RADIAL PROJECTIONS, CONVEX FEASIBILITY PROBLEMS AND MARGIN MAXIMIZATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Song Zhou

August 2023

RADIAL PROJECTIONS, CONVEX FEASIBILITY PROBLEMS AND MARGIN MAXIMIZATION

Song Zhou, Ph.D.

Cornell University 2023

This work comprises two parts. Part I focuses on the convex feasibility problem (finding or approximating a point in the intersection of finitely many closed convex sets). We avoid the need for orthogonal projections by using *radial projections*, introduced by Renegar[23]. The main requirement is that an interior point is known in each of the sets considered. By developing Renegar's theory, we obtain a family of radial projection-based algorithms for the convex feasibility problem which recover the linear convergence rates of orthogonal projection-based methods. Through studying different assumptions on the emptiness of the interior of the intersection set in the convex feasibility problem, we also exhibit how radial projections can be applied to solve constrained optimization problems when certain conditions are met.

Part II can be seen as an application of the theory of radial projections developed in Part I. Here, we revisit the notion of maximal-margin classifiers, from around 2000, but now from a general perspective – the intersections of generic closed convex cones, not just half-spaces (i.e., the perceptron). This requires extending concepts and establishing more general theory of the margin function, which is achieved by applying and refining the results in Part I in the conic case. Even more interestingly, we are led to the first $\tilde{O}(1/\epsilon)$ first-order method for approximating, within relative error ϵ, the margin-maximizer of the intersection cone. Previous results, only in the case of the perceptron, were $O(1/\epsilon^2)$, making our re-

sult a notable improvement even in the most basic of cases.

BIOGRAPHICAL SKETCH

Song Zhou, also known as Sam Zhou, was born and raised in Guangzhou, Guangdong, China. He completed his bachelor's degree in Mathematics at Tsinghua University in 2018. He then started his doctoral studies in the School of Operations Research & Information Engineering at Cornell University with a concentration in continuous optimization, advised by Professor James Renegar.

ACKNOWLEDGEMENTS

When I first embarked on my Ph.D. career, I did not really know what I had signed up for. I would like to take this opportunity to express my sincerest gratitude to all the people who made this journey possible.

First and foremost, I owe my deepest appreciation to my advisor, Jim Renegar. Humble, patient, and honest, he is a true inspiration and role model in both my academic and personal life. His unwavering trust and support guided me through this treacherous doctoral pursuit. The precious conversations I had with him will never be forgotten. I am truly fortunate to have Jim as my advisor.

I am also very thankful to the rest of the ORIE faculty, who offered a plethora of fascinating courses. In particular, I would like to acknowledge Professors Damek Davis and Adrian Lewis for showing me the breadth and depth of the world of continuous optimization.

Throughout my Ph.D., I have been helped by numerous friends during times of difficulty. While our paths might converge and diverge with the passing of time, I am grateful to have met and spent time with them in my life.

I would like to thank my parents and grandparents for their love and support. It is not always easy to bear with a free-spirited kid.

Lastly, thank you to my wife, Qinyi Luo. You are the best thing in my life.

TABLE OF CONTENTS

LIST OF FIGURES

Part I

Radial Projections and Convex Feasibility Problems

CHAPTER 1

INTRODUCTION

Given a finite collection of closed convex sets with nonempty intersection, the *convex feasibility problem* concerns approximating (or finding) a point in the intersection set. Traditionally, convex feasibility problems are solved via orthogonal projections onto the individual sets (see, e.g., [1, 2, 4, 11, 30]).

Orthogonal projections onto generic closed convex sets are difficult to compute, even in the case of ellipsoids (see [7, 16]). In this work, to avoid the need for orthogonal projections, we introduce *radial projections* and the γ function by assuming *a known interior point* (used as the *reference point*) in each of the individual sets, and develop algorithms for the convex feasibility problem accordingly.

The idea of radial projections and the γ function for a single closed convex cone were considered by Renegar in [23]. In Chapter 2, we generalize some previous results to generic closed convex sets and propose new ones, with an emphasis on the connection between orthogonal and radial projections. In order to apply the technique to convex feasibility problems, we develop the theory for multiple closed convex sets in Chapter 3. Our discussion features the function γ_0, which is defined to be the maximum of the γ functions of the individual sets. The analysis includes Theorem 3.1, a *linear regularity* result for generic closed convex sets when the intersection has non-empty interior, which is of independent interest. Chapter 4 briefly goes over the case where the convex feasibility problem involves an affine subspace, while the reference points do not necessary lie in that affine subspace.

With the theory in place, we propose and analyze two radial projection-based algorithms for the convex feasibility problem in Chapter 5. Both methods are

subgradient methods whose step sizes follow Polyak's rule [21]. By assuming an intersection with non-empty interior, both algorithms approximate the intersection set at linear rates, similar to orthogonal projection-based methods[1]. The first algorithm has γ_0 as its objective function, while the second algorithm cycles over the γ functions of the individual sets. This chapter concludes with numerical performances of these two algorithms applied to the feasibility problem of randomly generated ellipsoids.

Chapter 6 presents a stochastic algorithm for the convex feasibility problem, which is essentially an application of the methodology developed by Renegar and Zhou in [25]. While the methods presented in Chapter 5 and 6 only approximate the intersection, in Chapter 7, we discuss an algorithm which finds a feasible point in the intersection and terminates in finitely many iterations. The finiteness of the algorithm is achieved via sequential estimates of the minimum of γ_0.

The results in Chapter 5-7 all posit intersection sets with non-empty interior. In Chapter 8, when the normal cones to the individual sets at their intersection possess certain structures (see Assumption 8.1), we show that the convex feasibility problem can still be solved at a linear rate. By adopting the parallel scheme proposed by Renegar and Grimmer[24], a similar assumption, i.e., Assumption 9.1, allows us to tackle constrained optimization problems via radial projections, which is the subject of Chapter 9.

[1]See the discussion around (5.3) and the paragraph following the proof of Theorem 5.1 for more detailed comparisons of the canonical orthogonal projection-based algorithm and its radial projection-based counterpart.

CHAPTER 2

DEFINITION AND PROPERTIES OF THE γ FUNCTION FOR A SINGLE SET

Let $\mathcal{E}$ denote a finite-dimensional Euclidean space endowed with an inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\| \cdot \|$. Consider a closed, convex set $S \subset \mathcal{E}$. We assume that S has nonempty interior, and a *reference point* $e \in \text{int}(S)$ is given. For any $x \in \mathcal{E}$ and $\lambda > 0$, let

$$x(\lambda) := e + \lambda(x - e). \tag{2.1}$$

Define nonnegative functions

$$\alpha(x) := \sup\{\lambda > 0 \mid x(\lambda) \in S\} \tag{2.2}$$

and

$$\gamma(x) := 1/\alpha(x) = \inf\{\lambda^{-1} \mid x(\lambda) \in S,\ \lambda > 0\}. \tag{2.3}$$

Since $e \in \text{int}(S)$, we have $\alpha(x) > 0$ for all $x \in \mathcal{E}$. When $\alpha(x) = \infty$, let $\gamma(x) = 0$. Then we have

$$\gamma(x) = 0 \iff x - e \in \text{recc}(S - e), \tag{2.4}$$

where $\text{recc}(S - e) := \{d \in \mathcal{E} \mid \lambda d \in S - e, \forall \lambda > 0\}$ denotes the recession cone of the set $S - e$.

For any $x \in \mathcal{E}$ and $\lambda > 0$, we see that

$$\begin{aligned} \gamma(x(\lambda)) &= \inf\left\{\mu^{-1} \mid e + \mu(x(\lambda) - e) \in S,\ \mu > 0\right\} \\ &= \inf\left\{\mu^{-1} \mid e + \mu\lambda(x - e) \in S,\ \mu > 0\right\} \\ &= \lambda \inf\left\{\mu^{-1} \mid e + \mu(x - e) \in S,\ \mu > 0\right\} \\ &= \lambda\gamma(x). \end{aligned} \tag{2.5}$$

Therefore, γ is a positively homogeneous function if we replace the origin with e.

For $x \in \mathcal{E}$ such that $\gamma(x) > 0$ (i.e., $\alpha(x) < \infty$), define

$$\pi_S(x) := x(\alpha(x)) = e + \alpha(x)(x - e) = e + \frac{x - e}{\gamma(x)}. \tag{2.6}$$

Since S is closed and convex, by the definition of $\alpha(x)$, we immediately see that

$$S \cap \{x(\lambda) \mid \lambda > 0\} = (e, e + \alpha(x)(x - e)] = (e, \pi_S(x)]. \tag{2.7}$$

When $x \notin S$, we refer to $\pi_S(x)$ as the *radial projection* of x onto S. It should be noted that by (2.5), $\pi_S(x(\lambda))$ remains unchanged for all $\lambda > 0$:

$$\pi_S(x(\lambda)) = e + \frac{x(\lambda) - e}{\gamma(x(\lambda))} = e + \frac{\lambda(x - e)}{\lambda\gamma(x)} = e + \frac{x - e}{\gamma(x)} = \pi_S(x). \tag{2.8}$$

The α function defined in (2.2) is basically equivalent to the α_1 function considered by Renegar in [23]. The purpose of [23] is to develop a set of "projected" subgradient methods, with radial projections replacing the traditional orthogonal projections. Renegar also considered the $\lambda_{\min}$ function, which can be seen as the γ function for convex cones.

For general closed convex sets, the value of $\gamma(x)$ cannot be computed exactly. In [23], the author discussed the approximation of the α_1 function in detail. Similarly, in our case, the value of $\gamma(x)$ can also be approximated accurately. All that is required is a bisection on the half-line $\{x(\lambda) \mid \lambda > 0\}$ for accurately approximating where this half-line intersects the boundary of S. If such intersection does not exist, then $\gamma(x) = 0$, and the bisection halves its estimate of $\gamma(x)$ at each iteration. Otherwise, the bisection approximates the position of $\pi_S(x)$ (i.e., where the half-line intersects the boundary) and $\gamma(x)$ can be estimated accordingly. For most closed convex sets, this can be accomplished far more easily than traditional orthogonal projections.

In what follows, we assume $\gamma(x)$ can be computed exactly.

