ldots$ do
    Sample $d_k$ randomly in $\{d \in T_k\mathcal{M} \mid \|d\| = 1\}$
    if $f(R(x_k, \alpha_k d_k)) \leq f(x) - \gamma\alpha_k^2$ then
        $x_{k+1} = R(x_k, \alpha_k d_k)$, $\alpha_{k+1} = \gamma_2\alpha_k$
    else
        $x_{k+1} = x_k$, $\alpha_{k+1} = \gamma_1\alpha_k$
    end if
end for
```

Thanks to the theoretical tools previously introduced, and in particular to the relation between retractions and the Clarke directional derivative proved in Lemma 7.4.1, we can easily show that a suitable subsequence of unsuccessful iterations of the RDS-DD method converges to a Clarke stationary point.

Theorem 7.4.5. *Let $\{x_k\}$ be generated by Algorithm 16. If $\{x_{i(k)}\}$ is refining, with $x_{i(k)} \to x^*$, and $i(k)$ is an unsuccessful iteration for every $k \in \mathbb{N} \cup \{0\}$, x^* is Clarke stationary.*

Proof. By the same assumptions as in the smooth case $\alpha_k \to 0$ and in particular $\alpha_{i(k)} \to 0$. Since by assumption $i(k)$ is an unsuccessful step, we have, for every $i(k)$

$$f(R(x_{i(k)}, \alpha_{i(k)} d_{i(k)})) - f(x_{i(k)}) > -\gamma\alpha_{i(k)}^2 \,. \tag{7.4.13}$$

Let $\{j(i(k))\}$ be such that $d_{j(i(k))} \to d$, and let $y_k = x_{j(i(k))}$, $q_k = d_{j(i(k))}$, $t_k = \alpha_{j(i(k))}$. We have

$$\limsup_{k\to\infty} \frac{f(R(y_k, t_k q_k)) - f(y_k)}{t_k} \geq \limsup_{k\to\infty} -\gamma\alpha_{i(k)} = 0 \,, \tag{7.4.14}$$

thanks to (7.4.13), and by applying Lemma 7.4.1 we get

$$f^\circ(x^*, d) \geq \limsup_{k\to\infty} \frac{f(R(y_k, t_k q_k)) - f(y_k)}{t_k} \geq 0 \,, \tag{7.4.15}$$

which implies the thesis since d is arbitrary. □

7.4.4 Direct search with line search extrapolation for nonsmooth objectives

We present here our Riemannian Direct Search method with line search Extrapolation based on Dense Directions (RDSE-DD) for nonsmooth objectives. It can be seen as an extension to the Riemannian setting of the DFN$_{simple}$ algorithm introduced in [98] for the bound constrained setting. The detailed scheme is given in Algorithm 17. As we can easily see, the algorithm performs just two simple steps at an iteration k. First, a given search direction is suitably projected on the current tangent space. Then a line search is performed using Algorithm 15 to hopefully obtain a new point that guarantees a sufficient decrease.

Algorithm 17 RDSE-DD

```
Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 > 0$, $\gamma > 0, \gamma_1 \in (0,1), \gamma_2 \geq 1$, $\{\bar{d_k}\}$ dense in $S(0,1)$.
for $k = 0, 1, \ldots$ do
    Sample $d_k$ randomly in $\{d \in T_k\mathcal{M} \mid \|d\| = 1\}$
    Compute $\alpha_k, \tilde{\alpha}_{k+1}$ with Linesearchprocedure($\tilde{\alpha}_k, x_k, d_k, \gamma, \gamma_1, \gamma_2$)
    Set $x_{k+1} = R(x_k, \alpha_k d_k)$
end for
```

Once again, by exploiting the theoretical tools previously introduced, we can straightforwardly prove that a suitable subsequence of the RDSE-DD iterations converges to a Clarke stationary point. It is interesting to notice that, thanks to the use of the line search strategy, we are not restricted to considering unsuccessful iterations this time.

Theorem 7.4.6. *Let $\{x_k\}$ be generated by Algorithm 17. If $\{x_{i(k)}\}$ is refining, with $x_{i(k)} \to x^*$, then x^* is Clarke stationary.*

Proof. Let $\beta_k = \tilde{\alpha}_k/\gamma_2$ if the line search procedure exits before the loop, and $\beta_k = \gamma_1\tilde{\alpha}_{k+1}$ otherwise. Clearly $\beta_k \to 0$, and by definition of the line search procedure, for every k

$$f(R(x_k, \beta_k d_k)) - f(x_k) > -\gamma\beta_k^2 \,. \tag{7.4.16}$$

The rest of the proof is analogous to that of Theorem 7.4.5. □

7.5 Numerical results

We now report the results of some numerical experiments of the algorithms described in this chapter on a set of simple but illustrative example problems. The comparison among the algorithms is carried out by using data and performance profiles [186]. Specifically, let S be a set of algorithms and P a set of problems. For each $s \in S$ and $p \in P$, let $t_{p,s}$ be the number of function evaluations required by algorithm s on problem p to satisfy the condition

$$f(x_k) \leq f_L + \tau(f(x_0) - f_L)\,, \tag{7.5.1}$$

where $0 < \tau < 1$ and f_L is the best objective function value achieved by any solver on problem p. Then, the performance and data profiles of solver s are defined, respectively, by the following functions

$$\begin{aligned} \rho_s(\alpha) &= \frac{1}{|P|}\left|\left\{p \in P : \frac{t_{p,s}}{\min\{t_{p,s'} : s' \in S\}} \leq \alpha\right\}\right|, \\ d_s(\kappa) &= \frac{1}{|P|}\left|\left\{p \in P : t_{p,s} \leq \kappa(n_p + 1)\right\}\right|, \end{aligned}$$

where n_p is the dimension of problem p.

We used a budget of $100(n_p + 1)$ function evaluations in all cases and two different precisions for the condition (7.5.1), that is $\tau \in \{10^{-1}, 10^{-3}\}$. We consider randomly generated instances of well-known optimization problems over manifolds from [3,55,130]. The size of the ambient space for the instances varies from 2 to 200. For all the problems, the manifold structure we used was the one available in the MANOPT library [54]. After a basic tuning phase, we set the algorithm parameters as follows: we used $\gamma_1 = 0.61$, $\gamma_2 = 1$ and $\gamma = 0.77$ for Algorithm 13, $\gamma_1 = 0.81$, $\gamma_2 = 3.12$ and $\gamma = 0.11$ for Algorithm 14, and the stepsize $1.64/n$ (recall that n is the dimension of the ambient space) for the ZO-RGD method.

For the nonsmooth strategies RDS-DD+ and RDSE-DD+, we considered the same parameters of the smooth case for RDS-SB and RDSE-SB, setting $\alpha_\epsilon = 10^{-4}$, and for both RDS-DD and RDSE-DD used $\gamma_1 = 0.95$, $\gamma_2 = 2$, and $\gamma = 1$. When dealing with the nonsmooth case, the stepsize used for ZO-RGD was the same as the one considered in the smooth case.

The positive spanning basis was obtained both in Algorithm 13 and Algorithm 14 by projecting the positive spanning basis $(e_1, ..., e_n, -e_1, ..., -e_n)$ of the ambient space $\mathbb{R}^n$ on the tangent space. The initial stepsize was set to 1 for all the direct search methods, with no fine tuning.

We generated the starting point and the parameters related to the instances either with MATLAB rand function or by using the random element generators implemented in the MANOPT library.

7.5.1 Smooth problems

We describe here the 8 smooth instances of problem (7.2.1) from [3, 55].

Largest eigenvalue, singular value, and top singular values problem

In the largest eigenvalue problem [55, Section 2.3], given a symmetric matrix $A \in S(n,n) = \{A \in \mathbb{R}^{n\times n} \mid A = A^\top\}$, we are interested in computing

$$\max_{x\in\mathbb{S}^{n-1}} x^\top A x\,. \tag{7.5.2}$$

The largest singular value problem [55, Section 2.3] can be formulated generalizing (7.5.2): given $A \in \mathbb{R}^{m\times h}$, we are interested in

$$\max_{x\in\mathbb{S}^{m-1}, y\in\mathbb{S}^{h-1}} x^\top A y\,. \tag{7.5.3}$$

Notice how the domain in (7.5.2) and (7.5.3) are a sphere and the product of two spheres respectively.
Finally, to compute the sum of the top r singular values, as explained in [55, Section 2.5] it suffices to solve

$$\max_{X\in S(m,r), Y\in S(h,r)} X^\top A Y\,, \tag{7.5.4}$$

for $S(a,b)$ the Stiefel manifold with dimensions (a,b).

Dictionary learning

The dictionary learning problem [55, Section 2.4] can be formulated as

$$\begin{aligned} \min \quad & \|Y - DC\| + \lambda\|C\|_1, \\ \text{s.t.} \quad & D \in \mathbb{R}^{d\times h}, C \in \mathbb{R}^{h\times k},\ \|D_1\| = \dots = \|D_h\| = 1\,, \end{aligned} \tag{7.5.5}$$

for a fixed $Y \in \mathbb{R}^{d\times k}$, $\lambda > 0$, $\|\cdot\|_1$ the ℓ_1- norm, and $D_1, \dots, D_h$ the columns of D.
In our implementation we smooth the objective by using a smoothed version $\|\cdot\|_{1,\varepsilon}$ of $\|\cdot\|_1$

$$\|C\|_{1,\varepsilon} = \sum_{i,j} \sqrt{C_{i,j}^2 + \varepsilon^2}\,. \tag{7.5.6}$$

In our tests, we generated the solution $\bar{C}$ using MATLAB sprand function, with a density of 0.3, set the regularization parameter λ to 0.01 and ε to 0.001.

Synchronization of rotations

Let $\mathrm{SO}(d)$ be the special orthogonal group:

$$\mathrm{SO}(d) = \{R \in \mathbb{R}^{d\times d} \mid R^\top R = I_d \text{ and } \det(R) = 1\}\,. \tag{7.5.7}$$

In the synchronization of rotations problem [55, Section 2.6], we need to find rotations $R_1, \dots, R_h \in \mathrm{SO}(d)$ from noisy measurements H_{ij} of $R_i R_j^{-1}$, for every $(i, j) \in E$, a subset of $\binom{h}{2}$ (the set of couples of distinct elements in $[1 : h]$). The objective is then

$$\min_{\hat{R}_1,\dots,\hat{R}_h \in \mathrm{SO}(d)} \sum_{(i,j)\in E} \|\hat{R}_i - H_{ij}\hat{R}_j\|^2 \,. \tag{7.5.8}$$

In our tests, we considered the case $h = 2$ for simplicity.

Low-rank matrix completion

The low rank matrix completion problem [55, Section 2.7] can be written, for a fixed matrix $M \in \mathbb{R}^{m\times h}$, as

$$\begin{array}{rl} \min & \sum_{(i,j)\in\Omega}(X_{ij} - M_{ij})^2, \\ s.t. & X \in \mathbb{R}^{m\times h},\ \mathrm{rank}(X) = r\,, \end{array} \tag{7.5.9}$$

given a positive integer $r > 0$ and a subset of indices $\Omega \subset [1 : m] \times [1 : h]$. It can be proven that the optimization domain, that is the matrices in $\mathbb{R}^{m\times n}$ with fixed rank r, can be given a Riemannian manifold structure (see, e.g., [225]).

Gaussian mixture models

In the Gaussian mixture model problem [55, Section 2.8], we are interested in computing a maximum likelihood estimation for a given set of observations $x_1, \dots, x_h$:

$$\max_{\substack{\hat{u}_1,\dots,\hat{u}_k \in \mathbb{R}^d \\ \hat{\Sigma}_1,\dots,\hat{\Sigma}_k \in \mathrm{Sym}(d)^+, \\ w \in \Delta_+^{K-1}}} \sum_{i=1}^{h} \log\left(\sum_{k=1}^{K} w_k \frac{1}{\sqrt{2\pi \det(\Sigma_k)}} e^{\frac{(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)}{2}} \right), \tag{7.5.10}$$

where $\mathrm{Sym}(d)^+$ is the manifold of positive definite matrices

$$\mathrm{Sym}(d)^+ = \{X \in \mathbb{R}^{d\times d} \mid X = X^\top,\ X \succ 0\} \tag{7.5.11}$$

and Δ_+^{K-1} is the subset of strictly positive elements of the simplex Δ^{K-1}, which can be given a manifold structure. In our tests, we considered the case $K = 2$ and the reformulation proposed in [128], which does not use the unconstrained variables $(\hat{u}_1, \dots, \hat{u}_k)$.

Procrustes problem

The Procrustes problem [3] is the following linear regression problem, for fixed $A \in \mathbb{R}^{l\times n}$ and $B \in \mathbb{R}^{l\times p}$:

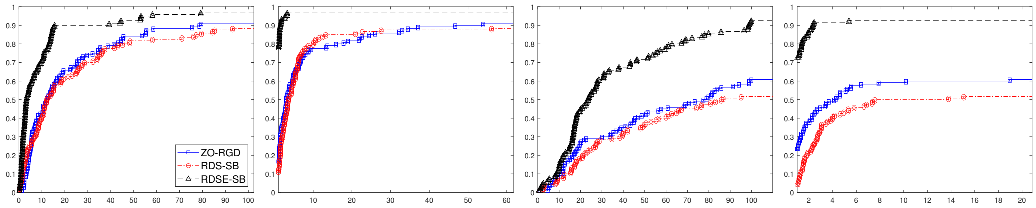
$$\min_{x\in\mathcal{M}} \|AX - B\|_F^2 \,, \tag{7.5.12}$$

In our tests, we assumed the variable $X \in \mathbb{R}^{n\times p}$ to be in the Stiefel manifold $\mathrm{St}(n, p)$, a choice leading to the so called unbalanced orthogonal Procrustes problem.

Results

In Figure 7.1, we include the results related to the 8 smooth instances of problem (7.2.1) discussed above, each with 15 different problem dimensions (from 2 to 200), for a total number of 60 tested instances. We compared our methods, that is RDS-SB and RDSE-SB, with the zeroth order gradient descent (ZO-RGD, [171, Algorithm 1]).

The results clearly show that RDSE-SB performs better than RDS-SB and ZO-RGD both in efficiency and reliability for both levels of precision. By taking a look at the detailed results in Section 7.5.4, we can also see how the gap between RDSE-SB and the other two algorithms gets larger as the problem dimension grows.