In the rest of this chapter, we develop some properties of the γ function for generic closed convex sets, while emphasizing its connection with the distance function. We start with some immediate observations:

Lemma 2.1. *For any $x \in \mathcal{E}$, we have*

$$\begin{cases} \gamma(x) < 1 \iff x \in \mathrm{int}(S); \\ \gamma(x) = 1 \iff x \in \mathrm{bdy}(S) \iff \pi_S(x) = x; \\ \gamma(x) > 1 \iff x \notin S, \end{cases}$$

and S is the 1-sublevel set of γ.

Proof. Since $e \in \mathrm{int}(S)$, there exists $r > 0$ such that $B(e, r) \subseteq \mathrm{int}(S)$. Here $B(e, r)$ denotes the open ball centered at e with radius r.

- When $\gamma(x) = 0$, we have $x - e \in \mathrm{recc}(S - e)$. According to Lemma A.1 in Appendix A.1, for any $w \in B(e, r) \subset S$, $x - e \in \mathrm{recc}(S - w) \subseteq (S - w)$. Hence

  $$x \in B(x, r) = B(e, r) + (x - e) \subseteq S,$$

  and $x \in \mathrm{int}(S)$.

- When $\gamma(x) \in (0, 1)$, note that $\pi_S(x) \in S$ and

  $$x = (1 - \gamma(x))e + \gamma(x)\left(e + \frac{(x - e)}{\gamma(x)}\right) = (1 - \gamma(x))e + \gamma(x)\pi_S(x),$$

  we get

  $$x \in B(x, (1 - \gamma(x))r) = (1 - \gamma(x))B(e, r) + \gamma(x)\pi_S(x) \subset \mathrm{int}(S),$$

  and $x \in \mathrm{int}(S)$.

- When $\gamma(x) = 1$, we immediately get $x \in \text{bdy}(S)$ by the definition of $\gamma(x)$.
- When $\gamma(x) > 1$, then $x \notin (e, \pi_S(x)]$. By (2.7), we see that $x \notin S$.

□

Lemma 2.1 shows that the 1-sublevel set of γ is S. To study generic sublevel sets of γ, for any $t \geq 0$, let

$$S(t) := \{x \in \mathcal{E} \mid \gamma(x) \leq t\} \tag{2.9}$$

denote the t-sublevel set of γ. Then by the definition of γ and the closedness of S, for any $t > 0$, we have

$$\gamma(x) \leq t \iff e + \frac{x-e}{t} \in S \iff x \in e + t(S-e).$$

Hence when $t > 0$,

$$S(t) = e + t(S-e). \tag{2.10}$$

Moreover, by (2.10) and Lemma 2.1, for all $t > 0$, we get

$$\begin{aligned} \text{int}(S(t)) &= \text{int}(e + t(S-e)) \\ &= e + t \cdot \text{int}(S-e) \\ &= \{e + t(x-e) \mid \gamma(x) < 1\} \\ &= \{x \in \mathcal{E} \mid \gamma(x) < t\}. \end{aligned} \tag{2.11}$$

We close this part with the following result:

Lemma 2.2. *The function γ is convex.*

Proof. Consider any $x, y \in \mathcal{E}$ and $\lambda \in (0,1)$. We wish to show

$$\gamma(\lambda x + (1-\lambda)y) \leq \lambda\gamma(x) + (1-\lambda)\gamma(y). \tag{2.12}$$

If $\gamma(x) = \gamma(y) = 0$, then (2.4) gives $x - e, y - e \in \mathrm{recc}(S - e)$. By the convexity of $\mathrm{recc}(S - e)$, we get

$$(\lambda x + (1 - \lambda)y) - e = \lambda(x - e) + (1 - \lambda)(y - e) \in \mathrm{recc}(S - e),$$

and $\gamma(\lambda x + (1 - \lambda)y) = 0$. Hence (2.12) holds.

Now assume $\gamma(x) > 0$. According to (2.5), by rescaling x and y with respect to e if necessary, we may assume $\gamma(x) = 1$ and $x \in \mathrm{bdy}(S)$. Let $t = \lambda + (1 - \lambda)\gamma(y)$ in (2.10), we see that (2.12) is true if and only if $\lambda x + (1 - \lambda)y \in S(\lambda + (1 - \lambda)\gamma(y))$.

- If $\gamma(y) = 0$, then $y - e \in \mathrm{recc}(S - e)$. Using (2.5), we see that $\gamma(x(\lambda)) = \lambda\gamma(x) = \lambda$. Hence by (2.10) and (2.4),

$$\begin{aligned} \lambda x + (1 - \lambda)y &= x(\lambda) + (1 - \lambda)(y - e) \\ &\in (e + \lambda(S - e)) + \mathrm{recc}(S - e) = e + \lambda(S - e) = S(\lambda), \end{aligned}$$

  and the statement holds.

- If $\gamma(y) > 0$, then $y = e + \gamma(y)(\pi_S(y) - e)$. Hence

$$\begin{aligned} &\lambda x + (1 - \lambda)y \\ &= \lambda(e + (x - e)) + (1 - \lambda)\left(e + \gamma(y)(\pi_S(y) - e)\right) \\ &= e + (\lambda + (1 - \lambda)\gamma(y))\left(\left(\frac{\lambda x}{\lambda + (1 - \lambda)\gamma(y)} + \frac{((1 - \lambda)\gamma(y))\,\pi_S(y)}{\lambda + (1 - \lambda)\gamma(y)}\right) - e\right) \\ &\in e + (\lambda + (1 - \lambda)\gamma(y))\,(S - e) \\ &= S(\lambda + (1 - \lambda)\gamma(y)), \end{aligned}$$

  where the last but one line follows from the fact that x and $\pi_S(y)$ are both in the convex set S, and the last equality is due to (2.10).

□

2.1 The γ function and the distance function

Recall that the distance between a vector x and a closed convex set S can be defined as

$$\mathrm{dist}(x, S) := \|x - \mathrm{P}_S(x)\|,$$

where $\mathrm{P}_S(x) := \mathrm{argmin}_{z \in S} \|z - x\|$ is the *orthogonal projection* of x onto S. One immediately sees that $\mathrm{P}_S(x) = x$ whenever $x \in S$. For generic closed convex sets, orthogonal projections are nontrivial to compute (e.g., [7] and [16] are dedicated to iterative methods for orthogonal projections onto ellipsoids).

In contrast, when a reference point $e \in \mathrm{int}(S)$ is known, for $x \notin S$, we can compute the radial projection $\pi_S(x) \in \mathrm{bdy}(S)$ by evaluating $\gamma(x)$.[1] The purpose of this section is to look into the connection between radial projections and orthogonal projections (i.e., $\gamma(x)$ v.s. $\mathrm{dist}(x, S)$) when $x \notin S$.

We start by studying the subgradients of γ. For any $z \in S$, let

$$\mathcal{N}_S(z) := \{d \in \mathcal{E} \mid \langle z, d\rangle \geq \langle x, d\rangle,\ \forall x \in S\}$$

denote the normal cone to S at z. Recall that when $x \notin S$, we have the following characterization of the subgradients of the distance function:

$$\partial_x \mathrm{dist}(x, S) = \{d \in \mathcal{E} \mid d \in \mathcal{N}_S(\mathrm{P}_S(x)) \text{ and } \|d\| = 1\}. \tag{2.13}$$

The subgradients of γ admit a similar formula, with $\pi_S(x)$ playing the role of $\mathrm{P}_S(x)$ in (2.13):

Proposition 2.1. *Given $x \in \mathcal{E}$, when $\gamma(x) > 0$, we have*

$$\partial\gamma(x) = \{d \in \mathcal{E} \mid d \in \mathcal{N}_S(\pi_S(x)) \text{ and } \langle \pi_S(x) - e, d\rangle = 1\}.$$

[1] See the remark following Proposition 2.1 for discussions on a first-order oracle for γ.

Remark. When $\gamma(x) > 0$, to get a subgradient of γ at x, it suffices to have an oracle which generates non-zero normal vectors to S at its boundary points.

Proof. Given $\lambda > 0$, for any $x, g \in \mathcal{E}$, we have

$$\begin{aligned} g \in \partial\gamma(x) &\iff \gamma(x) + \langle y - x, g\rangle \le \gamma(y),\ \forall y \in \mathcal{E} \\ &\iff \lambda\gamma(x) + \lambda\langle y - x, g\rangle \le \lambda\gamma(y),\ \forall y \in \mathcal{E} \\ &\overset{(2.5)}{\iff} \gamma(x(\lambda)) + \langle y(\lambda) - x(\lambda), g\rangle \le \gamma(y(\lambda)),\ \forall y \in \mathcal{E} \\ &\iff g \in \partial\gamma(x(\lambda)). \end{aligned}$$

When $\gamma(x) > 0$, let $\lambda = 1/\gamma(x)$, we get

$$\partial\gamma(x) = \partial\gamma(\pi_S(x)). \tag{2.14}$$

Consequently, by Lemma 2.1, we may assume $x \in \mathrm{bdy}(S)$ for the rest of this proof.

Consider any $d \in \mathcal{N}_S(x)$. Since $e \in \mathrm{int}(S)$, we have $\langle x - e, d\rangle > 0$ for all non-zero $d \in \mathcal{N}_S(x)$. Thus the set considered on right-hand side of Proposition 2.1 is non-empty (due to rescaling). Now pick any $d \in \mathcal{N}_S(\pi_S(x))$ satisfying $\langle x - e, d\rangle = 1$. For any $y \in \mathcal{E}$, we have

$$\gamma(x) + \langle y - x, d\rangle = 1 + \langle y - x, d\rangle = \langle x - e, d\rangle + \langle y - x, d\rangle = \langle y - e, d\rangle.$$

Here the first equality follows from Lemma 2.1. Thus

$$\begin{aligned} \left\langle e + \frac{y - e}{\gamma(x) + \langle y - x, d\rangle}, d \right\rangle &= \left\langle e + \frac{y - e}{\langle y - e, d\rangle}, d \right\rangle \\ &= \langle e, d\rangle + 1 = \langle e, d\rangle + \langle x - e, d\rangle = \langle x, d\rangle. \end{aligned}$$

Since $d \in \mathcal{N}_S(x)$, we see that

$$e + \frac{y - e}{\gamma(x) + \langle y - x, d\rangle} \notin \mathrm{int}(S).$$

Hence by (2.5) and Lemma 2.1,

$$\frac{\gamma(y)}{\gamma(x)+\langle y-x,d\rangle}=\gamma\left(e+\frac{y-e}{\gamma(x)+\langle y-x,d\rangle}\right)\geq 1,$$

and we conclude that $d\in\partial\gamma(x)$.

On the other hand, first note that for any $g\notin\mathcal{N}_S(x)$, there exists $y'\in S$ such that $\langle x,g\rangle<\langle y',g\rangle$. By Lemma 2.1, we get

$$\gamma(y')\leq 1=\gamma(x)<\gamma(x)+\langle y'-x,g\rangle.$$

Hence $g\notin\partial\gamma(x)$, and we conclude that $\partial\gamma(x)\subseteq\mathcal{N}_S(x)$.

When $g\in\mathcal{N}_S(x)$ but $\langle x-e,g\rangle>1$, by (2.5), we have

$$2=\gamma(x(2))\geq\gamma(x)+\langle(x(2)-x,g\rangle=\gamma(x)+\langle x-e,g\rangle>2,$$

which is a contradiction. When $\langle x-e,g\rangle<1$, we can obtain a similar contradiction by studying $x(\frac{1}{2})$. Hence we get $\langle x-e,g\rangle=1$ for all $g\in\partial\gamma(x)$. □

For any $x\in S$, define

$$r_S(x):=\max\{r\geq 0\mid\overline{B(x,r)}\subseteq S\}. \tag{2.15}$$

Then $e\in\operatorname{int}(S)$ implies $r_S(e)>0$. The next result is a consequence of Proposition 2.1:

Lemma 2.3. *For all $x\in\mathcal{E}$ satisfying $\gamma(x)>0$, we have*

$$\partial\gamma(x)\subset\overline{B(\vec{0},1/r_S(e))}.$$

Consequently, γ is $1/r_S(e)$-Lipschitz.

Proof. For any $x \notin S(0)$, consider any $g \in \partial\gamma(x)$. By Proposition 2.1, we have $g \neq \vec{0}$. Due to the definition of $r_S(e)$, we see that

$$e + \frac{r_S(e)}{\|g\|} g \in \overline{B(\vec{0}, r_S(e))} \subseteq S.$$

Again by Proposition 2.1, we get $g \in \mathcal{N}_S(\pi_S(x))$ and

$$\langle e, g\rangle + 1 = \langle \pi_S(x), g\rangle \geq \left\langle e + \frac{r_S(e)}{\|g\|} g, g \right\rangle = \langle e, g\rangle + r_S(e)\|g\|.$$

Hence $\|g\| \leq 1/r_S(e)$. For all $y \in \mathcal{E}$ satisfying $\gamma(y) \leq \gamma(x)$, we have

$$|\gamma(x) - \gamma(y)| = \gamma(x) - \gamma(y) \leq \langle x - y, g\rangle \leq \frac{\|x - y\|}{r_S(e)}.$$

As for $y \in \mathcal{E}$ such that $\gamma(y) > \gamma(x)$, we can show a similar inequality by studying the subgradients of γ at y, and the Lipschitzness of γ follows. □

On the other hand, we also have the following lower bound on the growth rate of γ outside of S:

Lemma 2.4. *When $x \notin S$, we have*

$$\frac{\gamma(x) - 1}{\operatorname{dist}(x, S)} \geq \frac{1}{\|\pi_S(x) - e\|} \geq \frac{1}{\|\mathrm{P}_S(x) - e\|} \geq \frac{1}{\|x - e\|}.$$

Proof. When $x \notin S$, by Lemma 2.1, we have $\gamma(x) > 1$ and $\pi_S(x) \in (e, x)$. Since $\pi_S(x) \in S$,

$$\gamma(x) - 1 = \frac{\|x - e\|}{\|\pi_S(x) - e\|} - 1 = \frac{\|x - \pi_S(e)\|}{\|\pi_S(x) - e\|} \geq \frac{\operatorname{dist}(x, S)}{\|\pi_S(x) - e\|}. \tag{2.16}$$

By the triangle inequality, we get

$$\|x - \pi_S(x)\| + \|\pi_S(x) - e\| = \|x - e\| \leq \|x - \mathrm{P}_S(x)\| + \|\mathrm{P}_S(x) - e\|.$$

Also note that by the definition of $\mathrm{P}_S(x)$, we have $\|x - \pi_S(x)\| \geq \|x - \mathrm{P}_S(x)\|$. Hence

$$\begin{aligned} \|\pi_S(x) - e\| &\leq (\|x - \mathrm{P}_S(x)\| + \|\mathrm{P}_S(x) - e\|) - \|x - \pi_S(x)\| \\ &\leq \|x - \mathrm{P}_S(x)\| + \|\mathrm{P}_S(x) - e\| - \|x - \mathrm{P}_S(x)\| \\ &= \|\mathrm{P}_S(x) - e\|. \end{aligned} \tag{2.17}$$

Combined with (2.16), we get

$$\gamma(x) - 1 \geq \frac{\mathrm{dist}(x, S)}{\|\pi_S(x) - e\|} \geq \frac{\mathrm{dist}(x, S)}{\|\mathrm{P}_S(x) - e\|} \geq \frac{\mathrm{dist}(x, S)}{\|x - e\|}.$$

□

For any $x \notin S$, combining Lemma 2.3 and Lemma 2.4 yields

$$\frac{1}{\|\pi_S(x) - e\|} \leq \frac{\gamma(x) - 1}{\mathrm{dist}(x, S)} \leq \frac{1}{r_S(e)}. \tag{2.18}$$

When S is bounded, let $R_S(e) := \max\{\|z - e\| \mid z \in S\}$. Then for $x \notin S$, we have

$$\frac{1}{R_S(e)} \leq \frac{\gamma(x) - 1}{\mathrm{dist}(x, S)} \leq \frac{1}{r_S(e)}.$$

By (2.18), when $x \notin S$, the quantity $\gamma(x) - 1$ can be seen as an approximation of a scalar multiple of $\mathrm{dist}(x, S)$, and the *condition number* $\frac{\|\pi_S(x) - e\|}{r_S(e)}$ characterizes the tightness of the approximation at x. It is straightforward to see that $\frac{\|\pi_S(x) - e\|}{r_S(e)} \geq 1$.

2.2 The γ function and support functions

The condition number deduced from (2.18) relies on the vector x. In this section, by referring to support functions, we develop another characterization of the γ function, which leads to another condition number of γ in terms of the *centrality* of the reference point e.

Given a closed convex set $S \subset \mathcal{E}$, consider the set

$$\Omega_S := \{d \mid \|d\| = 1,\ \exists z \in \mathrm{bdy}(S) \text{ and } d \in \mathcal{N}_S(z)\},$$

i.e., the set of unit vectors in the normal cones to S.

For any $d \in \Omega_S$, define the support function

$$f_S(d) := \sup\{\langle x, d\rangle \mid x \in S\}.$$

We immediately see that for any $z \in \mathrm{bdy}(S)$ and $d \in \Omega_S$,

$$\langle z, d\rangle = f_S(d) \iff d \in \mathcal{N}_S(z). \tag{2.19}$$

Since $e \in \mathrm{int}(S)$, we also have

$$\langle e, d\rangle < f_S(d),\ \forall d \in \Omega_S. \tag{2.20}$$

Let us denote the supporting hyperplanes of S and the corresponding half-spaces by

$$H(S,d) := \{x \in \mathcal{E} \mid \langle x, d\rangle = f_S(d)\},\ H^-(S,d) := \{x \in \mathcal{E} \mid \langle x, d\rangle \le f_S(d)\}.$$

For any $d \in \Omega_S$, since $\|d\| = 1$, by (2.20), we get

$$\mathrm{dist}(e, H(S,d)) = |f_S(d) - \langle e, d\rangle| = f_S(d) - \langle e, d\rangle. \tag{2.21}$$

Since S is closed and convex, by a standard application of the Hahn-Banach Theorem, one can show that

$$S = \cap_{d\in\Omega_S}\{x \in \mathcal{E} \mid \langle x, d\rangle \le f_S(d)\} = \cap_{d\in\Omega_S} H^-(S,d). \tag{2.22}$$

Consequently, for any $x \notin S$ and $d \in \Omega_S$, we have $\mathrm{P}_S(x) \in S \subseteq H^-(S,d)$. Hence

$$\mathrm{dist}(x, H^-(S,d)) \le \|x - \mathrm{P}_S(x)\| = \mathrm{dist}(x, S).$$

On the other hand, $\operatorname{dist}\left(x, H^-\left(S, \frac{x-\mathrm{P}_S(x)}{\|x-\mathrm{P}_S(x)\|}\right)\right) = \|x - \mathrm{P}_S(x)\| = \operatorname{dist}(x, S)$. Thus we conclude that

Lemma 2.5. *When $x \notin S$, we have*

$$\operatorname{dist}(x, S) = \max_{d \in \Omega_S} \operatorname{dist}(x, H^-(S, d)).$$

We next show how the γ function can be written in terms of support functions:

Proposition 2.2. *For any $x \in \mathcal{E}$, we have*

$$\gamma(x) = \left(\max_{d \in \Omega_S} \left\{ \frac{\langle x - e, d\rangle}{f_S(d) - \langle e, d\rangle} \right\}\right)_+.$$

Proof. When $\gamma(x) = 0$ and $x - e \in \operatorname{recc}(S - e)$, we have $x(\lambda) = e + \lambda(x - e) \in S$ for all $\lambda > 0$. Hence for any $d \in \Omega_S$, according to the definition of the support function, we have

$$\langle e, d\rangle + \lambda\langle x - e, d\rangle = \langle x(\lambda), d\rangle \leq f_S(d) < \infty.$$

Let $\lambda \nearrow \infty$, we get $\langle x - e, d\rangle \leq 0 = \gamma(x)$.