(a) Data p., $\tau = 10^{-1}$ **(b)** Perf. p., $\tau = 10^{-1}$ **(c)** Data p., $\tau = 10^{-3}$ **(d)** Perf. p., $\tau = 10^{-3}$

Figure 7.1: Smooth case: results for all the instances

7.5.2 Nonsmooth problems

We report two nonsmooth problems taken from [130].

Sparsest vector in a subspace

Given an orthonormal matrix $Q \in \mathbb{R}^{m\times n}$, the problem of finding the sparsest vector in the subspace generated by the columns of Q can be relaxed as

$$\min_{x\in\mathbb{S}^{n-1}} \|Qx\|_1 \,. \tag{7.5.13}$$

Nonsmooth low-rank matrix completion

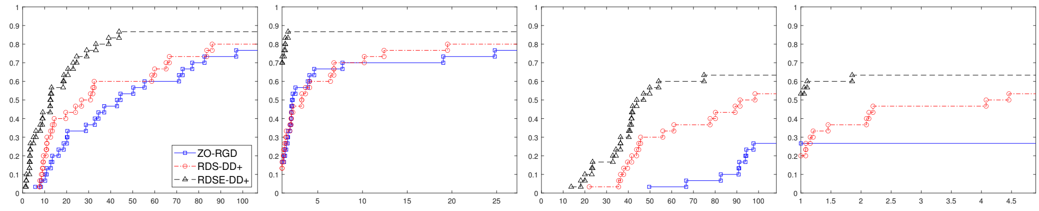
In the nonsmooth version of the low rank matrix completion problem (7.5.9) the Euclidean norm is replaced with the l_1 norm, so that in the objective we have a sum of absolute values:

$$\begin{array}{rl} \min & \sum_{(i,j)\in\Omega} |X_{ij} - M_{ij}|, \\ s.t. & X \in \mathbb{R}^{m\times n},\ \operatorname{rank}(X) = r\,. \end{array} \tag{7.5.14}$$

7.5.3 Results

We report here a preliminary comparison between a direct search strategy, a line search strategy and ZO-RGD on the two nonsmooth instances of (7.2.1) presented above, each with 15 different problem sizes (from 2 to 200), thus getting a total number of 30 tested instances. We remark that while in the unconstrained setting the performance of zeroth order (sub)gradient descent methods on nonsmooth objectives have been analyzed (see, e.g., [193]), there are, to the best of our knowledge, no convergence guarantees in the Riemannian setting.

In the direct search strategy (RDS-DD+), we apply the RDS-SB method until $\alpha_{k+1} \leq \alpha_\epsilon$, at which point we switch to the nonsmooth version RDS-DD. Analogously, in the line search strategy (RDSE-DD+), we apply the RDSE-SB method until $\max_{j\in[1:K]} \tilde{\alpha}^j_{k+1} \leq \alpha_\epsilon$, at which point we switch to the nonsmooth version RDSE-DD. Both strategies use a threshold parameter $\alpha_\epsilon > 0$ to switch from the smooth to the nonsmooth DFO algorithm. We refer the reader to [98] and references therein for other direct search strategies combining coordinate and dense directions. We report, in Figure 7.2, the comparison between the considered strategies. As in the smooth case, the line search based strategy outperforms both the simple direct search and the zeroth order one. By taking a look at the detailed results in Section 7.5.4, we can once again see how the gap between the algorithms gets larger as the problem dimension gets large enough.

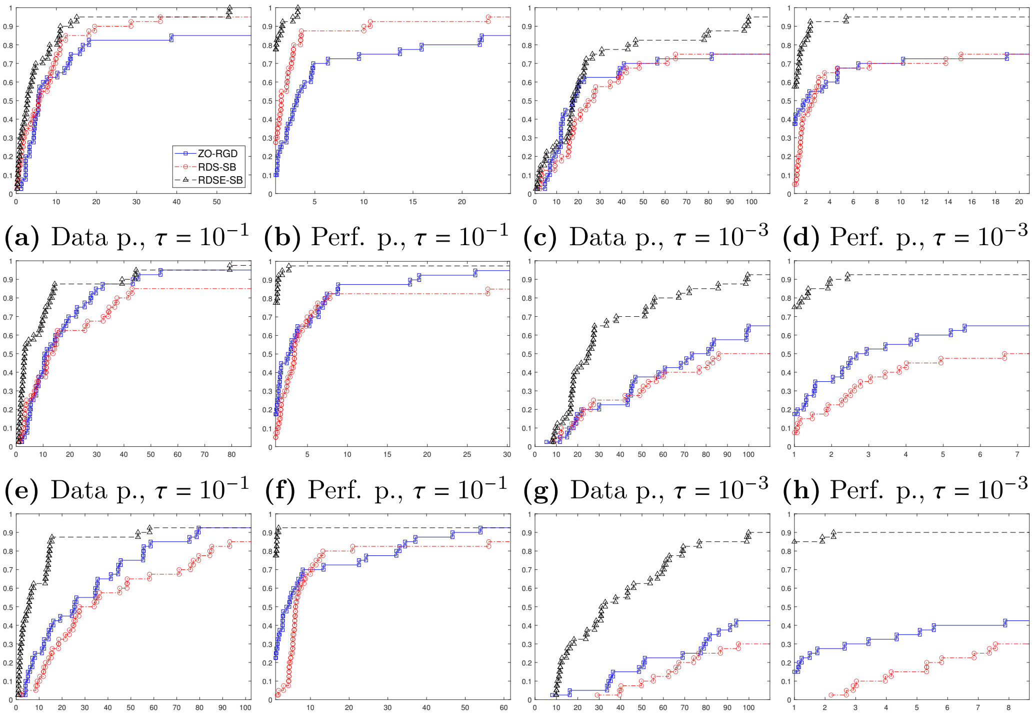


(a) Data p., $\tau = 10^{-1}$ **(b)** Perf. p., $\tau = 10^{-1}$ **(c)** Data p., $\tau = 10^{-3}$ **(d)** Perf. p., $\tau = 10^{-3}$

Figure 7.2: Nonsmooth case: results for all the instances

7.5.4 Data and performance profiles by ambient space dimension

We report here further detailed numerical results, splitting the problems by ambient space dimension: between 2 and 15 for small instances, between 16 and 50 for medium instances, and between 51 and 200 for large instances.

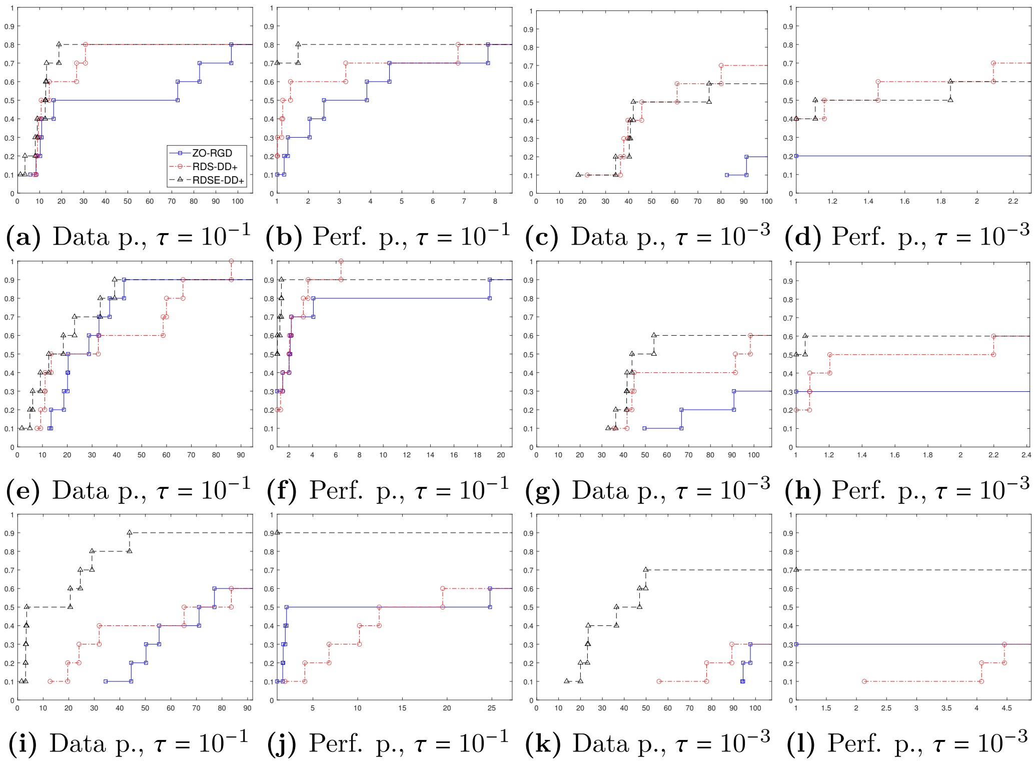


(a) Data p., $\tau = 10^{-1}$ **(b)** Perf. p., $\tau = 10^{-1}$ **(c)** Data p., $\tau = 10^{-3}$ **(d)** Perf. p., $\tau = 10^{-3}$

(e) Data p., $\tau = 10^{-1}$ **(f)** Perf. p., $\tau = 10^{-1}$ **(g)** Data p., $\tau = 10^{-3}$ **(h)** Perf. p., $\tau = 10^{-3}$

(i) Data p., $\tau = 10^{-1}$ **(j)** Perf. p., $\tau = 10^{-1}$ **(k)** Data p., $\tau = 10^{-3}$ **(l)** Perf. p., $\tau = 10^{-3}$

Figure 7.3: From top to bottom: results for small, medium and large instances in the smooth case.



(a) Data p., $\tau = 10^{-1}$ **(b)** Perf. p., $\tau = 10^{-1}$ **(c)** Data p., $\tau = 10^{-3}$ **(d)** Perf. p., $\tau = 10^{-3}$

(e) Data p., $\tau = 10^{-1}$ **(f)** Perf. p., $\tau = 10^{-1}$ **(g)** Data p., $\tau = 10^{-3}$ **(h)** Perf. p., $\tau = 10^{-3}$

(i) Data p., $\tau = 10^{-1}$ **(j)** Perf. p., $\tau = 10^{-1}$ **(k)** Data p., $\tau = 10^{-3}$ **(l)** Perf. p., $\tau = 10^{-3}$

Figure 7.4: From top to bottom: results for small, medium and large instances in the nonsmooth case.

Chapter 8

Convergence of direct search under a tail bound condition on the black box error

In this chapter, we use tail bounds to define a tailored probabilistic condition for function estimation that eases the theoretical analysis of a stochastic direct search method. In particular, we focus on the unconstrained minimization of a potentially non-smooth function, whose values can only be estimated via stochastic observations, and give a simplified convergence proof for a basic direct search scheme. We also study the trade-off between algorithm parameters, assumptions on the noise, and number of samples needed at every iteration for convergence.

8.1 Derivative free optimization with stochastic oracles

We consider the following unconstrained optimization problem

$$\min_{x\in\mathbb{R}^n} f(x)\,, \tag{8.1.1}$$

with f locally Lipschitz continuous and possibly non-smooth function such that $\inf f = f^* \in \mathbb{R}$. We assume that the original function $f(x)$ is not computable, and the only information available on f is given by a stochastic oracle producing an estimate $\tilde{f}(x)$ for any $x \in \mathbb{R}^n$. In some contexts, we can assume that the estimate is a random variable parameterized by x, that is

$$\tilde{f}(x) = F(x, \xi),$$

with the black-box oracle given by sampling on the ξ space. When dealing with, e.g., statistical learning problems, the function $F(x, \xi)$ evaluates the loss of the decision rule parametrized by x on a data point ξ (see, e.g., [160] for further details). In simulation-based engineering applications, the function $F(x, \xi)$ is simply related to some noisy computable version of the original function. In this case ξ represents the random variable that induces the noise (a classic example is given by Monte Carlo simulations). A detailed overview is given in, e.g., [11].

When this random variable is exact in expected value, problem (8.1.1) turns out to be the expected loss formulation

$$\min_{x\in\mathbb{R}^n} \mathbb{E}_\xi[F(x, \xi)], \tag{8.1.2}$$

a case addressed in recent literature, see, e.g., [162, 215], for further details.

Although the role of derivative-free optimization is particularly important when the black-box representing the function is somehow noisy or, in general, of a stochastic type, traditional DFO methods have been developed primarily for deterministic functions, and only recently adapted to deal with stochastic observations (see, e.g., [74] for a detailed discussion on this matter). We give here a brief overview of the main results available in the literature by first focusing on *trust region* strategies and then moving to *direct search* approaches.

In [162], the authors describe a trust-region algorithm to handle noisy objectives and prove convergence when f is sufficiently smooth (i.e., with Lipschitz continuous gradient) and the noise is drawn independently from a distribution with zero mean and finite variance, that is they aim at solving a smooth version of problem (8.1.2), when ξ is additive noise. In the same line of research, the authors in [215] developed a class of derivative-free trust-region algorithms, called ASTRO-DF, for unconstrained optimization problems whose objective function has Lipschitz continuous gradient and can only be implicitly expressed via a Monte Carlo oracle. The authors consider again an objective with noise drawn independently from a distribution with zero mean, finite variance and a bound on the $4v$-th moment (with $v \geq 2$), and prove the almost sure convergence of their method when using stochastic polynomial interpolation models. Another relevant reference in this context is given by [74], where the authors analyze a trust-region model-based algorithm for solving unconstrained stochastic optimization problems. They consider random models of a smooth objective function, obtained from stochastic observations of the function or its gradient. Convergence rates for this class of methods are reported in [34, 68]. The frameworks analyzed in [34, 68, 74] extend the trust region DFO method based on probabilistic models described in [23]. It is important to notice that the randomness in the models described in [23] comes from the way sample points are chosen, rather than from noise in the function evaluations.

All the above-mentioned model-based approaches consider functions with a certain degree of smoothness (e.g., with Lipschitz continuous gradient) and assume that a probabilistically accurate gradient estimate (e.g., some kind of probabilistically fully-linear model) can be generated, while of course such an estimate is not available when dealing with non-smooth functions.