Now consider $x \in \mathcal{E}$ such that $\gamma(x) > 0$. For any $d \in \Omega_S$, since $\pi_S(x) \in S$, by the definition of the support function, we have $\langle \pi_S(x) - e, d\rangle \leq f_S(d) - \langle e, d\rangle$. Also note that $f_S(d) - \langle e, d\rangle > 0$, we get

$$\gamma(x) \geq \gamma(x)\left(\frac{\langle \pi_S(x) - e, d\rangle}{f_S(d) - \langle e, d\rangle}\right) = \frac{\gamma(x)\langle \pi_S(x) - e, d\rangle}{f_S(d) - \langle e, d\rangle} = \frac{\langle x - e, d\rangle}{f_S(d) - \langle e, d\rangle},$$

where the last equality follows from

$$x - e = \gamma(x)(\pi_S(x) - e).$$

On the other hand, for any unit vector $d' \in \mathcal{N}_S(\pi_S(x))$, by (2.19), we have $\langle \pi_S(x), d'\rangle = f_S(d')$ and

$$\langle x - e, d'\rangle = \gamma(x)\langle \pi_S(x) - e, d'\rangle = \gamma(x)(f_S(d') - \langle e, d'\rangle).$$

Hence

$$\gamma(x) = \max_{d\in\Omega_S} \left\{ \frac{\langle x-e, d\rangle}{f_S(d) - \langle e, d\rangle} \right\},$$

and the maximum is attainable. □

Corollary 2.1. *When $x \notin S$, we have*

$$\gamma(x) - 1 = \max_{d\in\Omega_S} \left\{ \frac{\operatorname{dist}(x, H^-(S,d))}{\operatorname{dist}(e, H(S,d))} \right\}.$$

Proof. If $x \notin S$, then $\gamma(x) > 1$. By Proposition 2.2, we get

$$\begin{aligned}
\gamma(x) &= \max_{d\in\Omega_S} \left\{ \frac{\langle x-e, d\rangle}{f_S(d) - \langle e, d\rangle} \right\} \\
&= \max_{d\in\Omega_S} \left\{ 1 + \frac{\langle x, d\rangle - f_S(d)}{f_S(d) - \langle e, d\rangle} \right\} \\
&= 1 + \max_{d\in\Omega_S} \left\{ \frac{\langle x, d\rangle - f_S(d)}{f_S(d) - \langle e, d\rangle} \right\}.
\end{aligned}$$

Suppose $d' \in \Omega_S$ obtains the maximum here. Since $\gamma(x) > 1$, we see that

$$0 < \langle x, d'\rangle - f_S(d') = \operatorname{dist}(x, H^-(S,d')).$$

The final step of the proof follows immediately from (2.21). □

Given a reference point $e \in \operatorname{int}(S)$, define

$$g_S(e) := \min_{d\in\Omega_S} \operatorname{dist}\left(e, H(S,d)\right),^2 \; h_S(e) := \sup_{d\in\Omega_S} \operatorname{dist}\left(e, H(S,d)\right).$$

When $x \notin S$, recall that Lemma 2.5 states $\operatorname{dist}(x,S) = \max_{d\in\Omega_S} \operatorname{dist}(x, H^-(S,d))$. Compared with Corollary 2.1, we immediately see that

$$\frac{1}{h_S(e)} \le \frac{\gamma(x)-1}{\operatorname{dist}(x,S)} \le \frac{1}{g_S(e)}. \tag{2.23}$$

Let d' be any unit vector in $\mathcal{N}_S(\pi_S(x))$, then $\langle \pi_S(x) - e, d'\rangle > 0$. Noting that the three points $e, \pi_S(x), x$ are on the same line, we have

$$\gamma(x) - 1 = \frac{\|x - \pi_S(x)\|}{\|\pi_S(x) - e\|} = \frac{\langle x - \pi_S(x), d'\rangle}{\langle \pi_S(x) - e, d'\rangle} = \frac{\operatorname{dist}(x, H^-(S,d'))}{\operatorname{dist}(e, H(S,d'))}. \tag{2.24}$$

[2] In Lemma 8.1, we show that the minimum here is obtainable.

Hence either side of (2.23) is tight only when $\pi_S(x) = \mathrm{P}_S(x)$ and $\tilde{d} = \frac{x - \mathrm{P}_S(x)}{\|x - \mathrm{P}_S(x)\|} \in \mathcal{N}_S(\pi_S(x))$ achieves the extreme values in $h_S(e)$ or $g_S(e)$.

By (2.23), the *centrality* of e with respect to S, defined as

$$\tau_{(S,e)} := \frac{h_S(e)}{g_S(e)},$$

also provides a condition number of γ with respect to S. The definition of centrality implies that $\tau_{(S,e)} \geq 1$. When $\tau_{(S,e)}$ is small, $(\gamma(x) - 1)_+$ is a good proxy for the distance function.

Sandwich inequalities (2.18) and (2.23) both provide condition numbers of $(\gamma(x) - 1)_+$ as an approximation of a scalar multiple of the distance function $\mathrm{dist}(x, S)$. It should be noted that both inequalities share the same upper bound (i.e., $r_S(e) = g_S(e)$).[3] While the lower bound in (2.23) depends on x, the lower bound in (2.23) only involves e. When additional information regarding the positions of the points of interest is available (for instance, iterates of the algorithms introduced in Chapter 5), (2.18) could be very useful. In other cases where $\tau_{(S,e)}$ is bounded (e.g., when S is polyhedral), (2.23) could be more helpful, since it does not depend on the position of x. The following example is a case where (2.23) leads to a better bound than (2.18):

Example 2.1. Consider the set

$$S = \{(x, y) \mid xy \geq 1,\ x > 0\}$$

and a reference point $e = (2, 2)$. We can show that $g_S(e) = \sqrt{2}$ (obtained when the normal vector $d = (-\sqrt{1/2}, -\sqrt{1/2})$) and $h_S(e) = 2$ (obtained when the normal vectors tend to (0, -1) or (-1, 0)). Hence $\tau_{(S,e)} = \sqrt{2}$.

[3] Again, see Lemma 8.1.

In contrast, consider the points $\{x_n\}_{n\in\mathbb{N}}$ where $x_n = (n, 0)$. As $n \to \infty$, we have $\|\pi_S(x_n) - e\| \to \infty$, and the left-hand side of (2.18) tends to 0.

CHAPTER 3

DEFINITION AND PROPERTIES OF THE γ FUNCTION FOR MULTIPLE SETS

Now consider a finite number of closed convex sets $\{S_1, \ldots, S_m\} \subseteq \mathcal{E}$, where $m > 1$. For any S_i, assume a reference point $e_i \in \text{int}(S_i)$ is given. Let $\gamma_i : \mathcal{E} \to [0, \infty)$ be the γ function defined with respect to S_i and e_i. Define

$$S_0 := \cap_{i \in [m]} S_i \tag{3.1}$$

and

$$\gamma_0(x) := \max_{i \in [m]} \gamma_i(x). \tag{3.2}$$

Since $\gamma_i : \mathcal{E} \to [0, \infty)$ is convex for all $i \in [m]$, we see that $\gamma_0 : \mathcal{E} \to [0, \infty)$ is also convex.

In this chapter, we assume

$$\text{int}(S_0) \neq \emptyset. \tag{3.3}$$

By Lemma 2.1, this implies

$$\gamma_0^\star := \inf_{x \in \mathcal{E}} \gamma_0(x) < 1.$$

Recall that by (2.18) and (2.23), for any $i \in [m]$, $\gamma_i(x)$ and $\text{dist}(x, S_i)$ are closely connected via e_i. We would like to establish a similar sandwich inequality between $\gamma_0(x)$ and $\text{dist}(x, S_0)$. In order to do this, let us take a brief detour and dig into the relationship between $\text{dist}(x, S_0)$ and $\max_{i \in [m]} \text{dist}(x, S_i)$.

3.1 Linear regularity for closed convex sets

For a closed convex cone $K \neq \mathcal{E}$ with nonempty interior, the *inradius* of K[10] is

$$\begin{aligned} \operatorname{inrad}(K) &:= \max\{r_K(v) \mid v \in K \text{ and } \|v\| = 1\} \\ &= \max\left\{ \frac{r_K(v)}{\|v\|} \middle| v \in K \text{ and } \|v\| \neq 0 \right\}, \end{aligned} \tag{3.4}$$

i.e., the largest radius of balls which are subsets of K and centered at unit vectors. By the definition of $\operatorname{inrad}(K)$, we immediately have

$$\operatorname{inrad}(K) \leq 1.$$

The quantity defined in (3.4) is also referred to as the "width" of the cone K in the perceptron literature [8, 20]. Here we adopt the name inradius because our analysis in the following chapters often involve lower bounds on the radii of balls.

We rely on a related notion for the intersection $K \cap \mathcal{L}$, where $\mathcal{L}$ is a subspace intersecting the interior of K:

$$\operatorname{inrad}_{\mathcal{L}}(K) := \max\{r_K(v) \mid v \in K \cap \mathcal{L} \text{ and } \|v\| = 1\}.$$

Although the vectors v in the definition of $\operatorname{inrad}_{\mathcal{L}}(K)$ are restricted to the subspace, still the radius refers to balls that are full-dimensional and contained in K. Hence we have

$$\operatorname{inrad}_{\mathcal{L}}(K) \leq \operatorname{inrad}(K) \leq 1. \tag{3.5}$$

For $z \in S_0$, the tangent cone for S_0 at z is the closed convex cone

$$\mathcal{T}_{S_0}(z) = \operatorname{cl}\{t(x - z) \mid x \in S_0 \text{ and } t > 0\},$$

where "cl" denotes closure. The tangent cone is the smallest closed convex cone K for which $S_0 \subseteq z + K$. We have the following characterization of $\operatorname{inrad}_{\mathcal{L}}(\mathcal{T}_{S_0}(z))$:

Lemma 3.1. *For any $z \in \mathrm{bdy}(S_0)$ such that $\mathrm{int}(S_0) \cap (z + \mathcal{L}) \neq \emptyset$,*

$$\mathrm{inrad}_{\mathcal{L}}(\mathcal{T}_{S_0}(z)) = \sup \left\{ \frac{r_{S_0}(w)}{\|w - z\|} \mid w \in \mathrm{int}(S_0) \cap (z + \mathcal{L}) \right\}. \tag{3.6}$$

In particular, when $\mathcal{L} = \mathcal{E}$, we get

$$\mathrm{inrad}(\mathcal{T}_{S_0}(z)) = \sup \left\{ \frac{r_{S_0}(w)}{\|w - z\|} \mid w \in \mathrm{int}(S_0) \right\}. \tag{3.7}$$

Remark. See Example A.1 in Appendix A.2 for an instance where the supremum on the right-hand side of Lemma 3.1 is not obtainable.