A detailed convergence rate analysis of stochastic direct search variants is reported in [96] for the smooth case, i.e., for an objective function with Lipschitz continuous gradient. A stochastic mesh adaptive direct search for black-box nonsmooth optimization is proposed in [14]. The authors prove convergence with probability one to a Clarke stationary point (see [77]) of the objective function by assuming that stochastic observations are sufficiently accurate and satisfy a variance condition. The analysis adapts to the considered gradient-free framework the theoretical analysis given in [198] for a class of stochastic gradient-based methods. It is extended in [97] to the constrained case.

8.1.1 Contributions

The main goal of this chapter is to analyze some tail-bound probabilistic conditions for the error of a black box used within a general direct search scheme. We show how they can be used to obtain convergence and define a trade-off between noise, algorithm parameters, and number of samples.

Our algorithmic scheme is a simple direct search strategy obtained by replacing the function values with their estimates in the acceptance test of the deterministic counterpart. The scheme works as follows: it chooses a direction over the unit sphere; generates the new iterate by moving along the direction, and finally it uses a suitable acceptance test to decide if the new point can be accepted (successful iteration) or not. Convergence of the method is then carried out by simply assuming that our tail-bounds hold. The analysis has two main steps. In the first one, we show a result that implies convergence of the stepsize to zero almost surely. In the second one, we focus on the random sequence of the unsuccessful iterations and prove, by exploiting the first result, Clarke stationarity at limit points.

We will see how:

- our conditions are implied by the variance conditions considered in [14] and by the probabilistically accurate function estimate assumption used in [14,74,198];
- one of our conditions is implied by a tail bound used in [162];
- the finite variance oracle usually considered in the literature (see, e.g., [162, 215]) can be replaced by a finite moment oracle (see Section 8.2.5 for further details) when constructing estimates satisfying our conditions.
- we can compute the number of samples needed for convergence as a function of the stepsize exponent used in the acceptance test and the moments of the noise. One of our results is that if all the moments are finite like in the case of gaussian noise we only need $O(\Delta_k^{-2-\varepsilon})$ samples with $\varepsilon > 0$ for a suitable choice of the stepsize exponent, instead of the $O(\Delta_k^{-4})$ samples required in previous works on stochastic trust region (see, e.g., [34, 74, 215]) and direct search (see [14, 96, 97]) methods, where Δ_k is the stepsize at the step k (see Remark 8.2.10).

8.2 A weak tail-bound probabilistic condition for function estimation

In order to give convergence results for our algorithm, we first need some probabilistic assumptions on the accuracy of the oracle. In this section, we hence describe our tail-bound conditions and compare them with other existing conditions from the literature. The stochastic quantities defined hereafter lie in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with probability measure $\mathbb{P}$ and σ-algebra $\mathcal{F}$ containing subsets of Ω, that is the space of the realizations of the algorithms under analysis. Any single outcome of the sample space Ω will be denoted by w. For a random variable X defined in Ω we use the shorthand $\{X \in A\}$ to denote $\{w \mid X(w) \in A\}$.
Our algorithm generates a random process with the following random variables and corresponding realizations. The search direction and the stepsize are denoted with Δ_k and G_k, with realizations δ_k and g_k respectively. The function values $f(x_k)$ and $f(x_k + \Delta_k G_k)$ are denoted with F_k and F_k^g, with realizations f_k and f_k^g respectively. We define $\mathcal{F}_{k-1}$ as the $\sigma-$algebra of events up to the choice of G_k (so that in particular G_k is measurable with respect to $\mathcal{F}_{k-1}$). More explicitly, we define $\mathcal{F}_{k-1}$ as

the σ-algebra generated by $(F_j, F_j^g)_{j=0}^{k-1}$ and $(G_j)_{j=0}^{k}$. Finally, we use $\mathbb{E}$ to denote expectation and conditional expectation, and a.s. as a shorthand for "almost surely".

8.2.1 The weak tail-bound probabilistic condition

We now introduce our tail bound assumptions.

Assumption 8.1. For every $\alpha > 0$ and some $\varepsilon_f > 0, q > 1$ (independent of α, k), a.s.:

$$\mathbb{P}\left(|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \geq \alpha \Delta_k^q \; |\mathcal{F}_{k-1}\right) \leq \frac{\varepsilon_f}{\alpha}\,. \tag{A1}$$

Assumption 8.2. For every $\alpha > 0$ and some $\varepsilon_q > 0, p > 1, q > 1$ (independent of α, k), a.s.:

$$\mathbb{P}\left(|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \geq \alpha \Delta_k^{1+q/p} \; |\mathcal{F}_{k-1}\right) \leq \frac{\varepsilon_q}{\alpha^p}\,. \tag{A2}$$

Notice that we are only assuming error bounds for the estimate of the difference $f(x_k) - f(x_k + \Delta_k G_k)$ and not for the estimates of $f(x_k)$ and $f(x_k + \Delta_k G_k)$ taken individually; we basically want to bound the probability that the error in that estimate is large, as such an estimation plays a crucial role in the acceptance tests of our algorithm. If $p = q = 2$, condition (A2) implies (A1) for $\varepsilon_f = \max(1, \varepsilon_q)$, as it can be seen using that the LHS are the same while the RHS are $O(\frac{1}{\alpha^2})$ and $O(\frac{1}{\alpha})$ respectively.

In our convergence arguments we will need Assumptions 8.1 and 8.2 with a $\mathcal{F}_{k-1}$ measurable random variable A rather than a real number α. This is justified by the following lemma.

Lemma 8.2.1. *Let* A *be a positive* $\mathcal{F}_{k-1}$ *measurable random variable. If* (A1) *holds, then it holds also with* A *instead of* α*, and an analogous result is true for* (A2).

Proof. We prove the result in the case where A is a discrete random variable with a countable set of possible realizations $\{a_i\}_{i\in\mathbb{N}}$, which is sufficient since the general case then follows by approximation.

Let $X = |F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))|/\Delta_k^q$. By the definition of conditional probability, (A1) holds with A instead of α iff, for every $F \in \mathcal{F}_{k-1}$:

$$\mathbb{E}[\mathbb{1}_F \mathbb{1}_{\{X \leq \mathrm{A}\}}] \leq \mathbb{E}[\mathbb{1}_F \frac{\varepsilon_f}{\mathrm{A}}]\,. \tag{8.2.1}$$

Indeed we have

$$\begin{aligned}\mathbb{E}[\mathbb{1}_F \mathbb{1}_{\{X \leq \mathrm{A}\}}] &= \sum_{i \in \mathbb{N}} \mathbb{E}[\mathbb{1}_F \mathbb{1}_{\{X \leq \mathrm{A}\}} \mathbb{1}_{\{\mathrm{A}=a_i\}}] = \sum_{i \in \mathbb{N}} \mathbb{E}[\mathbb{1}_{F \cap \{\mathrm{A}=a_i\}} \mathbb{1}_{\{X \leq a_i\}}] \\ &\leq \sum_{i \in \mathbb{N}} \mathbb{E}[\mathbb{1}_{F \cap \{\mathrm{A}=a_i\}} \frac{\varepsilon_f}{a_i}] = \sum_{i \in \mathbb{N}} \mathbb{E}[\mathbb{1}_F \mathbb{1}_{\{\mathrm{A}=a_i\}} \frac{\varepsilon_f}{a_i}] = \mathbb{E}[\mathbb{1}_F \frac{\varepsilon_f}{\mathrm{A}}]\end{aligned} \tag{8.2.2}$$

as desired, where we used that $F \cap \{\mathrm{A} = a_i\}$ is measurable w.r.t. $\mathcal{F}_{k-1}$ together with (A1) for $\alpha = a_i$ in the inequality.
This proves the Lemma for (A1), and an analogous argument holds for (A2). □

8.2.2 Conditional Chebycheff's inequality

We briefly recall here for completeness the conditional Chebycheff's inequality, which will be a key tool to relate our assumptions with other used in previous works. Thanks to the properties of conditional expectations, this inequality can be proved in the same way as the standard Chebycheff's inequality.

Proposition 8.2.2. *Given random variables X, ϵ defined on $\mathbb{R}^n$ with $\epsilon > 0$ measurable with respect to a sub σ-field $\mathcal{F}$, we have*

$$\mathbb{P}(|X| \geq \epsilon \mid \mathcal{F}) \leq \frac{\mathbb{E}[|X| \mid \mathcal{F}]}{\epsilon}.$$

Proof. We have

$$\begin{aligned}\epsilon \mathbb{P}(|X| \geq \epsilon \mid \mathcal{F}) &= \epsilon \mathbb{E}[\mathbb{1}_{|X| \geq \epsilon} \mid \mathcal{F}] \\ &= \mathbb{E}[\epsilon \mathbb{1}_{|X| \geq \epsilon} \mid \mathcal{F}] \leq \mathbb{E}[|X| \mid \mathcal{F}],\end{aligned}$$

where we used that ϵ is $\mathcal{F}$ measurable in the second equality and the monotonicity of the conditional expectation together with $\epsilon \mathbb{1}_{|X| \geq \epsilon} \leq |X|$ in the inequality. □

Remark 8.2.3. Alternative proofs to Lemma 8.2.1 without approximation arguments and to Proposition 8.2.2 can be given using [52, Theorem 3.1.1] in a straightforward way (see, e.g., [52, Corollary 3.1.1]).

8.2.3 Comparison with the existing conditions

Our conditions are weaker than the ones imposed in [14]. More precisely, they are implied by [14, Equation (2)], rewritten in our notation as

$$\begin{aligned}\mathbb{E}[|F_k^g - f(x_k + \Delta_k G_k)|^2 \mid \mathcal{F}_{k-1}] &\leq k_f^2 \Delta_k^4 \\ \mathbb{E}[|F_k - f(x_k)|^2 \mid \mathcal{F}_{k-1}] &\leq k_f^2 \Delta_k^4,\end{aligned} \tag{8.2.3}$$

for a constant $k_f > 0$. The k_f-variance condition in (8.2.3) is a gradient free version of [198, Assumption 2.4, (iii)], and more precisely can be obtained from the latter by removing the gradient related terms in the right hand side. However, in [198] as well as in other works on smooth stochastic derivative free optimization (see, e.g., [74,162,215] and references therein), a probabilistically accurate gradient estimate is also used, while of course such an estimate is not available in a possibly non-smooth setting.

Proposition 8.2.4. *Condition* (8.2.3) *implies Assumption 8.1 and Assumption 8.2 for* $\varepsilon_f = 2k_f$ *and* $\varepsilon_q = 4k_f^2, p = 2$ *respectively, and* $q = 2$.

Proof. First, notice that

$$\begin{aligned}&\mathbb{E}[|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))|^2 \mid \mathcal{F}_{k-1}]\\ &\leq 2(\mathbb{E}[|F_k^g - f(x_k + \Delta_k G_k)|^2 \mid \mathcal{F}_{k-1}] + \mathbb{E}[|F_k - f(x_k)|^2 \mid \mathcal{F}_{k-1}])\\ &\leq 4k_f^2\Delta_k^4,\end{aligned} \tag{8.2.4}$$

where we used $(a+b)^2 \leq 2(a^2+b^2)$ for $a, b \in \mathbb{R}$ in the first inequality, and (8.2.3) in the second.

We now prove (A1). In order to do so, we only need a bound on the first moment $\mathbb{E}[|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \mid \mathcal{F}_{k-1}]$, implied by the bound on the second moment (8.2.4) thanks to conditional Jensen's inequality:

$$\begin{aligned}&\mathbb{E}[|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \mid \mathcal{F}_{k-1}]\\ &\leq \sqrt{\mathbb{E}[|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))|^2 \mid \mathcal{F}_{k-1}]} \leq 2k_f\Delta_k^2.\end{aligned} \tag{8.2.5}$$

We can now conclude by noticing

$$\begin{aligned}&\mathbb{P}(|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \geq \alpha\Delta_k^2 \, |\mathcal{F}_{k-1})\\ &\leq \frac{\mathbb{E}(|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \, |\mathcal{F}_{k-1})}{\alpha\Delta_k^2} \leq \frac{2k_f}{\alpha},\end{aligned} \tag{8.2.6}$$

where we used the conditional Chebyshev's inequality in the first inequality, and (8.2.5) in the second inequality. In particular, (8.2.3) implies (A1) for $\varepsilon_f = 2k_f$.

As for (A2), we have

$$\begin{aligned}&\mathbb{P}[|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \geq \alpha\Delta_k^2 \mid \mathcal{F}_{k-1}]\\ &= \mathbb{P}[|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))|^2 \geq \alpha^2\Delta_k^4 \mid \mathcal{F}_{k-1}]\\ &\leq \frac{\mathbb{E}[|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))|^2 \mid \mathcal{F}_{k-1}]}{\alpha^2\Delta_k^4} \leq \frac{4k_f^2}{\alpha^2},\end{aligned}$$

where we used the conditional Chebyshev's inequality in the first inequality, and (8.2.4) in the second inequality. By setting $\varepsilon_q = 4k_f^2$ in the above equation we obtain

$$\mathbb{P}[|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \geq \alpha \Delta_k^2 \mid \mathcal{F}_{k-1}] \leq \frac{\varepsilon_q}{\alpha^2} \,. \tag{8.2.7}$$

as desired. □

Remark 8.2.5. In the direct search algorithm proposed in [14] the search direction at iteration k is chosen before the function estimates to be used in the acceptance test are computed. Thus our analysis can be extended also to that algorithm.

Remark 8.2.6. As a corollary of Proposition 8.2.4, our assumptions can always be satisfied if the variance of the oracle is finite (see Section 8.2.4 in for details). In Section 8.2.5 this is proved for finite moment oracles as well.

We now describe the relation between our assumptions and the β-probabilistically accurate function estimate assumption

$$\mathbb{P}(\{|F_k - f(x_k)| \leq \tau_f \Delta_k^2\} \cap \{|F_k^g - f(x_k + \Delta_k G_k)| \leq \tau_f \Delta_k^2|\} \mid \mathcal{F}_{k-1}) \geq \beta \,, \tag{8.2.8}$$

used in [14, 74, 198] in combination with other assumptions. In particular, conditions (8.2.3) are used in [14] and [198] (as discussed above), and a probabilistic assumption on the accuracy of random models for the objective is considered in [74].

We show that if (8.2.8) is satisfied for every β in a certain interval, with τ_f depending on an accuracy parameter ε, then also our assumptions are satisfied with $\varepsilon_f, \varepsilon_q$ dependent on ε. Note that the parameter τ_f is upper bounded by a function of β, arbitrarily large for β close to 1, but the result holds for any positive τ_f within the prescribed interval.

Proposition 8.2.7. *Let $\varepsilon > 0$ and $\bar{p} \in (0, 1)$. Assume that* (8.2.8) *holds for every $\beta \in [1 - \bar{p}, 1)$.*

- *If $\tau_f < \frac{\varepsilon}{2(1-\beta)}$, then Assumption 8.1 holds with $\varepsilon_f = \frac{\varepsilon}{\bar{p}}$ and $q = 2$.*
- *If $\tau_f < \frac{1}{2}\sqrt{\frac{\varepsilon}{1-\beta}}$, then Assumption 8.2 holds with $\varepsilon_q = \sqrt{\frac{\varepsilon}{\bar{p}}}$ and $p = q = 2$.*

Proof. First observe that by the triangular inequality

$$|F_k - f(x_k)| + |F_k^g - f(x_k + \Delta_k G_k)| \geq |F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \,.$$

Let $\alpha > \varepsilon_f$ be arbitrary. Then, for any $\tau_f < \frac{\alpha}{2}$,

$$\begin{aligned} &\{|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| < \alpha\Delta_k^2\} \\ &\supset \{|F_k - f(x_k)| \leq \tau_f\Delta_k^2\} \cap \{|F_k^g - f(x_k + \Delta_k G_k)| \leq \tau_f\Delta_k^2\}\,. \end{aligned} \tag{8.2.9}$$

Therefore, for $\beta = 1 - \frac{\varepsilon_f}{\alpha}\bar{p}$,

$$\begin{aligned} &\mathbb{P}(|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \geq \alpha\Delta_k^2 \; |\mathcal{F}_{k-1}) \\ &= (1 - \mathbb{P}(|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| < \alpha\Delta_k^2 \; |\mathcal{F}_{k-1})) \\ &\leq (1 - \mathbb{P}(\{|F_k - f(x_k)| \leq \tau_f(\beta)\Delta_k^2\} \cap \{|F_k^g - f(x_k + \Delta_k G_k)| \leq \tau_f(\beta)\Delta_k^2\} \mid \mathcal{F}_{k-1})) \\ &\leq 1 - \beta = \frac{\varepsilon_f}{\alpha}\bar{p} \leq \frac{\varepsilon_f}{\alpha}\,, \end{aligned}$$

where we were able to apply (8.2.9) in the first inequality since by assumption $\tau_f(\beta) < \frac{\varepsilon}{2(1-\beta)} = \frac{\alpha}{2}$, and the second inequality follows from (8.2.8). Given that $\alpha > \varepsilon_f$ is arbitrary, this proves the first point of the thesis, and an analogous reasoning holds for the second. □

We now show how the tail bound [162, Condition 2] is stronger than (a slight modification of) Assumption 8.1 for $q = 2$. We remark that in [162] this tail bound is combined with a probabilistically accurate difference estimate assumption and fully linear local model in order to prove convergence. We first recall the tail bound assumption [162, Condition 2]:

$$\mathbb{P}(F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k)) > (\beta\eta + \varepsilon)\min\{\Delta_k, \Delta_k^2\} \; |\mathcal{F}_{k-1}) \leq \frac{\theta}{\varepsilon}, \tag{8.2.10}$$

for every $\varepsilon > 0$, $k > \hat{k}$, and some $\beta, \eta, \theta > 0$. We now introduce the following modification of Assumption 8.1 for $q = 2$, essentially equivalent for our purposes:

$$\mathbb{P}(F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k)) > \alpha\Delta_k^2 \; |\mathcal{F}_{k-1}) \leq \frac{\varepsilon_f}{\alpha}\,, \tag{8.2.11}$$

for every $\alpha \geq \varepsilon_f$. It is straightforward to check that all of our results still hold if we replace (A1) with (8.2.11).

Proposition 8.2.8. *If* (8.2.10) *holds with*

$$\theta + \beta\eta < \varepsilon_f, \tag{8.2.12}$$

then (8.2.11) *holds.*

Proof. First, for every $\alpha \geq \varepsilon_f$ we have

$$\frac{\theta}{\alpha - \eta\beta} \leq \frac{\varepsilon_f}{\alpha} \tag{8.2.13}$$

under (8.2.12), since

$$\frac{\theta}{1 - \eta\beta/\alpha} \leq \frac{\theta}{1 - \eta\beta/\varepsilon_f} = \frac{\varepsilon_f \theta}{\varepsilon_f - \eta\beta} \leq \varepsilon_f \, ,$$

where we used $\alpha \geq \varepsilon_f$ in the first inequality and (8.2.12) in the last inequality.

Now, for every $\alpha \geq \varepsilon_f$:

$$\begin{aligned}
&\mathbb{P}(F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k)) > \alpha \Delta_k^2 \; |\mathcal{F}_{k-1}) \\
&= \mathbb{P}(F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k)) > (\eta\beta + (\alpha - \eta\beta))\Delta_k^2 \; |\mathcal{F}_{k-1}) \\
&\leq \mathbb{P}(F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k)) > (\eta\beta + (\alpha - \eta\beta)) \min\{\Delta_k, \Delta_k^2\} \; |\mathcal{F}_{k-1}) \\
&\leq \frac{\theta}{\alpha - \eta\beta} \leq \frac{\varepsilon_f}{\alpha},
\end{aligned} \tag{8.2.14}$$

where we used (8.2.10) with $\varepsilon = \alpha - n\beta$ in the first inequality and (8.2.13) in the last inequality. □

8.2.4 Finite variance oracle

A common assumption in stochastic derivative-free optimization is that the stochastic oracle is exact in expected value and with bounded variance [162, 215]:

$$\begin{aligned}
f(x) &= \mathbb{E}_\xi[F(x, \xi)] \, , \\
\mathrm{Var}_\xi[F(x, \xi)] &\leq V < +\infty \, .
\end{aligned} \tag{8.2.15}$$

In other words, the objective is assumed to be the expected value of a random variable $F(x, \xi)$ parametrized by x, with the black-box oracle given by sampling on the ξ space. The estimate F_k can then be computed by averaging on p_k i.i.d. samples $\{\xi_{k,i}\}_{i=1}^{p_k}$ of ε:

$$F_k = \frac{1}{p_k} \sum_{i=1}^{p_k} F(x_k, \xi_{k,i}) \, , \tag{8.2.16}$$

and analogously F_k^g can be computed by averaging on p_k^g random samples $\{\xi_{k,i}^g\}_{i=1}^{p_k^g}$.

Denoting with $\lceil \cdot \rceil$ the upper integer approximation, we have that $\lceil V/(k_f^2 \Delta_k^4) \rceil$ samples are enough to satisfy (8.2.3) and therefore in particular our conditions for $\varepsilon_f = 2k_f$, $\varepsilon_q = 4k_f^2$, and $p = q = 2$ thanks to Proposition 8.2.4. Indeed for $p_k \geq \lceil V/(k_f^2 \Delta_k^4) \rceil$ we have

$$\begin{aligned} \mathbb{E}[|F_k - f(x_k)|^2 \mid \mathcal{F}_{k-1}] =& \mathbb{E}\left[\left(\frac{1}{p_k}\sum_{i=1}^{p_k} F(x_k, \xi_{k,i}) - f(x_k)\right)^2 \mid \mathcal{F}_{k-1}\right] \\ =& \frac{1}{p_k}\mathbb{E}\left[\frac{1}{p_k}\sum_{i=1}^{p_k}(F(x_k, \xi_{k,i}) - f(x_k))^2 \mid \mathcal{F}_{k-1}\right] \\ =& \frac{1}{p_k}\mathrm{Var}[F(x_k, \xi)] \leq \frac{V}{p_k} \leq k_f^2 \Delta_k^4 , \end{aligned}$$

where we used the $\mathcal{F}_{k-1}$ measurability of p_k in the second equality, that $\{\xi_{k,i}\}_{i=1}^{p_k}$ are i.i.d. and also independent of $\mathcal{F}_{k-1}$ in the third equality, and the assumption (8.2.15). The inequality for F_k^g can be proved analogously when $p_k^g \geq \lceil V/(k_f^2 \Delta_k^4) \rceil$.

8.2.5 Finite moment oracle

We now describe the more general case where instead of having finite variance we have finite $r-$th moment for some $r > 1$:

$$\begin{aligned} f(x) &= \mathbb{E}_\xi[F(x, \xi)] , \\ \mathrm{E}_\xi[|F(x, \xi) - f(x)|^r] &\leq M_r < +\infty . \end{aligned} \tag{8.2.17}$$

Recall that finite $r-$th moment implies finite $r'-$th moment for any $r' \in (1, r]$. Thus for $r < 2$ assumption (8.2.17) is weaker than (8.2.15), while for $r > 2$ (8.2.17) is stronger than (8.2.15). The next result describes the number of samples needed asymptotically to satisfy our tail bound conditions as a function of r, q.

Theorem 8.2.9. *If $r \in (1, 2]$, then Assumptions 8.1 and 8.2 for $p = r$ can be satisfied with*

$$O\left(\Delta_k^{\min(-\frac{qr}{r-1}, -\frac{r+q}{r-1})}\right) \tag{8.2.18}$$

samples, while if $r \in [2 + \infty)$, they can be satisfied with

$$O\left(\Delta_k^{\min(-2q, -\frac{2(r+q)}{r})}\right) \tag{8.2.19}$$

samples.

We start with a lemma derived from classic results on the convergence rate for the law of large numbers from [229, 230].

Proof. Let $\bar{F}_k = F_k - f(x_k)$ and $\bar{F}_k^g = F_k^g - f(x_k + \alpha_k d_k)$, for F_k and F_k^g average of p_k samples as in Section 8.2.4.
We start with the case $r \in (1, 2]$. By the conditional version of [229, Theorem 2], we have

$$\mathbb{E}[|\bar{X}_k|^r \mid \mathcal{F}_{k-1}] \le 2M_r p_k^{1-r} \tag{8.2.20}$$

for $\bar{X}_k = \bar{F}_k, \bar{F}_k^g$. Let now $X_k = \bar{F}_k - \bar{F}_k^g$. We have

$$\mathbb{E}[|X_k|^r \mid \mathcal{F}_{k-1}] \le 2^{r-1}\mathbb{E}[|\bar{F}_k|^r + |\bar{F}_k^g|^r \mid \mathcal{F}_{k-1}] \le 2^r M_r p_k^{1-r}\,, \tag{8.2.21}$$

where we used $||a| + |b||^r \le 2^{r-1}(|a|^r + |b|^r)$ for $a, b \in \mathbb{R}$ in the first inequality, and (8.2.20) in the second. Now by Jensen's inequality

$$\mathbb{E}[|X_k| \mid \mathcal{F}_{k-1}] \le \sqrt[r]{\mathbb{E}[|X_k|^r \mid \mathcal{F}_{k-1}]} \le 2\sqrt[r]{M_r} p_k^{\frac{1-r}{r}}\,. \tag{8.2.22}$$

We can finally obtain our first tail bound:

$$\mathbb{P}(|X_k| \ge \alpha\Delta_k^q \mid \mathcal{F}_{k-1}) =\le \frac{\mathbb{E}[|X_k| \mid \mathcal{F}_{k-1}]}{\alpha\Delta_k^q} \le 2\sqrt[r]{M_r}\frac{p_k^{\frac{1-r}{r}}}{\alpha\Delta_k^q} \tag{8.2.23}$$

where we used the conditional Chebycheff inequality in the first inequality, and (8.2.22) in the second inequality. For $p_k = O(\Delta_k^{-\frac{qr}{r-1}})$ in particular the RHS of (8.2.23) is $O(1/\alpha)$, implying Assumption 8.1 as desired. As for Assumption 8.2, reasoning as for (8.2.23) and applying (8.2.21) we obtain

$$\begin{aligned}\mathbb{P}(|X_k| \ge \alpha\Delta_k^{1+\frac{q}{r}} \mid \mathcal{F}_{k-1}) &= \mathbb{P}(|X_k|^r \ge \alpha\Delta_k^{r+q} \mid \mathcal{F}_{k-1}) \\ \le \frac{\mathbb{E}[|X_k|^r \mid \mathcal{F}_{k-1}]}{\alpha^r\Delta_k^{r+q}} &\le 2^r M_r \frac{p_k^{1-r}}{\alpha^r\Delta_k^{r+q}}\,,\end{aligned} \tag{8.2.24}$$

where for $p_k = O(\Delta_k^{-\frac{q+r}{r-1}})$ the RHS of (8.2.24) is $O(1/\alpha^r)$ and Assumption 8.2 follows.
In the case $r \in (2, +\infty)$, by the conditional version of the first moment bound in [230, Section 5], we have

$$\mathbb{E}[|\bar{X}_k|^r \mid \mathcal{F}_{k-1}] \le K p_k^{-\frac{r}{2}} \tag{8.2.25}$$

for some constant K dependent from the distribution of the error, and for $\bar{X}_k = \bar{F}_k, \bar{F}_k^g$. Then reasoning as for the case $r \in (1, 2]$, we obtain, analogously to (8.2.23):

$$\begin{aligned}\mathbb{P}(|X_k| \ge \alpha\Delta_k^q \mid \mathcal{F}_{k-1}) &\le \frac{\mathbb{E}[|X_k| \mid \mathcal{F}_{k-1}]}{\alpha\Delta_k^q} \\ \le \frac{\sqrt[r]{\mathbb{E}[|X_k|^r \mid \mathcal{F}_{k-1}]}}{\alpha\Delta_k^q} &\le \frac{\sqrt[r]{K} p_k^{-\frac{1}{2}}}{\alpha\Delta_k^q}\,,\end{aligned} \tag{8.2.26}$$

so that in particular for $p_k = O(\Delta_k^{-2q})$ we retrieve Assumption 8.1. We then obtain, analogously to (8.2.24):

$$\begin{aligned} \mathbb{P}(|X_k| \geq \alpha\Delta_k^{1+\frac{q}{r}} \mid \mathcal{F}_{k-1}) &= \mathbb{P}(|X_k|^r \geq \alpha^r \Delta_k^{r+q} \mid \mathcal{F}_{k-1}) \\ \leq \frac{\mathbb{E}[|X_k|^r \mid \mathcal{F}_{k-1}]}{\alpha^r \Delta_k^{r+q}} &\leq \frac{K p_k^{-\frac{r}{2}}}{\alpha^r \Delta_k^{r+q}}\,. \end{aligned} \tag{8.2.27}$$

so that in particular for $p_k = O(\Delta_k^{\frac{-2(r+q)}{r}})$ we retrieve Assumption 8.2. The result then follows immediately taking the worst case of the bounds proved above for p_k. □

Remark 8.2.10. Let $\varepsilon > 0$. Applying (8.2.19) with $r_\varepsilon = \max(2, \frac{2q}{\varepsilon})$ and $q_\varepsilon = 1 + \frac{\varepsilon}{2}$ we can conclude that $O(\Delta_k^{-2-\varepsilon})$ samples are sufficient to satisfy assumptions 8.1 and 8.2 for $p = r_\varepsilon$ and $q = q_\varepsilon$, under the finite moment assumption (8.2.17) for $r = r_\varepsilon$.