Proof. By the definition of $\mathcal{T}_{S_0}(z))$, we have $S \subseteq z + \mathcal{T}_{S_0}(z))$. Hence for any $w \in \mathrm{int}(S_0) \cap (z + \mathcal{L})$,

$$\mathrm{inrad}(\mathcal{T}_{S_0}(z)) \geq \frac{r_{S_0}(w)}{\|w - z\|}.$$

On the other hand, note that for any compact convex set T,

$$T \subset \mathrm{int}(\mathcal{T}_{S_0}(z)) \Longrightarrow \exists t' > 0 \text{ for which } z + t' \cdot T \subset S_0.^1 \tag{3.8}$$

Now assume $v \in \mathcal{T}_{S_0}(z)$ satisfies $\|v\| = 1$ and $r_{\mathcal{T}_{S_0}(z)}(v) = \mathrm{inrad}(\mathcal{T}_{S_0}(z))$. Then for any $r < \mathrm{inrad}(\mathcal{T}_{S_0}(z))$, we have

$$\overline{B(v, r)} \subset \mathrm{int}(\mathcal{T}_{S_0}(z)).$$

By (3.8), there exists $t' > 0$ such that

$$z + t'(\overline{B(v, r)}) = \overline{B(z + t'v, t'r)} \subset S_0$$

and

$$\frac{r_{S_0}(z + t'v)}{\|(z + t'v) - z\|} = \frac{r_{S_0}(z + t'v)}{\|t'v\|} \geq \frac{\|t'r\|}{\|t'v\|} = r.$$

Let $r \nearrow \mathrm{inrad}(\mathcal{T}_{S_0}(z))$, we get

$$\mathrm{inrad}(\mathcal{T}_{S_0}(z)) \leq \sup \left\{ \frac{r_{S_0}(w)}{\|w - z\|} \mid w \in \mathrm{int}(S_0) \right\}. \tag{3.9}$$

□

[1] See the proof in Appendix A.2.

The purpose of this part is to establish and discuss the following theorem. The theorem is significant in its own regard, as we clarify following the proof. On a first reading, we recommend assuming $L = \mathcal{E}$, in which case "$S_0 \cap (x + L)$" becomes simply "S_0", and "$\mathrm{dist}_{\mathcal{L}}$" and "$\mathrm{inrad}_{\mathcal{L}}$" become "dist" and "inrad", respectively.

Theorem 3.1. *Assume $x \notin S_0$ and $\mathrm{int}(S_0) \cap (x + \mathcal{L}) \neq \emptyset$, where $\mathcal{L}$ is a subspace (possibly $\mathcal{L} = \mathcal{E}$). Let $z = \mathrm{P}_{S_0 \cap (x+\mathcal{L})}(x)$ and $\mathcal{I} := \{i \in [m] \mid z \in \mathrm{bdy}(S_i)\}$. Then*

$$\mathrm{dist}_{\mathcal{L}}(x, S_0) := \mathrm{dist}(x, S_0 \cap (x + \mathcal{L})) = \|x - z\| \leq \frac{\max_{i \in \mathcal{I}} \mathrm{dist}(x, S_i)}{\mathrm{inrad}_{\mathcal{L}}(\mathcal{T}_{S_0}(z))}. \tag{3.10}$$

Remark. We make two remarks to facilitate readers' understanding of Theorem 3.1:

1. To understand that (3.10) cannot in general be strengthened, assume $\mathcal{L} = \mathcal{E}$ (then $\mathrm{inrad}_{\mathcal{L}} \mathcal{T}_{S_0}(z) = \mathrm{inrad}(\mathcal{T}_{S_0}(z))$), and $\mathrm{bdy}(S_0)$ is a smooth manifold in a neighborhood of z, in which case $\mathcal{T}_{S_0}(z)$ is a half-space so $\mathrm{inrad}(\mathcal{T}_{S_0}(z)) = 1$. Consequently (3.10) gives $\mathrm{dist}(x, S_0) \leq \max_{i \in \mathcal{I}} \mathrm{dist}(x, S_i)$. Since we always have $\mathrm{dist}(x, S_0) \geq \max_{i \in [m]} \mathrm{dist}(x, S_i)$, it follows that

   $$\mathrm{dist}(x, S_0) = \max_{i \in \mathcal{I}} \mathrm{dist}(x, S_i) = \max_{i \in [m]} \mathrm{dist}(x, S_i),$$

   a bound on $\mathrm{dist}(x, S_0)$ which is impossible to improve. See Example A.2 in Appendix A.2 for another example.

2. The reason for introducing a subspace $\mathcal{L}$ is that we allow convex feasibility problems which include linear equations, $Ax = b$, in which case $\mathcal{L}$ will be the null-space of A. The iterates of our algorithms will satisfy the linear equations. Consequently, in analyzing improvements over the course of iterations, we need "$\mathrm{dist}_{\mathcal{L}}$" on the left of (3.10), not "dist". However, to avoid making the strong assumption that the points e_i are solutions to the linear equations in addition to being interior points of the sets S_i, on the right we need "dist" rather than "$\mathrm{dist}_{\mathcal{L}}$".

Proof. Consider any $w \in \mathrm{int}(S) \cap (x + \mathcal{L})$. Since z is the orthogonal projection of x onto $S_0 \cap (x + \mathcal{L})$, we have

$$x - z \in \mathcal{N}_{S_0 \cap (x+\mathcal{L})}(z) = \mathrm{cl}\left(\sum_{i\in\mathcal{I}} \mathcal{N}_{S_i}(z) + \mathcal{L}^\perp\right) = \sum_{i\in\mathcal{I}} \mathcal{N}_{S_i}(z) + \mathcal{L}^\perp.$$

The last equality is a consequence to the assumption $\mathrm{int}(S_0) \cap (x + \mathcal{L}) \neq \emptyset$ (c.f., Theorem 6.42 in [27]). Hence we can write

$$x - z = \sum_{i\in\mathcal{I}} \lambda_i d_i + v, \text{ where } \lambda_i \geq 0,\ d_i \in \mathcal{N}_{S_i}(z),\ \|d_i\| = 1,\ v \in L^\perp. \tag{3.11}$$

Since $\overline{B(w, r_{S_0}(w))} \subset S_0$, we have $w + r_{S_0}(w)d_i \in S_i$ for all $i \in \mathcal{I}$. Thus, because $d_i \in \mathcal{N}_{S_i}(z)$ and $\|d_i\| = 1$, we find

$$\langle z, d_i\rangle \geq \langle w + r_{S_0}(w)d_i, d_i\rangle = \langle w, d_i\rangle + r_{S_0}(w).$$

That is, $\langle d_i, z - w\rangle \geq r_{S_0}(w)$. Consequently,

$$\langle x - z, z - w\rangle = \left\langle \sum_{i\in\mathcal{I}} \lambda_i d_i + v, z - w \right\rangle = \sum_{i\in\mathcal{I}} \lambda_i \langle d_i, z - w\rangle \geq r_{S_0}(w) \sum_{i\in\mathcal{I}} \lambda_i,$$

where we have made use of $z - w \in L$ and $v \in L^\perp$. Hence, by Cauchy-Schwartz,

$$\|x - z\| \geq \left(\sum_{i\in\mathcal{I}} \lambda_i\right) \frac{r_{S_0}(w)}{\|w - z\|}. \tag{3.12}$$

On the other hand, note that $x - z \in L$ and $v \in L^\perp$,

$$\langle x-z, x-z\rangle = \left\langle \sum_{i\in\mathcal{I}} \lambda_i d_i + v, x - z \right\rangle = \sum_{i\in\mathcal{I}} \lambda_i \langle d_i, x-z\rangle \leq \left(\sum_{i\in\mathcal{I}} \lambda_i\right) \max_{i\in\mathcal{I}} \langle d_i, x-z\rangle. \tag{3.13}$$

For any $i \in [m]$, note that $\|d_i\| = 1$ and $d_i \in \mathcal{N}_{S_i}(z)$, by Lemma 2.5, we have

$$\langle x - z, d_i\rangle \leq (\langle x - z, d_i\rangle)_+ = \mathrm{dist}(x, H^-(S_i, d_i)) \leq \mathrm{dist}(x, S_i).$$

Substituting into (3.13) yields

$$\|x - z\|^2 \leq \left(\sum_{i\in\mathcal{I}} \lambda_i\right) \max_{i\in\mathcal{I}} \mathrm{dist}(x, S_i). \tag{3.14}$$

Combining the lower bound on $\|x - z\|$ given by (3.12) and and the upper bound on $\|x - z\|^2$ given by (3.14), for any $w \in \text{int}(S_0) \cap (x + \mathcal{L})$, we get

$$\|x - z\| = \frac{\|x - z\|^2}{\|x - z\|} \leq \frac{\left(\sum_{i \in \mathcal{I}} \lambda_i\right) \max_{i \in \mathcal{I}} \text{dist}(x, S_i)}{\left(\sum_{i \in \mathcal{I}} \lambda_i\right) \frac{r_{S_0}(w)}{\|w - z\|}} = \frac{\max_{i \in \mathcal{I}} \text{dist}(x, S_i)}{\frac{r_{S_0}(w)}{\|w - z\|}}.$$

Note that $z \in (x + \mathcal{L})$, we get $z + \mathcal{L} = x + \mathcal{L}$. Taking the supremum of $\frac{r_{S_0}(w)}{\|w - z\|}$ over all $w \in \text{int}(S) \cap (x + \mathcal{L})$, by Lemma 3.1, we see that

$$\|x - z\| \leq \frac{\max_{i \in \mathcal{I}} \text{dist}(x, S_i)}{\sup\left\{\frac{r_{S_0}(w)}{\|w - z\|} \mid w \in \text{int}(S_0) \cap (x + \mathcal{L})\right\}} = \frac{\max_{i \in \mathcal{I}} \text{dist}(x, S_i)}{\text{inrad}_{\mathcal{L}}(\mathcal{T}_{S_0}(z))}.$$

□

Given a convex feasibility problem of finding (or approximating) a point in $S_0 = \cap_{i \in [m]} S_i$, let $\kappa(x) > 0$ satisfy

$$\text{dist}(x, S_0) = \kappa(x) \max_{i \in [m]} \text{dist}(x, S_i) \tag{3.15}$$

for all $x \notin S$. If $\sup_{x \notin S} \kappa(x) < \infty$, then the problem is *linearly regular* [3]. Otherwise, if $\kappa(x)$ is upper bounded for x restricted to any bounded regions in $\mathcal{E}$, we say that the convex feasibility problem exhibits *bounded linear regularity*.