8.3 Direct search for stochastic non-smooth functions

In this section, we first describe a simple stochastic direct search algorithm for the unconstrained minimization problem given in (8.1.1), where f is possibly non-smooth, and then analyze its convergence.

8.3.1 A simple stochastic direct search scheme

A detailed description of our stochastic direct search method is given in Algorithm 18. At each iteration, we generate a direction g_k in the unitary sphere (independently of the estimates of the objective function generated so far; see Step 3), and perform a step along the direction g_k with stepsize δ_k. Then, at Step 4, we compute the estimate values f_k^g and f_k of the function at the resulting trial point $x_k + \delta_k g_k$ and also at x_k. We then accept or reject the trial point based on a sufficient decrease condition, imposing that the improvement on the objective estimate at the trial point is at least $\theta\delta_k^q$. If the sufficient decrease condition is satisfied, we have a successful iteration. We hence update our iterate x_{k+1} by setting it equal to the trial point and expand or keep the same stepsize at Step 5. Otherwise, the iteration is unsuccessful, so we do not move (i.e., $x_{k+1} = x_k$) and shrink the stepsize (see Step 6).

Algorithm 18 `Stochastic direct search`

```
Initialization. Choose a point $x_0$, $\Delta_0, \theta > 0$, $\tau \in (0,1)$, $\bar{\tau} \in [1, 1+\tau]$, $q > 1$.
For $k = 0, 1 \dots$
    Select a direction $g_k$ in the unitary sphere.
    Compute estimates $f_k$ and $f_k^g$ for $f$ in $x_k$ and $x_k + \delta_k g_k$.
    If $f_k - f_k^g \geq \theta \delta_k^q$, Then set SUCCESS = true, $x_{k+1} = x_k + \delta_k g_k$, $\Delta_{k+1} = \bar{\tau}\delta_k$.
    Else set SUCCESS = false, $x_{k+1} = x_k$, $\Delta_{k+1} = (1-\tau)\delta_k$.
    End if
End for
```

In order for the method to convergence to Clarke stationary points, the sequence $\{g_k\}$ must be dense in the unit sphere on certain subsequences (see Theorem 8.3.3). We remark that a dense sequence on the unit sphere can be generated using a suitable quasirandom sequence (see, e.g., [121, 172]).

8.3.2 Convergence analysis under the tail-bound probabilistic condition

The following theorem, which implies that the stepsize sequence $\{\Delta_k\}$ converges to zero almost surely, is a key result in the convergence analysis. By taking a look at the proof, we can see how the use of the tail-bound probabilistic condition (A1) allows us to give a unified argument for unsuccessful and successful steps.

We define now for convenience the positive constants $\tau_q^+ = (1+\tau)^q - 1$, $\tau_q^- = 1 - (1-\tau)^q$, and $\tau_q^{(\Delta)} = \tau_q^+ + \tau_q^-$. To obtain our result we need the following lower bound on the parameter θ defining the sufficient decrease condition, dependent on the stepsize update parameter τ and the tail bound parameter ε_f:

$$\theta > \frac{\varepsilon_f \tau_q^{(\Delta)}}{\tau_q^-}. \tag{8.3.1}$$

Notice that since $\tau \in (0,1)$ we must always have $\theta > 0$. The bound (8.3.1) allows us to relate stepsize expansions to improvements of the objective.

Theorem 8.3.1. *Under Assumption 8.1, if* (8.3.1) *holds then*

$$\sum_{k \in \mathbb{N}_0} \mathbb{E}[\Delta_k^q] < \infty \tag{8.3.2}$$

a.s. in $\mathbf{\Omega}$.

Proof. Let $\Phi_k = f(x_k) - f^* + \eta\Delta_k^q$, with $\eta = \frac{\theta}{\tau_q^{(\Delta)}}$, and

$$\varepsilon = -\varepsilon_f + \frac{\tau_q^- \theta}{\tau_q^{(\Delta)}} > 0\,,$$

where the inequality follows by (8.3.1).

We will prove, for every $k \geq 0$, that

$$\mathbb{E}[\Phi_k - \Phi_{k+1} \mid \mathcal{F}_{k-1}] \geq \varepsilon\Delta_k^q\,. \tag{8.3.3}$$

The thesis then follows as in [14, Theorem 1] (or directly by Robbins-Siegmund theorem [210]).

It remains to prove (8.3.3). Let ρ_k be the random variable such that $f(x_k) - f(x_k + \Delta_k G_k) = (\theta - \rho_k)\Delta_k^q$, and let J_k be the event that the step k is successful. We have

$$\begin{aligned}
&\mathbb{E}[(\Phi_k - \Phi_{k+1})|\mathcal{F}_{k-1}] = \mathbb{E}[(\Phi_k - \Phi_{k+1})(\mathbb{1}_{J_k} \;\; + (1 - \mathbb{1}_{J_k}))|\mathcal{F}_{k-1}] \\
&= (f(x_k) - f(x_{k+1}) + \eta(\Delta_k^q - \Delta_{k+1}^q))\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&\quad + (f(x_k) - f(x_{k+1}) + \eta(\Delta_k^q - \Delta_{k+1}^q))\mathbb{E}[\; 1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&= (f(x_k) - f(x_k + \Delta_k G_k) + \eta(\Delta_k^q - \Delta_{k+1}^q))\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&\quad + \eta(\Delta_k^q - \Delta_{k+1}^q)\mathbb{E}[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&\geq (((\theta - \rho_k) - \eta\tau_q^+)\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta\tau_q^-\mathbb{E}[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}])\Delta_k^q,
\end{aligned} \tag{8.3.4}$$

where we used $x_k = x_{k+1}$ for unsuccessful steps in the second equality, and $\Delta_{k+1} = \bar{\tau}\Delta_k \leq (1+\tau)\Delta_k$ for successful steps in the inequality. In turn,

$$\begin{aligned}
&(((\theta - \rho_k) - \eta\tau_q^+)\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta\tau_q^-\mathbb{E}[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}])\Delta_k^q \\
&= ((\theta - \rho_k - \eta\tau_q^{(\Delta)})\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta\tau_q^-)\Delta_k^q \\
&= -\rho_k\Delta_k^q\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta\tau_q^-\Delta_k^q\,,
\end{aligned} \tag{8.3.5}$$

where we used $\mathbb{E}[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}] = 1 - \mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}]$ in the first equality, and $\theta = \eta\tau_q^{(\Delta)}$ in the second one. By combining (8.3.4) and (8.3.5) we can therefore conclude

$$\mathbb{E}[(\Phi_k - \Phi_{k+1})|\mathcal{F}_{k-1}] \geq -\rho_k\Delta_k^q\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta\tau_q^-\Delta_k^q\,. \tag{8.3.6}$$

Notice that if the step is successful then $f_k - f_k^g \geq \theta\Delta_k^q$, which implies

$$f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k G_k)) \geq \theta\Delta_k^q - (\theta - \rho_k)\Delta_k^q = \rho_k\Delta_k^q\,.$$

In particular

$$J_k \subset \{|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \geq \rho_k \Delta_k^q\},$$

and we can write

$$\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] = \mathbb{P}(J_k|\mathcal{F}_{k-1}) \leq \mathbb{P}(|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \geq \rho_k \Delta_k^q|\mathcal{F}_{k-1}). \tag{8.3.7}$$

We now have

$$-\rho_k \mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \geq -\rho_k \mathbb{P}(|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \geq \rho_k \Delta_k^q|\mathcal{F}_{k-1}) \geq -\varepsilon_f, \tag{8.3.8}$$

where we applied (8.3.7) in the first inequality, the last inequality is a direct consequence of (A1) for $\alpha = \rho_k$. Hence,

$$-\rho_k \Delta_k^q \mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta \tau_q^- \Delta_k^q \geq (-\varepsilon_f + \eta \tau_q^-)\Delta_k^q = \varepsilon \Delta_k^q, \tag{8.3.9}$$

where we used (8.3.8) in the inequality.

Claim (8.3.3) can finally be obtained by concatenating (8.3.6) and (8.3.9). □

The lemma we now state will be useful for the proof of the optimality result of Theorem 8.3.3 which is based on the Clarke generalized directional derivative. We notice that Assumption 8.2 plays a key role in this result, allowing us to upper bound the error of the reduction estimate by a quantity that depends on the stepsize Δ_k.

Lemma 8.3.2. *Let K be the set of indices of unsuccessful iterations (notice that K is random). Then under Assumptions 8.1–8.2 and* (8.3.1) *we have a.s. in* Ω

$$\liminf_{k \in K,\, k \to \infty} \frac{f(x_k + \Delta_k G_k) - f(x_k)}{\Delta_k} \geq 0. \tag{8.3.10}$$

Proof. Clearly it suffices to show that, for any given $m \in \mathbb{N}$ and a.s.,

$$\liminf_{k \in K,\, k \to \infty} \frac{f(x_k + \Delta_k G_k) - f(x_k)}{\Delta_k} \geq -\frac{1}{m}. \tag{8.3.11}$$

To start with, by applying (A2) with $\alpha = \frac{\Delta_k^{-\frac{q}{p}}}{m}$ we have

$$\mathbb{P}(|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \geq \frac{\Delta_k}{m} \mid \mathcal{F}_{k-1}) \leq m^p \Delta_k^q \varepsilon_q.$$

and therefore taking expectations on both sides

$$\mathbb{P}(|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \geq \frac{\Delta_k}{m}) \leq m^p \mathbb{E}[\Delta_k^q] \varepsilon_q \,. \tag{8.3.12}$$

We can now deduce

$$\sum_{k \in \mathbb{N}_0} \mathbb{P}(|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \geq \frac{\Delta_k}{m}) \leq \sum_{k \in \mathbb{N}_0} m^p \mathbb{E}[\Delta_k^q] \varepsilon_q < \infty \,, \tag{8.3.13}$$

where we applied Theorem 8.3.1 in the last inequality. In particular, by the Borel-Cantelli's First Lemma

$$\mathbb{P}\left(\left\{|F_k - f_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \geq \frac{\Delta_k}{m}\right\} \text{ i.o.}\right) = 0 \,,$$

where "i.o." stands for *infinitely often.* Hence, we have a.s.

$$|F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))| \leq \frac{\Delta_k}{m} \quad \text{for } k \text{ large enough.} \tag{8.3.14}$$

From this we can infer that a.s., for every $k \in K$ large enough

$$\begin{aligned} \frac{f(x_k + \Delta_k G_k) - f(x_k)}{\Delta_k} &\geq \frac{F_k^g - F_k - |F_k - F_k^g - (f(x_k) - f(x_k + \Delta_k G_k))|}{\Delta_k} \\ &\geq -\theta \Delta_k - \frac{1}{m} \,, \end{aligned} \tag{8.3.15}$$

where we used (8.3.14) combined with the unsuccessful step condition of Algorithm 18 in the second inequality. Finally, (8.3.11) follows passing to the liminf for $k \to \infty$ in (8.3.15). □

We now report the main convergence result for our stochastic direct search scheme.

Theorem 8.3.3. *Assume that f is Lipschitz continuous with constant L_f^* around any limit point of the sequence of iterates $\{x_k\}$. Let K be the set of indices of unsuccessful iterations. Under Assumptions 8.1–8.2, the following property holds a.s. in Ω: if $L \subset K$ (notice that L, K are random) is such that $\{G_k\}_{k \in L}$ is dense in the unit sphere and*

$$\lim_{k \in L,\, k \to \infty} x_k = x^* \,,$$

then the point x^* is Clarke stationary.

Proof. Let d be a direction in the unitary sphere, and for $w \in \Omega$ let $S(w) \subset L(w)$ be such that

$$\lim_{k \in S(w),\, k \to \infty} G_k = d\,.$$

By definition of Clarke stationarity, we just need to prove that a.s. (for an event independent of d)

$$\limsup_{k \in S(w),\, k \to \infty} \frac{f(x_k + \Delta_k d) - f(x_k)}{\Delta_k} \geq 0\,.$$

Then on V' we can write

$$\limsup_{k \in S(w),\, k \to \infty} \frac{f(x_k + \Delta_k G_k) - f(x_k)}{\Delta_k} \geq \liminf_{k \in K(w),\, k \to \infty} \frac{f(x_k + \Delta_k G_k) - f(x_k)}{\Delta_k} \geq 0, \tag{8.3.16}$$

where the last inequality follows by (8.3.10).

Now using the Lipschitz property of f we can write, for $k \in S(w)$ large enough,

$$\begin{aligned} \frac{f(x_k + \Delta_k d) - f(x_k)}{\Delta_k} &= \frac{f(x_k + \Delta_k G_k) - f(x_k)}{\Delta_k} + \frac{f(x_k + \Delta_k d) - f(x_k + \Delta_k G_k)}{\Delta_k} \\ &\geq \frac{f(x_k + \Delta_k G_k) - f(x_k)}{\Delta_k} - L_f^* \|G_k - d\|. \end{aligned}$$

Passing to the limsup for $k \in S(w) \subset L(w)$ we get

$$\limsup_{k \in S(w),\ k \to \infty} \frac{f(x_k + \Delta_k d) - f(x_k)}{\Delta_k} \geq \limsup_{k \in S(w),\ k \to \infty} \frac{f(x_k + \Delta_k G_k) - f(x_k)}{\Delta_k} \geq 0\,,$$

for every $w \in V'$, where we used $\|G_k - d\| \to 0$ by construction in the first inequality and (8.3.16) in the second. □

Chapter 9

Conclusion

In this thesis, several convergence results were given for first order projection free and direct search methods. A recurring theme was the use of relatively "cheap" local search directions relaxing some properties of the (projected) negative gradient and still achieving comparable convergence results. For instance, in Chapter 3 a framework to obtain linear convergence for constrained smooth optimization problems using directions satisfying the angle condition (3.3.2) was employed to improve several FW variants. In Chapters 4 and 5 it was proved that using the FW direction combined with away or in face steps one can obtain local identification properties analogous to those of the projected gradient method. In Chapter 7 qualitative convergence results on Riemannian manifolds were given for methods applying retractions to tentative descent directions chosen from a poll set with positive cosine measure on the tangent space of the current iterate. Finally, the direct search method in Chapter 8 allows the directions to be taken uniformly at random in the unit sphere, while still showing convergence properties to Clarke stationary points for stochastic objectives.

We now discuss some possible future works. First, concerning the framework introduced in Chapter 3 for projection free optimization, a possible extension consists in its adaptation to problems on product domains, i.e. of the form

$$\min_{x \in \Omega^{(1)} \times \ldots \times \Omega^{(m)}} f(x) \,. \tag{9.0.1}$$

As explained in Section 2.8, a block coordinate version of the classic FW method for problem (9.0.1) was given in [158]. With respect to that work, an adaptation of our framework to problem (9.0.1) would also cover FW variants, and different block selection strategies from the randomized one like parallel and Gauss-Southwell block selection (see, e.g., [195]). The main idea here is that by applying the SSC separately to each of the blocks one can retrieve descent properties analogous to those proved in the single block case in Chapter 3. For instance, for the parallel update

$$x_{k+1}^{(i)} = \mathrm{SSC}(x_k^{(i)}, -\nabla f(x_k)^{(i)}) \quad \text{for } i \in [1:m] \tag{9.0.2}$$

it is possible to prove a property analogous to (3.4.7), that is

$$\|x_{k+1} - x_k\|^2 \geq \frac{\tau^2}{2(1+\tau^2)L^2} \|\pi(T_\Omega(\tilde{x}_k), -\nabla f(\tilde{x}_k))\|^2 , \tag{9.0.3}$$

for a suitably chosen $\tilde{x}_k$.

For direct search methods, a possible future work consists in combining the method analyzed in Chapter 8 with the nonmonotone linesearch technique. This technique, introduced in [114] for Newton's method and analyzed in [248] for gradient descent, consists in considering an upper bound E_k on $f(x_k)$ instead of $f(x_k)$ itself in sufficient decrease conditions, thus enabling more aggressive exploration strategies. It has been extended to direct search methods in the recent work [173], but only for deterministic objectives. A promising approach for the stochastic case appears to be considering E_k as an exponential moving average of the past function estimates, adapting an idea introduced in [248] for deterministic gradient descent. One important obstacle is that in the stochastic case we cannot ensure that E_k is an upper bound on $f(x_k)$. A possible solution is to prove instead

$$\mathbb{P}(E_k - f(x_k) \leq -\alpha\Delta_k^2 \mid \mathcal{F}_{k-1}) \leq \frac{\varepsilon_E}{\alpha^2} \tag{9.0.4}$$

for every $\alpha > 0$ and some $\varepsilon_E > 0$, that is a tail bound condition along the lines of those introduced in Chapter 8.

Another possible development concerning the tail bound conditions 8.1 and 8.2 is their extensions to model based derivative free optimization methods. A suitable setting for such an extension appears to be the trust region method proposed in [172]. The sufficient decrease condition $f(x_k) - f(x_k + s_k) \geq \eta_1 \theta \|s_k\|^q$ used in [172], with s_k solution of the trust region subproblem and $\eta_1 > 0$ constant, can be easily extended to the stochastic setting using function estimates f_k and f_k^s in place of exact function values. Then 8.1 and 8.2 can be adapted by using $\|s_k\|$ and $\hat{s}_k$ in place of Δ_k and g_k respectively. However, it is still unclear to the authors if these conditions can be extended, beside to the function estimates used in the sufficient decrease condition, to the trust region model itself.

Other future works include the extension of the methods studied in Chapters 3, 4 and 7 to the stochastic case, as well as more numerical tests on real world data science problems.

Bibliography

[1] Mark A Abramson, Charles Audet, G Couture, John E Dennis Jr, Sébastien Le Digabel, and C Tribes. The nomad project. *Software available at http://www. gerad. ca/nomad*, 115, 2011.

[2] P-A Absil, Robert Mahony, and Benjamin Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.

[3] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, 2009.

[4] P-A Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

[5] Selin Damla Ahipaşaoğlu, Peng Sun, and Michael J Todd. Linear convergence of a modified Frank–Wolfe algorithm for computing minimum-volume enclosing ellipsoids. *Optimisation Methods and Software*, 23(1):5–19, 2008.

[6] Selin Damla Ahipaşaoğlu and Michael J Todd. A modified Frank–Wolfe algorithm for computing minimum-area enclosing ellipsoidal cylinders: Theory and algorithms. *Computational Geometry*, 46(5):494–519, 2013.

[7] Martin Aigner, Günter M Ziegler, Karl H Hofmann, and Paul Erdős. *Proofs from the Book*, volume 274. Springer, 2010.

[8] Pedro Alberto, Fernando Nogueira, Humberto Rocha, and Luís N Vicente. Pattern search methods for user-provided points: Application to molecular geometry problems. *SIAM Journal on Optimization*, 14(4):1216–1236, 2004.

[9] Ralph Alexander. The width and diameter of a simplex. *Geometriae Dedicata*, 6(1):87–94, 1977.

[10] Zeyuan Allen-Zhu, Elad Hazan, Wei Hu, and Yuanzhi Li. Linear convergence of a Frank-Wolfe type algorithm over trace-norm balls. *Advances in Neural Information Processing Systems*, 2017:6192–6201, 2017.

[11] S. Amaran, N. V. Sahinidis, B. Sharda, and S. J. Bury. Simulation optimization: a review of algorithms and applications. *Ann. Oper. Res.*, 240:351–380, 2016.

[12] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[13] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[14] C. Audet, K. J. Dzahini, M. Kokkolaras, and S. Le Digabel. Stochastic mesh adaptive direct search for blackbox optimization using probabilistic estimates. 79:1–34, 2021.

[15] C. Audet and W. Hare. *Derivative-Free and Blackbox Optimization*, volume 2 of *Springer Series in Operations Research and Financial Engineering*. Springer, Cham, Switzerland, 2017.

[16] Charles Audet. A survey on direct search methods for blackbox optimization and their applications. *Mathematics without boundaries*, pages 31–56, 2014.

[17] Charles Audet and John E Dennis Jr. Analysis of generalized pattern searches. *SIAM J. Optim.*, 13(3):889–903, 2002.

[18] Charles Audet and John E Dennis Jr. Mesh adaptive direct search algorithms for constrained optimization. *SIAM J. Optim.*, 17(1):188–217, 2006.

[19] Charles Audet and Dominique Orban. Finding optimal algorithmic parameters using derivative-free optimization. *SIAM Journal on Optimization*, 17(3):642–664, 2006.

[20] Daniel Azagra, Juan Ferrera, and Fernando López-Mesas. Nonsmooth analysis and hamilton–jacobi equations on riemannian manifolds. *J. Funct. Anal.*, 220(2):304–361, 2005.

[21] Francis Bach et al. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends® in Machine Learning*, 6(2-3):145–373, 2013.

[22] Maxim Balashov, Boris Polyak, and Andrey Tremba. Gradient projection and conditional gradient methods for constrained nonconvex minimization. *arXiv preprint arXiv:1906.11580*, 41(7):822–849, 2019.

[23] A. S. Bandeira, K. Scheinberg, and L. N. Vicente. Convergence of trust-region methods based on probabilistic models. 24:1238–1264, 2014.

[24] Mohammad Ali Bashiri and Xinhua Zhang. Decomposition-invariant conditional gradient for general polytopes with line search. In *Advances in Neural Information Processing Systems*, pages 2690–2700, 2017.

[25] Leonard E Baum and George Sell. Growth transformations for functions on manifolds. *Pacific Journal of Mathematics*, 27(2):211–227, 1968.

[26] Amir Beck, Edouard Pauwels, and Shoham Sabach. The cyclic block conditional gradient method for convex optimization problems. *SIAM Journal on Optimization*, 25(4):2024–2049, 2015.

[27] Amir Beck and Shimrit Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 164(1-2):1–27, 2017.

[28] Leonard Berrada, Andrew Zisserman, and M Pawan Kumar. Deep Frank-Wolfe for neural network optimization. In *International Conference on Learning Representations*, 2018.

[29] Dimitri P Bertsekas. Projected newton methods for optimization problems with simple constraints. *SIAM J. Control Optim.*, 20(2):221–246, 1982.

[30] Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

[31] Dimitri P Bertsekas. *Convex optimization algorithms.* Athena Scientific, Nashua, 2015.

[32] Mathieu Besançon, Alejandro Carderera, and Sebastian Pokutta. Frankwolfe. jl: A high-performance and flexible toolbox for frank–wolfe algorithms and conditional gradients. *INFORMS Journal on Computing*, 2022.

[33] Ernesto G Birgin and José Mario Martínez. Large-scale active-set box-constrained optimization method with spectral projected gradients. *Comput. Optim. Appl.*, 23(1):101–125, 2002.

[34] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg. Convergence rate analysis of a stochastic trust-region method via supermartingales. *INFORMS J. Optim.*, 1:92–119, 2019.

[35] Jacek Bochnak, Michel Coste, and Marie-Françoise Roy. *Real algebraic geometry*, volume 36. Springer Science & Business Media, 2013.

[36] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

[37] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

[38] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

[39] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

[40] Immanuel M Bomze. Evolution towards the maximum clique. *Journal of Global Optimization*, 10(2):143–164, 1997.

[41] Immanuel M Bomze. On standard quadratic optimization problems. *J. Global Optim.*, 13(4):369–387, 1998.

[42] Immanuel M. Bomze. Regularity versus degeneracy in dynamics, games, and optimization: a unified approach to different aspects. *SIAM Review*, 44(3):394–414, 2002.

[43] Immanuel M Bomze, Marco Budinich, Panos M Pardalos, and Marcello Pelillo. The maximum clique problem. In *Handbook of Combinatorial Optimization*, pages 1–74. Springer, 1999.

[44] Immanuel M. Bomze and Etienne de Klerk. Solving standard quadratic optimization problems via linear, semidefinite and copositive programming. *Journal of Global Optimization*, 24(2):163–185, 2002.

[45] Immanuel M Bomze, Mirjam Dür, Etienne De Klerk, Cornelis Roos, Arie J Quist, and Tamás Terlaky. On copositive programming and standard quadratic optimization problems. *J. Global Optim.*, 18(4):301–320, 2000.

[46] Immanuel M Bomze, Francesco Rinaldi, and Samuel Rota Bulò. First-order methods for the impatient: Support identification in finite time with convergent Frank-Wolfe variants. *SIAM Journal on Optimization*, 29(3):2211–2226, 2019.

[47] Immanuel M Bomze, Francesco Rinaldi, and Damiano Zeffiro. Active set complexity of the away-step Frank–Wolfe algorithm. *SIAM Journal on Optimization*, 30(3):2470–2500, 2020.

[48] Immanuel M Bomze, Francesco Rinaldi, and Damiano Zeffiro. Frank–wolfe and friends: a journey into projection-free first-order optimization methods. *4OR*, 19(3):313–345, 2021.

[49] Immanuel M Bomze, Francesco Rinaldi, and Damiano Zeffiro. Fast cluster detection in networks by first order optimization. *SIAM Journal on Mathematics of Data Science*, 4(1):285–305, 2022.

[50] Immanuel M Bomze and Volker Stix. Genetic engineering via negative fitness: Evolutionary dynamics for global optimization. *Ann. Oper. Res.*, 89:297–318, 1999.

[51] Andrew J Booker, JE Dennis, Paul D Frank, David B Serafini, and Virginia Torczon. Optimization using surrogate objectives on a helicopter test example. In *Computational Methods for Optimal Design and Control*, pages 49–58. Springer, 1998.

[52] V. S. Borkar. *Probability Theory: An Advanced Course*. Springer Science & Business Media, New York, 2012.

[53] Fani Boukouvala, Ruth Misener, and Christodoulos A Floudas. Global optimization advances in mixed-integer nonlinear programming, minlp, and constrained derivative-free optimization, cdfo. *European Journal of Operational Research*, 252(3):701–727, 2016.

[54] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(42):1455–1459, 2014.

[55] Nicolas Boumal. An introduction to optimization on smooth manifolds, 2022.

[56] Nicolas Boumal, Pierre-Antoine Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA J. Numer. Anal.*, 39(1):1–33, 2019.

[57] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[58] Gábor Braun, Sebastian Pokutta, Dan Tu, and Stephen Wright. Blended conditonal gradients. In *International Conference on Machine Learning*, pages 735–743. PMLR, 2019.

[59] Gábor Braun, Sebastian Pokutta, and Daniel Zink. Lazifying conditional gradient algorithms. In *ICML*, pages 566–575, 2017.

[60] James V Burke and Jorge J Moré. On the identification of active constraints. *SIAM Journal on Numerical Analysis*, 25(5):1197–1211, 1988.

[61] James V Burke and Jorge J Moré. Exposing constraints. *SIAM Journal on Optimization*, 4(3):573–595, 1994.

[62] Jim Burke. On the identification of active constraints II: The nonconvex case. *SIAM Journal on Numerical Analysis*, 27(4):1081–1102, 1990.

[63] Elcin Aleixo Calado, Marco Leite, and Arlindo Silva. Selecting composite materials considering cost and environmental impact in the early phases of aircraft structure design. *Journal of Cleaner Production*, 186:113–122, 2018.

[64] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[65] Michael D Canon and Clifton D Cullum. A tight upper bound on the rate of convergence of Frank-Wolfe algorithm. *SIAM Journal on Control*, 6(4):509–516, 1968.