(Bounded) linear regularity is usually essential to establishing the linear convergence rate of projection algorithms for convex feasibility problems (see, e.g., [1, 2, 4, 10, 11]). Hoffman's lemma [13] shows the linear regularity of the feasibility problem where $S_0 \neq \emptyset$ and S_i's are half spaces.

As for convex feasibility problems involving generic closed convex sets, fix $x \notin S_0$ and let $z = \text{P}_{S_0}(x)$. Assume $\text{int}(S_0) \neq \emptyset$, and pick any $w \in \text{int}(S_0)$. In Theorem 7 of [3], Bauschke, Borwein and Li showed $\frac{\|z - w\|}{r_{S_0}(w)} \sum_{i \in [m]} \text{dist}(x, S_i) \geq \text{dist}(x, S_0)$, which implies $\kappa(x) \leq \frac{m}{r_{S_0}(w)/\|z - w\|}$. Similar results involving m, the number of sets considered, have been discovered by Beck and Teboulle [4] (where

$\kappa(x) \leq \frac{2\sqrt{m}+1/\|x-w\|}{r_{S_0}(w)/\|x-w\|}$) and Bolte et al. [6] (where $\kappa(x) \leq (\frac{2+1/\|x-w\|}{r_{S_0}(w)/\|z-w\|})^{m-1}$). In [19], Nedić's analysis led to $\kappa(x) \leq \frac{\|x-w\|}{r_{S_0}(w)}$. All these results imply bounded linear regularity.

In the proof of Theorem 3.1, we essentially showed

$$\kappa(x) \leq \frac{\|z-w\|}{r_{S_0}(w)}, \tag{3.16}$$

which is independent of m.

When S_i's are half-spaces and S_0 is a polyhedral cone, let x be a vector in the polar cone of S_0 (i.e., $\langle x, w\rangle \leq 0$ for all $w \in S_0$). Then $z = \vec{0}$, and Theorem 3.1 recovers Theorem 3.5.2 in [10] by Goffin, which was derived with proof techniques similar to those used for Theorem 3.1.

3.2 First-order properties of γ function for multiple sets

We now study the properties of γ_0. Recall that $\gamma_0(x) = \max_{i\in[m]} \gamma_i(x)$. By Proposition 2.1, we immediately obtain the following characterization of the subgradients of γ_0:

Lemma 3.2. *Given $x \in \mathcal{E}$, when $\gamma_0(x) > 0$,*

$$\begin{aligned}\partial\gamma_0(x) &= \operatorname{conv}(\{d \in \partial\gamma_i(x) \mid \gamma_i(x) = \gamma_0(x)\}) \\ &= \operatorname{conv}(\{d \in \mathcal{E} \mid \gamma_i(x) = \gamma_0(x),\ d \in \mathcal{N}_{S_i}(\pi_{S_i}(x)) \text{ and } \langle \pi_{S_i}(x) - e_i, d\rangle = 1\}).\end{aligned}$$

Let

$$r_0 := \min_{i\in[m]} r_i,\ r_i := r_{S_i}(e_i),\ \forall i \in [m].$$

By Lemma 3.2 and Lemma 2.3, we get

Lemma 3.3. *Given $x \in \mathcal{E}$ such that $\gamma_0(x) > 0$,*

$$\partial\gamma_0(x) \subset \overline{B(0, 1/r_0)}.$$

Consequently, the function γ_0 is $1/r_0$-Lipschitz.

When $\mathrm{int}(S_0) \neq \emptyset$ and $x \notin S_0$, let $z = \mathrm{P}_{S_0}(x)$. By taking $\mathcal{L} = \mathcal{E}$ in Theorem 3.1, with the growth rate of $\gamma_i(x) - 1$ in Lemma 2.4, we obtain the following growth rate of $\gamma_0(x) - 1$ with respect to $\mathrm{dist}(x, S_0)$:

$$\begin{aligned}
\frac{\gamma_0(x) - 1}{\mathrm{dist}(x, S_0)} &= \frac{\max_{\gamma_i(x)>1}(\gamma_i(x) - 1)}{\mathrm{dist}(x, S_0)} \\
&\geq \frac{\max_{\gamma_i(x)>1} \left(\mathrm{dist}(x, S_i)/\|\pi_{S_i}(x) - e_i\|\right)}{\mathrm{dist}(x, S_0)} \\
&\geq \max_{i\in[m]} \left\{ \frac{\mathrm{dist}(x, S_i)}{\mathrm{dist}(x, S_0)} \right\} \frac{1}{\max_{\gamma_i(x)>1} \|\pi_{S_i}(x) - e_i\|} \\
&\geq \frac{\mathrm{inrad}(\mathcal{T}_{S_0}(z))}{\max_{\gamma_i(x)>1} \|\pi_{S_i}(x) - e_i\|}.
\end{aligned}$$

By (2.17), we get

Proposition 3.1. *Assume $S_0 \neq \emptyset$ and $x \notin S_0$. Let $z = \mathrm{P}_{S_0}(x)$. We have*

$$\begin{aligned}
\frac{\gamma_0(x) - 1}{\mathrm{dist}(x, S_0)} &\geq \frac{\mathrm{inrad}(\mathcal{T}_{S_0}(z))}{\max_{\gamma_i(x)>1} \|\pi_{S_i}(x) - e\|} \\
&\geq \frac{\mathrm{inrad}(\mathcal{T}_{S_0}(z))}{\max_{\gamma_i(x)>1} \|\mathrm{P}_{S_i}(x) - e\|} \\
&\geq \frac{\mathrm{inrad}(\mathcal{T}_{S_0}(z))}{\max_{\gamma_i(x)>1} \|x - e_i\|}.
\end{aligned}$$

Given a finite number of closed convex sets and the corresponding reference points, define the centrality of the reference points by

$$\tau_{\{(S_i,e_i)\}_{i\in[m]}} := \frac{\max_{i\in[m]} h_{S_i}(e_i)}{\min_{i\in[m]} g_{S_i}(e_i)}. \tag{3.17}$$

Then we get

$$\max_{i\in[m]} \tau_{(S_i,e_i)} = \max_{i\in[m]} \left\{ \frac{h_{S_i}(e)}{g_{S_i}(e)} \right\} \leq \frac{\max_{i\in[m]} h_{S_i}(e_i)}{\min_{i\in[m]} g_{S_i}(e_i)} = \tau_{\{(S_i,e_i)\}_{i\in[m]}}. \tag{3.18}$$

For any $x \notin S_0 \neq \emptyset$, let $z = \mathrm{P}_{S_0}(x)$. Since $r_i = g_{S_i}(e_i)$,[2] we can combine Lemma 3.3 and Proposition 3.1 to get

$$\frac{\mathrm{inrad}(\mathcal{T}_{S_0}(z))}{\max_{i\in[m]} h_{S_i}(e_i)} \leq \frac{\gamma_0(x)-1}{\mathrm{dist}(x,S_0)} \leq \frac{1}{\min_{i\in[m]} g_{S_i}(e_i)} = \frac{1}{r_0}. \tag{3.19}$$

Hence a *condition number of* γ_0 (with respect to S_0) is

$$\mathrm{inrad}(\mathcal{T}_{S_0}(z))^{-1}\tau_{\{(S_i,e_i)\}_{i\in[m]}}. \tag{3.20}$$

Consequently, for generic $x \notin S_0$, we see that the condition number of γ_0 is determined by the geometry of S_0 and the centrality of $\{e_i\}_{i\in[m]}$ with respect to $\{S_i\}_{i\in[m]}$.

We close this chapter by showing the connection between $r_{S_0}(x)$ and $\gamma_0(x)$:

Lemma 3.4. *When $x \in S_0$, then $\gamma_0(x) \leq 1$ and we have*

$$r_{S_0}(x) \geq (1-\gamma_0(x))r_0.$$

Proof. For any $i \in [m]$, if $\gamma_i(x) = 0$, then by (2.4) and Lemma A.1 in Appendix A.1, we have

$$\overline{B(x,r_i)} = \overline{B(e,r_i)} + (x-e) \subset S_i.$$

When $\gamma_i(x) > 0$, by (2.7), we get $\pi_{S_i}(x) \in S_i$. Since $x \in S_0 \subseteq S_i$, we have $\gamma_i(x) \leq 1$ and $x \in (e, \pi_{S_i}(x)]$. Hence

$$\overline{B(x,(1-\gamma_i(x))r_i)} = (1-\gamma_i(x))B(e,r_i) + \gamma_i(x)\pi_{S_i}(x) \subset S_i.$$

Consequently, $r_{S_i}(x) \geq (1-\gamma_i(x))r_i$. The rest of the proof follows from the definition of γ_0 and r_0. □

[2] See Lemma 8.1.

CHAPTER 4

THE γ FUNCTION ON AN AFFINE SUBSPACE

In Section 3.1, we developed a linear regularity result involving $\text{dist}_{\mathcal{L}}$ and $\text{inrad}_{\mathcal{L}}$. As alluded to in the remarks following Theorem 3.1, in this chapter, we study the properties of the γ function when restricted to an affine subspace of $\mathcal{E}$. This allows us to consider convex feasibility problems including linear equations, $Ax = b$. Let $\mathcal{L}$ denote the null-space of A, and assume $u \in \mathcal{E}$ satisfies $Au = b$. Then the solutions to the linear equations are $u + \mathcal{L}$.