[66] Alejandro Carderera and Sebastian Pokutta. Second-order conditional gradient sliding. *arXiv preprint arXiv:2002.08907*, 2020.

[67] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[68] Coralia Cartis and Katya Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, 169(2):337–375, 2018.

[69] Deeparnab Chakrabarty, Prateek Jain, and Pravesh Kothari. Provable submodular minimization using Wolfe's algorithm. *Advances in Neural Information Processing Systems*, 27:802–809, 2014.

[70] Ibrahim M Chamseddine, Hermann B Frieboes, and Michael Kokkolaras. Design optimization of tumor vasculature-bound nanoparticles. *Scientific Reports*, 8(1):1–15, 2018.

[71] Ibrahim M Chamseddine and Michael Kokkolaras. A dual nanoparticle delivery strategy for enhancing drug distribution in cancerous tissue. *Journal of Biomechanical Engineering*, 142(12), 2020.

[72] Jinghui Chen, Dongruo Zhou, Jinfeng Yi, and Quanquan Gu. A Frank-Wolfe framework for efficient and effective adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3486–3494, 2020.

[73] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017.

[74] R. Chen, M. Menickelly, and K. Scheinberg. Stochastic optimization using a trust-region method and random models. 169(2):447–487, 2018.

[75] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[76] Xiaoyu Chen, Yi Zhou, Jin-Kao Hao, and Mingyu Xiao. Computing maximum k-defective cliques in massive graphs. *Comput. Oper. Res.*, 127:105131, 2021.

[77] F. H. Clarke. *Optimization and Nonsmooth Analysis.* John Wiley & Sons, New York, 1983. Reissued by SIAM, Philadelphia, 1990.

[78] Kenneth L Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Transactions on Algorithms*, 6(4):1–30, 2010.

[79] Cyrille Combettes and Sebastian Pokutta. Boosting frank-wolfe by chasing gradients. In *International Conference on Machine Learning*, pages 2111–2121. PMLR, 2020.

[80] Cyrille W Combettes and Sebastian Pokutta. Complexity of linear minimization and projection on some sets. *arXiv preprint arXiv:2101.10040*, 2021.

[81] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization.* MPS-SIAM Series on Optimization. SIAM, Philadelphia, 2009.

[82] Rixon Crane and Fred Roosta. Invexifying regularization of non-linear least-squares problems. *arXiv preprint arXiv:2111.11027*, 2021.

[83] Andrea Cristofari, Marianna De Santis, Stefano Lucidi, and Francesco Rinaldi. An active-set algorithmic framework for non-convex optimization problems over the simplex. *arXiv preprint arXiv:1703.07761v2*, 2018.

[84] Andrea Cristofari, Marianna De Santis, Stefano Lucidi, and Francesco Rinaldi. An active-set algorithmic framework for non-convex optimization problems over the simplex. *Computational Optimization and Applications*, 77:57–89, 2020.

[85] Andrea Cristofari and Francesco Rinaldi. A derivative-free method for structured optimization problems. *SIAM J. Optim.*, 31(2):1079–1107, 2021.

[86] Ana Luísa Custódio, JF Aguilar Madeira, A Ismael F Vaz, and Luís Nunes Vicente. Direct multisearch for multiobjective optimization. *SIAM Journal on Optimization*, 21(3):1109–1140, 2011.

[87] Dragiša Cvetković and Peter Rowlinson. The largest eigenvalue of a graph: A survey. *Linear Multilinear Algebra*, 28(1-2):3–33, 1990.

[88] Marianna De Santis, Gianni Di Pillo, and Stefano Lucidi. An active set feasible method for large-scale minimization problems with bound constraints. *Computational Optimization and Applications*, 53(2):395–423, 2012.

[89] Vladimir Fedorovich Demyanov and Aleksandr Moiseevich Rubinov. *Approximate methods in optimization problems*, volume 32. American Elsevier, 1970.

[90] Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.

[91] Lijun Ding, Yingjie Fei, Qiantong Xu, and Chengrun Yang. Spectral Frank-Wolfe algorithm: Strict complementarity and linear convergence. In *International Conference on Machine Learning*, pages 2535–2544. PMLR, 2020.

[92] David W Dreisigmeyer. Equality constraints, riemannian manifolds and direct search methods. *https://optimization-online.org/?p=9135/*, 2006.

[93] David W Dreisigmeyer. Direct search methods on reductive homogeneous spaces. *J. Optim. Theory Appl.*, 176(3):585–604, 2018.

[94] Joseph C Dunn. Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. *SIAM Journal on Control and Optimization*, 17(2):187–211, 1979.

[95] Joseph C Dunn and S Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444, 1978.

[96] K. J. Dzahini. Expected complexity analysis of stochastic direct-search. 81:179–200, 2022.

[97] Kwassi Joseph Dzahini, Michael Kokkolaras, and Sébastien Le Digabel. Constrained stochastic blackbox optimization using a progressive barrier and probabilistic estimates. *Mathematical Programming*, pages 1–58, 2022.

[98] Giovanni Fasano, Giampaolo Liuzzi, Stefano Lucidi, and Francesco Rinaldi. A linesearch-based derivative-free approach for nonsmooth constrained optimization. *SIAM J. Optim.*, 24(3):959–992, 2014.

[99] E Fermi and N Metropolis. Los alamos unclassified report ls–1492. *Rapport technique, Los Alamos National Laboratory, Los Alamos, New Mexico*, 1952.

[100] OP Ferreira and WS Sosa. On the Frank–Wolfe algorithm for non-compact constrained optimization problems. *Optimization*, pages 1–15, 2021.

[101] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

[102] Robert M Freund and Paul Grigas. New analysis and results for the Frank–Wolfe method. *Mathematical Programming*, 155(1-2):199–230, 2016.

[103] Robert M Freund, Paul Grigas, and Rahul Mazumder. An extended Frank–Wolfe method with in-face directions, and its application to low-rank matrix completion. *SIAM Journal on Optimization*, 27(1):319–346, 2017.

[104] Satoru Fujishige. Lexicographically optimal base of a polymatroid with respect to a weight vector. *Mathematics of Operations Research*, 5(2):186–196, 1980.

[105] Masao Fukushima. A modified Frank-Wolfe algorithm for solving the traffic assignment problem. *Transportation Research Part B: Methodological*, 18(2):169–177, 1984.

[106] Dan Garber. Linear convergence of Frank-Wolfe for rank-one matrix recovery without strong convexity. *arXiv preprint arXiv:1912.01467*, 2019.

[107] Dan Garber. Revisiting frank-wolfe for polytopes: Strict complementarity and sparsity. *Advances in Neural Information Processing Systems*, 33:18883–18893, 2020.

[108] Dan Garber and Elad Hazan. Faster rates for the Frank-Wolfe method over strongly-convex sets. In *ICML*, volume 15 of *ICML'15*, pages 541–549. JMLR.org, 2015.

[109] Dan Garber and Elad Hazan. A linearly convergent variant of the conditional gradient algorithm under strong convexity, with applications to online and stochastic optimization. *SIAM Journal on Optimization*, 26(3):1493–1528, 2016.

[110] Dan Garber and Ofer Meshi. Linear-memory and decomposition-invariant linearly convergent conditional gradient algorithm for structured polytopes. *Advances in neural information processing systems*, 29, 2016.

[111] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[112] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[113] Serge Gratton, Clément W Royer, Luís Nunes Vicente, and Zaikun Zhang. Direct search based on probabilistic descent. *SIAM J. Optim.*, 25(3):1515–1541, 2015.

[114] Luigi Grippo, Francesco Lampariello, and Stephano Lucidi. A nonmonotone line search technique for newton's method. *SIAM Journal on Numerical Analysis*, 23(4):707–716, 1986.

[115] Peter Gritzmann and Marek Lassak. Estimates for the minimal width of polytopes inscribed in convex bodies. *Discrete & Computational Geometry*, 4(6):627–635, 1989.

[116] Jacques Guelat and Patrice Marcotte. Some comments on Wolfe's away step. *Mathematical Programming*, 35(1):110–119, 1986.

[117] David H Gutman and Javier F Pena. The condition number of a function relative to a set. *Mathematical Programming*, pages 1–40, 2020.

[118] William W Hager, Dzung T Phan, and Hongchao Zhang. Gradient-based methods for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):146–165, 2011.

[119] William W Hager and Hongchao Zhang. A new active set algorithm for box constrained optimization. *SIAM J. Optim.*, 17(2):526–557, 2006.

[120] William W Hager and Hongchao Zhang. An active set algorithm for nonlinear optimization with polyhedral constraints. *Science China Mathematics*, 59(8):1525–1542, 2016.

[121] J. H. Halton. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numer. Math.*, 2:84–90, 1960.

[122] Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1):75–112, 2015.

[123] Warren L Hare and Adrian S Lewis. Identifying active constraints via partial smoothness and prox-regularity. *Journal of Convex Analysis*, 11(2):251–266, 2004.

[124] Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pages 1263–1271, 2016.

[125] William W Hogan. Convergence results for some extensions of the Frank-Wolfe method. Technical report, CALIFORNIA UNIV LOS ANGELES WESTERN MANAGEMENT SCIENCE INST, 1971.

[126] Charles A Holloway. An extension of the Frank and Wolfe method of feasible directions. *Mathematical Programming*, 6(1):14–27, 1974.

[127] Robert Hooke and Terry A Jeeves. “direct search”solution of numerical and statistical problems. *Journal of the ACM (JACM)*, 8(2):212–229, 1961.

[128] Reshad Hosseini and Suvrit Sra. Matrix manifold optimization for gaussian mixtures. *Advances in Neural Information Processing Systems*, 28:910–918, 2015.

[129] S Hosseini and MR Pouryayevali. Nonsmooth optimization techniques on riemannian manifolds. *J. Optim. Theory Appl.*, 158(2):328–342, 2013.

[130] Seyedehsomayeh Hosseini, Boris Sholimovich Mordukhovich, and André Uschmajew. *Nonsmooth optimization and its applications.* International Series of Numerical Mathematics. Springer International Publishing, 2019.

[131] Seyedehsomayeh Hosseini and André Uschmajew. A riemannian gradient sampling algorithm for nonsmooth optimization on manifolds. *SIAM J. Optim.*, 27:173–189, 2017.

[132] Changwu Huang, Yuanxiang Li, and Xin Yao. A survey of automatic parameter tuning methods for metaheuristics. *IEEE transactions on evolutionary computation*, 24(2):201–216, 2019.

[133] James T Hungerford and Francesco Rinaldi. A general regularized continuous formulation for the maximum clique problem. *Mathematics of Operations Research*, 44(4):1161–1173, 2019.

[134] Alfredo N Iusem. On the convergence properties of the projected gradient method for convex optimization. *Computational & Applied Mathematics*, 22(1):37–52, 2003.

[135] Martin Jaggi. *Sparse convex optimization methods for machine learning.* PhD thesis, ETH Zurich, 2011.

[136] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

[137] Martin Jaggi and Marek Sulovský. A simple algorithm for nuclear norm regularized problems. In *ICML*, pages 471–478, 2010.

[138] Rujun Jiang and Xudong Li. Hölderian error bounds and kurdyka-łojasiewicz inequality for the trust region subproblem. *Mathematics of Operations Research*, 2022.

[139] Carl Johnell and Morteza Haghir Chehreghani. Frank-Wolfe optimization for dominant set clustering. *arXiv preprint arXiv:2007.11652*, 2020.

[140] David S Johnson. Cliques, coloring, and satisfiability: second dimacs implementation challenge. *DIMACS series in discrete mathematics and theoretical computer science*, 26:11–13, 1993.

[141] Hikaru G Jolliffe, Samir Diab, and Dimitrios I Gerogiorgis. Nonlinear optimization via explicit nrtl model solubility prediction for antisolvent mixture selection in artemisinin crystallization. *Organic Process Research & Development*, 22(1):40–53, 2018.

[142] Armand Joulin, Kevin Tang, and Li Fei-Fei. Efficient image and video co-localization with Frank-Wolfe algorithm. In *European Conference on Computer Vision*, pages 253–268. Springer, 2014.

[143] Zoran Kadelburg, Dusan Dukic, Milivoje Lukic, and Ivan Matic. Inequalities of Karamata, Schur and Muirhead, and some applications. *The Teaching of Mathematics*, 8(1):31–45, 2005.

[144] Jovan Karamata. Sur une inégalité relative aux fonctions convexes. *Publications de l'Institut Mathématique*, 1(1):145–147, 1932.

[145] Hamid R Karbasian and Brian C Vermeire. Gradient-free aerodynamic shape optimization using large eddy simulation. *Computers & Fluids*, 232:105185, 2022.

[146] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[147] Ehsan Kazemi, Thomas Kerdreux, and Liquang Wang. Generating structured adversarial attacks using Frank-Wolfe method. *arXiv preprint arXiv:2102.07360*, 2021.

[148] Thomas Kerdreux, Alexandre d'Aspremont, and Sebastian Pokutta. Restarting Frank-Wolfe. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1275–1283. PMLR, 2019.

[149] Thomas Kerdreux, Alexandre d'Aspremont, and Sebastian Pokutta. Projection-free optimization on uniformly convex sets. In *International Conference on Artificial Intelligence and Statistics*, pages 19–27. PMLR, 2021.

[150] Thomas Kerdreux, Lewis Liu, Simon Lacoste-Julien, and Damien Scieur. Affine invariant analysis of frank-wolfe on strongly convex sets. In *International Conference on Machine Learning*, pages 5398–5408. PMLR, 2021.

[151] T. G. Kolda, R. M. Lewis, and V. Torczon. Optimization by direct search: New perspectives on some classical and modern methods. 45(3):385–482, 2003.

[152] Tamara G Kolda, Robert Michael Lewis, and Virginia Torczon. Stationarity results for generating set search for linearly constrained optimization. *SIAM Journal on Optimization*, 17(4):943–968, 2007.

[153] Vladimir Kolmogorov. Practical Frank-Wolfe algorithms. *arXiv preprint arXiv:2010.09567*, 2020.

[154] IV Konnov. Simplified versions of the conditional gradient method. *Optimization*, 67(12):2275–2290, 2018.

[155] Piyush Kumar, Joseph SB Mitchell, and E Alper Yıldırım. Approximate minimum enclosing balls in high dimensions using core-sets. *Journal of Experimental Algorithmics*, 8:1–1, 2003.