Given a collection of closed convex sets $\{S_i\}_{i\in[m]}$, if our reference points $\{e_i\}_{i\in[m]}$ satisfy $\{e_i\}_{i\in[S]} \subset (u + \mathcal{L})$, then the previous results are immediately applicable by replacing the full-dimensional space $\mathcal{E}$ with $u + \mathcal{L}$. Consequently, the purpose of this chapter is to explore the properties of the γ function without assuming $\{e_i\}_{i\in[m]} \subset (u + \mathcal{L})$.

Given a closed convex S and $e \in \text{int}(S)$, let

$$r_{(S,\mathcal{L})}(e) := \max\{r > 0 \mid (\overline{B(e, r)} \cap (e + \mathcal{L})) \subseteq (S \cap (e + \mathcal{L})\}.$$

By the definition of $r_S(e)$, we have

$$\overline{B(e, r_S(e))} \cap (e + \mathcal{L}) \subseteq S \cap (e + \mathcal{L}).$$

Hence $r_{(S,\mathcal{L})}(e) \geq r_S(e)$. The γ function defined with respect to S and e has the following Lipschitz continuity when restricted to affine subspaces:

Lemma 4.1. *For any $x, y \in \mathcal{E}$ such that $x - y \in \mathcal{L}$, we have*

$$|\gamma(x) - \gamma(y)| \leq \frac{\|x - y\|}{r_{(S,\mathcal{L})}(e)}.$$

Proof. The result holds trivially when $\gamma(x) = \gamma(y) = 0$, so let us assume $\gamma(x) > 0$.

First assume $r_{(S,\mathcal{L})}(e) = \infty$. Then $(e + \mathcal{L}) \subseteq S$. Consequently,

$$y - x \in \mathcal{L} \subseteq \operatorname{recc}(S - e) = \operatorname{recc}(S - \pi_S(x)),$$

where the equality follows from Lemma A.1 in Appendix A.1. Hence $\pi_S(x) + (y - x)/\gamma(x) \in S$, and

$$y\left(\frac{1}{\gamma(x)}\right) = e + \frac{y-e}{\gamma(x)} = \left(e + \frac{x-e}{\gamma(x)}\right) + \frac{y-x}{\gamma(x)} = \pi_S(x) + \frac{y-x}{\gamma(x)} \in S.$$

By (2.5) and Lemma 2.1, we get

$$\frac{\gamma(y)}{\gamma(x)} = \gamma\left(y\left(\frac{1}{\gamma(x)}\right)\right) \leq 1. \tag{4.1}$$

Now note that $(e + \mathcal{L}) \subseteq S$ implies $\mathcal{L} \subseteq \operatorname{recc}(S - e)$. If $\gamma(y) = 0$, then

$$x - e = (x - y) + (y - e) \in \mathcal{L} + \operatorname{recc}(S - e) = \operatorname{recc}(S - e),$$

which implies $\gamma(x) = 0$, a contradiction. Hence $\gamma(y) > 0$, and we can show $\gamma(x) \leq \gamma(y)$ by a argument similar to that of (4.1).

Now assume $r_{(S,\mathcal{L})}(e) < \infty$. For any $g \in \partial\gamma(x)$, we have $\mathrm{P}_{\mathcal{L}}(g) \in \mathcal{L}$. According to the definition of $r_{(S,\mathcal{L})}(e)$,

$$e + \frac{r_{(S,\mathcal{L})}(e)}{\|\mathrm{P}_{\mathcal{L}}(g)\|}\mathrm{P}_{\mathcal{L}}(g) \in S \cap (e + \mathcal{L}) \subset S.$$

Recall that Proposition 2.1 implies $\langle \pi_S(x) - e, g\rangle = 1$ and $g \in \mathcal{N}_S(\pi_S(x))$. Hence

$$\begin{aligned}
\langle e, g\rangle + 1 &= \langle \pi_S(x), g\rangle \\
&\geq \left\langle e + \frac{r_{(S,\mathcal{L})}(e)}{\|\mathrm{P}_{\mathcal{L}}(g)\|}\mathrm{P}_{\mathcal{L}}(g), g\right\rangle \\
&= \langle e, g\rangle + \frac{r_{(S,\mathcal{L})}(e)}{\|\mathrm{P}_{\mathcal{L}}(g)\|}\langle \mathrm{P}_{\mathcal{L}}(g), g\rangle \\
&= \langle e, g\rangle + r_{(S,\mathcal{L})}(e)\|\mathrm{P}_{\mathcal{L}}(g)\|.
\end{aligned}$$

Thus $\|\mathrm{P}_{\mathcal{L}}(g)\| \leq 1/r_{(S,\mathcal{L})}(e)$, and the rest of the proof is similar to that of Lemma 2.3. □

In (2.23), we established the growth rate of $\gamma(x)$ with respect to $\mathrm{dist}(x, S)$. When the convex feasibility problem of interest includes linear equations, $\gamma(x)$ has the following growth rate with respect to $\mathrm{dist}_{\mathcal{L}}(x, S)$:

Lemma 4.2. *Assume $x \notin S$ and $S \cap (x + \mathcal{L}) \neq \emptyset$. Let $z = \mathrm{P}_{S\cap(x+\mathcal{L})}(x)$. Then we have*

$$\frac{\gamma(x) - 1}{\mathrm{dist}_{\mathcal{L}}(x, S)} \geq \frac{\mathrm{inrad}_{\mathcal{L}}(\mathcal{T}_S(z))}{\|\pi_S(x) - e\|}.$$

Remark. When restricted to an affine subspace, both the Lipschitz constant (Lemma 4.1) and growth rate (Lemma 4.2) become smaller than the corresponding results for the full space $\mathcal{E}$.

Proof. By Lemma 2.4,

$$\frac{\gamma(x) - 1}{\mathrm{dist}_{\mathcal{L}}(x, S)} = \frac{\gamma(x) - 1}{\mathrm{dist}(x, S)} \frac{\mathrm{dist}(x, S)}{\mathrm{dist}_{\mathcal{L}}(x, S)} \geq \frac{1}{\|\pi_S(x) - e\|} \frac{\mathrm{dist}(x, S)}{\mathrm{dist}_{\mathcal{L}}(x, S)}.$$

By Theorem 3.1, we have

$$\frac{\mathrm{dist}(x, S)}{\mathrm{dist}_{\mathcal{L}}(x, S)} \geq \mathrm{inrad}_{\mathcal{L}}(\mathcal{T}_S(z)),$$

and the statement is true. □

Now consider a finite number of closed convex sets $\{S_1, \ldots, S_m\} \subseteq \mathcal{E}$, and $S_0 = \cap_{i\in[m]} S_i$ satisfying $S_0 \neq \emptyset$. Following arguments similar to those in Section 3.2,[1] we can show the Lipschitzness and growth rate of γ_0 when restricted to an affine subspace. It is worth noting that both bounds become smaller when restricted to affine subspaces.

[1] See Appendix A.3 for the proofs.

Lemma 4.3. *For any $x, y \in \mathcal{E}$ such that $x - y \in \mathcal{L}$, we have*

$$\frac{|\gamma_0(x) - \gamma_0(y)|}{\|x - y\|} \leq \frac{1}{\min_{i\in[m]} r_{(S_i,\mathcal{L})}(e_i)}.$$

Lemma 4.4. *Assume $x \notin S_0$ and $S_0 \cap (x + \mathcal{L}) \neq \emptyset$. Let $z = \mathrm{P}_{S_0\cap(x+\mathcal{L})}(x)$. Then we have*

$$\begin{aligned}\frac{\gamma_0(x) - 1}{\mathrm{dist}_{\mathcal{L}}(x, S_0)} &\geq \frac{\mathrm{inrad}_{\mathcal{L}}(\mathcal{T}_{S_0}(z))}{\max_{\gamma_i(x)>1} \|\pi_{S_i}(x) - e_i\|} \\ &\geq \frac{\mathrm{inrad}_{\mathcal{L}}(\mathcal{T}_{S_0}(z))}{\max_{\gamma_i(x)>1} \|\mathrm{P}_{S_i}(x) - e_i\|} \\ &\geq \frac{\mathrm{inrad}_{\mathcal{L}}(\mathcal{T}_{S_0}(z))}{\max_{\gamma_i(x)>1} \|x - e_i\|}.\end{aligned}$$

CHAPTER 5

DETERMINISTIC ALGORITHMS FOR THE CONVEX FEASIBILITY PROBLEM

In this chapter, we discuss two radial projection-based deterministic algorithms for *the convex feasibility problem* (approximating the set S_0).

For any closed convex set $S \subset \mathcal{E}$, when $x \notin S$, the vector $x - \mathrm{P}_S(x)$ plays a dual role in a first-order oracle of the distance function $\mathrm{dist}(x, S)$:

- $\mathrm{dist}(x, S) = \|x - \mathrm{P}_S(x)\|$;
- $\frac{x-\mathrm{P}_S(x)}{\|x-\mathrm{P}_S(x)\|} \in \partial_x \mathrm{dist}(x, S)$.

Consequently, the orthogonal projection $x \mapsto \mathrm{P}_S(x)$ can be seen as a subgradient update of the distance function, with the step size being $\mathrm{dist}(x, S)$.

By Lemma 3.3 and Proposition 3.1, for any $x \notin S_0$, we have

$$\frac{\mathrm{inrad}\, \mathcal{T}_{S_0}(z)}{\max_{\gamma_i(x_k)>1} \|\pi_{S_i}(x) - e_i\|} \leq \frac{\gamma_0(x) - 1}{\mathrm{dist}(x, S_0)} \leq \frac{1}{r_0}, \tag{5.1}$$

where $z = \mathrm{P}_{S_0}(x)$. If we can lower bound the left-hand side of (5.1) by a positive constant at all the x of interest, then $(\gamma_0(x) - 1)_+$ becomes a proxy for $\mathrm{dist}(x, S_0)$. In this chapter, under the assumption $\mathrm{int}(S_0) \neq \emptyset$, we discuss two subgradient algorithms that solve the convex feasibility problem by minimizing $(\gamma_0(x) - 1)_+$.