[156] Simon Lacoste-Julien. Convergence rate of Frank-Wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

[157] Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *Advances in Neural Information Processing Systems*, volume 28, pages 496–504, 2015.

[158] Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 53–61, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[159] Dounia Lakhmiri, Sébastien Le Digabel, and Christophe Tribes. Hypernomad: Hyperparameter optimization of deep neural networks using mesh adaptive direct search. *ACM Transactions on Mathematical Software (TOMS)*, 47(3):1–27, 2021.

[160] Guanghui Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Data Sciences. Springer, Switzerland, 2020.

[161] Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.

[162] J. Larson and S. C. Billups. Stochastic derivative-free optimization using a trust region framework. 64:619–645, 2016.

[163] Jeffrey Larson, Matt Menickelly, and Stefan M Wild. Derivative-free optimization methods. *Acta Numer.*, 28:287–404, 2019.

[164] Larry J LeBlanc, Edward K Morlok, and William P Pierskalla. An efficient approach to solving the road network equilibrium traffic assignment problem. *Transportation Research*, 9(5):309–318, 1975.

[165] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Math. Program.*, 176(1):311–337, 2019.

[166] Evgeny S Levitin and Boris T Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966.

[167] Kfir Levy and Andreas Krause. Projection free online learning over smooth sets. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1458–1466, 2019.

[168] Robert Michael Lewis, Anne Shepherd, and Virginia Torczon. Implementing generating set search methods for linearly constrained minimization. *SIAM Journal on Scientific Computing*, 29(6):2507–2530, 2007.

[169] Robert Michael Lewis, Virginia Torczon, and Michael W Trosset. Direct search methods: then and now. *Journal of computational and Applied Mathematics*, 124(1-2):191–207, 2000.

[170] Guoyin Li and Ting Kei Pong. Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5):1199–1232, 2018.

[171] Jiaxiang Li, Krishnakumar Balasubramanian, and Shiqian Ma. Zeroth-order optimization on riemannian manifolds. 2020.

[172] G. Liuzzi, S. Lucidi, F. Rinaldi, and L. N. Vicente. Trust-region methods for the derivative-free optimization of nonsmooth black-box functions. 29:3012–3035, 2019.

[173] Giampaolo Liuzzi, Stefano Lucidi, and Francesco Rinaldi. An algorithmic framework based on primitive directions and nonmonotone line searches for black-box optimization problems with integer variables. *Mathematical Programming Computation*, 12(4):673–702, 2020.

[174] Giampaolo Liuzzi, Stefano Lucidi, and Marco Sciandrone. Sequential penalty derivative-free methods for nonlinear constrained optimization. *SIAM J. Optim.*, 20(5):2614–2635, 2010.

[175] Francesco Locatello, Rajiv Khanna, Michael Tschannen, and Martin Jaggi. A unified optimization view on generalized matching pursuit and frank-wolfe. In *Artificial Intelligence and Statistics*, pages 860–868. PMLR, 2017.

[176] Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[177] R Duncan Luce and Albert D Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949.

[178] Stefano Lucidi and Marco Sciandrone. A derivative-free algorithm for bound constrained optimization. *Comput. Optim. Appl.*, 21(2):119–142, 2002.

[179] Stefano Lucidi and Marco Sciandrone. On the global convergence of derivative-free methods for unconstrained optimization. *SIAM J. Optim.*, 13:97–116, 2002.

[180] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[181] OL Mangasarian. Machine learning via polyhedral concave minimization. In *Applied Mathematics and Parallel Computing*, pages 175–188. Springer, 1996.

[182] Alessio Massaro, Marcello Pelillo, and Immanuel M Bomze. A complementary pivoting approach to the maximum weight clique problem. *SIAM J. Optim*, 12(4):928–948, 2002.

[183] Juan C. Meza and Monica L. Martinez. Direct search methods for the molecular conformation problem. *Journal of Computational Chemistry*, 15(6):627–632, 1994.

[184] BF Mitchell, Vladimir Fedorovich Demyanov, and VN Malozemov. Finding the point of a polyhedron closest to the origin. *SIAM Journal on Control*, 12(1):19–26, 1974.

[185] Maria Mitradjieva and Per Olov Lindberg. The stiff is moving-conjugate direction frank-wolfe methods with applications to traffic assignment. *Transportation Science*, 47(2):280–293, 2013.

[186] Jorge J Moré and Stefan M Wild. Benchmarking derivative-free optimization algorithms. *SIAM J. Optim.*, 20(1):172–191, 2009.

[187] Hassan Mortagy, Swati Gupta, and Sebastian Pokutta. Walking in the shadow: A new perspective on descent directions for constrained minimization. *Advances in Neural Information Processing Systems*, 33, 2020.

[188] Theodore S Motzkin and Ernst G Straus. Maxima for graphs and a new proof of a theorem of Turán. *Canad. J. Math.*, 17:533–540, 1965.

[189] Cun Mu, Yuqian Zhang, John Wright, and Donald Goldfarb. Scalable robust matrix recovery: Frank–Wolfe meets proximal methods. *SIAM Journal on Scientific Computing*, 38(5):A3291–A3317, 2016.

[190] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[191] Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 1998.

[192] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[193] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527–566, 2017.

[194] Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer Science & Business Media, 2006.

[195] Julie Nutini, Issam Laradji, and Mark Schmidt. Let's make block coordinate descent go fast: Faster greedy rules, message-passing, active-set complexity, and superlinear convergence. *arXiv preprint arXiv:1712.08859*, 2017.

[196] Julie Nutini, Mark Schmidt, and Warren Hare. "active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern? *Optimization Letters*, 13(4):645–655, 2019.

[197] Anton Osokin, Jean-Baptiste Alayrac, Isabella Lukasewitz, Puneet Dokania, and Simon Lacoste-Julien. Minding the gaps for block Frank-Wolfe optimization of structured svms. In *International Conference on Machine Learning*, pages 593–602. PMLR, 2016.

[198] C. Paquette and K. Scheinberg. A stochastic line search method with expected complexity analysis. 30:349–376, 2020.

[199] Jeffrey Pattillo, Nataly Youssef, and Sergiy Butenko. On clique relaxation models in network analysis. *European Journal of Operational Research*, 226(1):9–18, 2013.

[200] Javier Peña and Daniel Rodriguez. Polytope conditioning and linear convergence of the Frank-Wolfe algorithm. *Mathematics of Operartions Research*, 44(1):1–18, 2018.

[201] Fabian Pedregosa, Geoffrey Negiar, Armin Askari, and Martin Jaggi. Linearly convergent Frank-Wolfe with backtracking line-search. In *International Conference on Artificial Intelligence and Statistics*, pages 1–10. PMLR, 2020.

[202] Olga Perederieieva, Matthias Ehrgott, Andrea Raith, and Judith YT Wang. A framework for and empirical study of algorithms for traffic assignment. *Computers & Operations Research*, 54:90–107, 2015.

[203] Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[204] Michael JD Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, 7(2):155–162, 1964.

[205] Chao Qu, Yan Li, and Huan Xu. Non-convex conditional gradient sliding. In *International Conference on Machine Learning*, pages 4208–4217. PMLR, 2018.

[206] Luis Rademacher and Chang Shu. The smoothed complexity of Frank-Wolfe methods via conditioning of random matrices and polytopes. *arXiv preprint arXiv:2009.12685*, 2020.

[207] Francesco Rinaldi, Fabio Schoen, and Marco Sciandrone. Concave programming for minimizing the zero-norm over polyhedral sets. *Computational Optimization and Applications*, 46(3):467–486, 2010.

[208] Francesco Rinaldi and Damiano Zeffiro. A unifying framework for the analysis of projection-free first-order methods under a sufficient slope condition. *arXiv preprint arXiv:2008.09781*, 2020.

[209] Francesco Rinaldi and Damiano Zeffiro. Avoiding bad steps in Frank Wolfe variants. *Computational Optimization and Applications*, 2022.

[210] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

[211] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, Berlin, 2009.

[212] HoHo Rosenbrock. An automatic method for finding the greatest or least value of a function. *The computer journal*, 3(3):175–184, 1960.

[213] Anit Kumar Sahu and Soummya Kar. Decentralized zeroth-order constrained stochastic optimization algorithms: Frank–Wolfe and variants with applications to black-box adversarial attacks. *Proceedings of the IEEE*, 108(11):1890–1905, 2020.

[214] Neel Shah, Vladimir Kolmogorov, and Christoph H Lampert. A multi-plane block-coordinate Frank-Wolfe algorithm for training structural svms with a costly max-oracle. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2737–2745, 2015.

[215] S. Shashaani, F. S. Hashemi, and R. Pasupathy. ASTRO-DF: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. 28:3145–3176, 2018.

[216] WGRFR Spendley, George R Hext, and Francis R Himsworth. Sequential application of simplex designs in optimisation and evolutionary operation. *Technometrics*, 4(4):441–461, 1962.

[217] Vladimir Stozhkov, Austin Buchanan, Sergiy Butenko, and Vladimir Boginski. Continuous cubic formulations for cluster detection problems in networks. *Math. Program.*, online, 2020.

[218] Yifan Sun, Halyun Jeong, Julie Nutini, and Mark Schmidt. Are we there yet? manifold identification of gradient-related proximal methods. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1110–1119, 2019.

[219] Yijia Sun, Nikolaos V Sahinidis, Anantha Sundaram, and Myun-Seok Cheon. Derivative-free optimization for chemical product design. *Current Opinion in Chemical Engineering*, 27:98–106, 2020.

[220] Arie Tamir. A strongly polynomial algorithm for minimum convex separable quadratic cost flow problems on two-terminal series-parallel networks. *Math. Program.*, 59:117–132, 1993.

[221] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[222] Virginia Torczon. On the convergence of pattern search algorithms. *SIAM Journal on optimization*, 7(1):1–25, 1997.

[223] Klaus Truemper. Unimodular matrices of flow problems with additional constraints. *Networks*, 7(4):343–358, 1977.

[224] Svyatoslav Trukhanov, Chitra Balasubramaniam, Balabhaskar Balasundaram, and Sergiy Butenko. Algorithms for detecting optimal hereditary structures in graphs, with application to clique relaxations. *Comput. Optim. Appl.*, 56(1):113–130, 2013.

[225] Bart Vandereycken. Riemannian and multilevel optimization for rank-constrained matrix problems. PhD thesis, Department of Computer Science, KU Leuven, 2010.

[226] Vladimir Vapnik. *The Nature of Statistical Learning Theory.* Springer, 2013.

[227] Bruno Veloso, João Gama, and Benedita Malheiro. Self hyper-parameter tuning for data streams. In *International Conference on Discovery Science*, pages 241–255. Springer, 2018.

[228] Luís Nunes Vicente. Worst case complexity of direct search. *EURO J. Comput. Optim.*, 1(1):143–153, 2013.

[229] B. von Bahr and C.-G. Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36:299–303, 1965.

[230] Bengt Von Bahr. On the convergence of moments in the central limit theorem. *The Annals of Mathematical Statistics*, pages 808–818, 1965.

[231] Balder Von Hohenbalken. Simplicial decomposition in nonlinear programming algorithms. *Mathematical Programming*, 13(1):49–68, 1977.

[232] Haoyue Wang, Haihao Lu, and Rahul Mazumder. Frank-Wolfe methods with an unbounded feasible region and applications to structured learning. *arXiv preprint arXiv:2012.15361*, 2020.

[233] Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of admm in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78(1):29–63, 2019.

[234] Yu-Xiang Wang, Veeranjaneyulu Sadhanala, Wei Dai, Willie Neiswanger, Suvrit Sra, and Eric Xing. Parallel and distributed block-coordinate Frank-Wolfe algorithms. In *International Conference on Machine Learning*, pages 1548–1557. PMLR, 2016.

[235] John Glen Wardrop. Road paper. some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers*, 1(3):325–362, 1952.

[236] Andrés Weintraub, Carmen Ortiz, and Jaime González. Accelerating convergence of the Frank-Wolfe algorithm. *Transportation Research Part B: Methodological*, 19(2):113–122, 1985.

[237] Philip Wolfe. Convergence theory in nonlinear programming. In J. Abadie, editor, *Integer and nonlinear programming*, pages 1–36. North Holland, 1970.

[238] Philip Wolfe. Finding the nearest point in a polytope. *Mathematical Programming*, 11(1):128–149, 1976.

[239] Margaret H Wright. Direct search methods: Once scorned, now respectable. *Pitman Research Notes in Mathematics Series*, pages 191–208, 1996.

[240] Stephen J Wright. Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization*, 31(4):1063–1079, 1993.

[241] Qinghua Wu and Jin-Kao Hao. A review on algorithms for maximum clique problems. *European Journal of Operational Research*, 242(3):693–709, 2015.

[242] Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

[243] Yi Xu and Tianbao Yang. Frank-Wolfe method is automatically adaptive to error bound condition. *arXiv preprint arXiv:1810.04765*, 2018.

[244] Teng-Teng Yao, Zhi Zhao, Zheng-Jian Bai, and Xiao-Qing Jin. A riemannian derivative-free polak–ribiére–polyak method for tangent vector field. *Numerical Algorithms*, 86(1):325–355, 2021.

[245] Guixin Ye, Zhanyong Tang, Dingyi Fang, Zhanxing Zhu, Yansong Feng, Pengfei Xu, Xiaojiang Chen, and Zheng Wang. Yet another text captcha solver: A generative adversarial network based approach. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pages 332–348, 2018.

[246] E Alper Yıldırım. Two algorithms for the minimum enclosing ball problem. *SIAM Journal on Optimization*, 19(3):1368–1391, 2008.

[247] Haiyuan Yu, Alberto Paccanaro, Valery Trifonov, and Mark Gerstein. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, 22(7):823–829, 2006.

[248] Hongchao Zhang and William W Hager. A nonmonotone line search technique and its application to unconstrained optimization. *SIAM journal on Optimization*, 14(4):1043–1056, 2004.

[249] Li Zhang, Weijun Zhou, and Dong-Hui Li. A descent modified Polak-Ribière-Polyak conjugate gradient method and its global convergence. *IMA Journal of Numerical Analysis*, 26(4):629–640, 2006.