Before proceeding to the algorithms, let us recall a fundamental result in the subgradient method literature:

Lemma 5.1. *Given a convex function $f : \mathcal{E} \to \mathbb{R}$, consider any $x \in \mathcal{E}$ and $g \in \partial f(x)$. Define $x_+ := x - ag$, where $a \geq 0$ is a step size. Then for any $z \in \mathcal{E}$*

such that $f(z) \leq f(x)$*, we have*

$$\|x_+ - z\| \leq \|x - z\|, \ \forall a \in [0, (2f(x) - f(z))/\|g\|^2].$$

Proof. By the definition of subgradients,

$$\begin{aligned} \|x_+ - z\|^2 &= \|x - z\|^2 + a(a\|g\|^2 + 2\langle z - x, g\rangle) \\ &\leq \|x - z\|^2 + a(a\|g\|^2 + 2(f(z) - f(x))). \end{aligned} \tag{5.2}$$

The statement follows immediately. □

5.1 Polyak's rule

Algorithm 1 is a direct application of Polyak's rule [21]:

Algorithm 1: Subgradient Method with Polyak's Rule

input : target accuracy $\epsilon > 0$ and an initial iterate $x_0 \in \mathcal{E}$

output: a point $\bar{x} \in \mathcal{E}$ satisfying $\gamma_0(\bar{x}) \leq 1 + \epsilon$

initialization: let $k = 0$

while $\gamma_0(x_k) > 1 + \epsilon$ **do**

- let $i_k \in \operatorname{argmax}_{i \in [m]} \gamma_i(x_k)$;
- compute $g_k \in \partial\gamma_{i_k}(x_k) \subseteq \partial\gamma_0(x_k)$;
- $x_{k+1} := x_k - \frac{\gamma_0(x_k) - 1}{\|g_k\|^2} g_k$;
- compute $\gamma_0(x_{k+1}) = \max_{i \in [m]} \gamma_0(x_{k+1})$;
- $k = k + 1$;

return $\bar{x} := x_k$

In Algorithm 1, when $x_k \notin S_0$, by Proposition 2.1, we have $\langle \pi_{S_{i_k}}(x_k) - e_{i_k}, g_k\rangle = 1$. Noting that the three points x_k, $\pi_{S_{i_k}}(x_k)$ and e_{i_k} are on the same line, we get

$$\gamma_0(x_k) - 1 = \frac{\|x_k - \pi_{S_{i_k}}(x_k)\|}{\|\pi_{S_{i_k}}(x_k) - e_{i_k}\|} = \frac{\langle x_k - \pi_{S_{i_k}}(x_k), g_k\rangle}{\langle \pi_{S_{i_k}}(x_k) - e_{i_k}, g_k\rangle} = \langle x_k - \pi_{S_{i_k}}(x_k), g_k\rangle,$$

and

$$\begin{aligned}\langle x_{k+1}, g_k\rangle &= \langle x_k, g_k\rangle - \frac{\gamma_0(x_k)-1}{\|g_k\|^2}\langle g_k, g_k\rangle \\ &= \langle x_k, g_k\rangle - \langle x_k - \pi_{S_{i_k}}(x_k), g_k\rangle \\ &= \langle \pi_{S_{i_k}}(x_k), g_k\rangle.\end{aligned}$$

Since $g_k \in \mathcal{N}_{S_{i_k}}(\pi_{S_{i_k}}(x_k))$ (by Proposition 2.1), we have $\langle \pi_{S_{i_k}}(x_k), g_k\rangle = f_{S_{i_k}}(g_k)$ and

$$x_{k+1} = \mathrm{P}_{H^-(S_{i_k}, g_k)}(x_k). \tag{5.3}$$

Consequently, in each iteration of Algorithm 1, we orthogonally project the current iterate onto a supporting half-space of one of the individual sets to which the current iterate does not belong.

In contrast, in the method of orthogonal projections where each iterate is orthogonally projected onto the farthest set (see, e.g., [1, 11]), by Lemma 2.5, one is essentially projecting onto the *farthest* supporting half-space of the individual sets. Hence when orthogonal projections onto the individual sets are available, this method makes more progress per iteration than Algorithm 1.

Theorem 5.1. *When* $\mathrm{int}(S_0) \neq \emptyset$*, for any* x_k *in Algorithm 1, let* $\mu_k > 0$ *satisfy*

$$\gamma_0(x_k) - 1 = \mu_k \cdot \mathrm{dist}(x_k, S_0).$$

Then for any $\epsilon > 0$*, when*

$$k \geq 2\left(\frac{1}{r_0 \min_{k' \leq k} \mu_{k'}}\right)^2 \log_2\left(\frac{\mathrm{dist}(x_0, S_0)}{\epsilon r_0}\right),$$

we must have $\min_{k' \leq k} \gamma_0(x_{k'}) \leq 1 + \epsilon$.

Remark. For any $k \in \mathbb{N}$, by (3.19), we have

$$\frac{1}{r_0 \mu_k} \leq \frac{1}{r_0} \frac{\max_{i\in[m]} h_{S_i}(e_i)}{\mathrm{inrad}(\mathcal{T}_{S_0}(z_k))} = \mathrm{inrad}(\mathcal{T}_{S_0}(z_k))^{-1} \tau_{\{(S_i, e_i)\}_{i\in[m]}}. \tag{5.4}$$

Let $z_k = \mathrm{P}_{S_0}(x_k)$. Then by the left-hand side of (5.1), we also have

$$\mu_k \geq \operatorname{inrad} \mathcal{T}_{S_0}(z_k) \left(\frac{1}{\max_{\gamma_i(x_k)>1} \|\pi_{S_i}(x_k) - e_i\|} \right). \tag{5.5}$$

In Section 5.1.1, we discuss ways to lower bound μ_k via (5.5).

Proof. In Algorithm 1, when $\gamma_0(x_k) \geq 1 + \epsilon$, let $z_k = \mathrm{P}_{S_0}(x_k)$. By (5.2), we have

$$\begin{aligned} \operatorname{dist}(x_{k+1}, S_0)^2 &\leq \|x_{k+1} - z_k\|^2 \\ &\leq \|x_k - z_k\|^2 + \frac{\gamma_0(x_k) - 1}{\|g_k\|^2} \left(\frac{\gamma_0(x_k) - 1}{\|g_k\|^2} \|g_k\|^2 + 2(1 - \gamma_0(x_k)) \right) \\ &= \operatorname{dist}(x_k, S_0)^2 - \left(\frac{\gamma_0(x_k) - 1}{\|g_k\|} \right)^2. \end{aligned} \tag{5.6}$$

According to Lemma 3.3, we see that γ_0 is $1/r_0$-Lipschitz. Hence

$$\operatorname{dist}(x_{k+1}, S_0)^2 \leq (1 - (r_0\mu_k)^2) \operatorname{dist}(x_k, S_0)^2. \tag{5.7}$$

By induction, if Algorithm 1 does not terminate after the first k iterations, then by the Lipschitzness of γ_0, we have

$$\gamma_0(x_{k+1}) - 1 \leq \frac{\operatorname{dist}(x_{k+1}, S_0)}{r_0} \leq \frac{\sqrt{\prod_{k' \leq k} (1 - (r_0\mu_{k'})^2)} \operatorname{dist}(x_0, S_0)}{r_0}.$$

The rest of the analysis is standard to the subgradient method literature, and is deferred to Appendix A.4. □

In (5.5), the first term depends on the geometry of S_0 near z_k, whereas the second term is affected by the positions of the reference points. For an algorithm where iterates are orthogonally projected onto the farthest set in each iteration, we can drop the second terms in (5.4) and (5.5) and obtain a faster rate.

5.1.1 Lower bounds on μ_k

The bound in Theorem 5.1 relies on $\min_{k' \le k} \mu_k$, which can be lower bounded by a strictly positive constant involving x_0, $\{S_i\}_{i\in[m]}$ and $\{e_i\}_{i\in[m]}$:

Lemma 5.2. *Consider any $w \in \operatorname{int}(S_0)$. Given an initial iterate x_0, let $z_0 := \mathrm{P}_{S_0}(x_0)$. If Algorithm 1 does not terminate after the first k iteration, then we have*

$$\operatorname{inrad} \mathcal{T}_{S_0}(z_k) \ge \frac{r_{S_0}(w)}{\|x_0 - w\|}$$

and

$$\max_{\gamma_i(x)>1} \|\pi_{S_i}(x_k) - e_i\| \le \|x_0 - z_0\| + \max_{i\in[m]} \|z_0 - e_i\|.$$

Hence by (5.5)*,*

$$\mu_k \ge \frac{r_{S_0}(w)}{\|x_0 - w\|(\|x_0 - z_0\| + \max_{i\in[m]} \|z_0 - e_i\|)}.$$

Remark. In the proof of Lemma 5.2, we only make use of the fact that $z_0 \in S_0$. Hence the statements remain true when we replace z_0 with any $z \in S_0$.

Proof. Let $z_k := \mathrm{P}_{S_0}(x_k)$. Note that we have $\gamma_0(w) < 1 < \gamma_0(x_{k'})$ for all $k' \le k$. Substituting $x_{k'}$ and w into Lemma 5.1, by induction, we get

$$\|z_k - w\| < \|x_k - w\| \le \|x_{k-1} - w\| \le \cdots \le \|x_0 - w\|.$$

Due to the characterization of $\operatorname{inrad} \mathcal{T}_{S_0}(z_k)$ in (3.7), we get $\operatorname{inrad} \mathcal{T}_{S_0}(z_k) \ge \frac{r_{S_0}(w)}{\|x_0 - w\|}$.

Similarly, we have $\|x_k - z_0\| \le \|x_0 - z_0\|$. For all $i \in [m]$ such that $\gamma_i(x_k) > 1$, we have $\pi_{S_i}(x_k) \in (e_i, x_k]$. Hence

$$\begin{aligned}
\max_{\gamma_i(x_k)>1} \|\pi_{S_i}(x_k) - e_i\| &\le \max_{\gamma_i(x_k)>1} \|x_k - e_i\| \\
&\le \|x_k - z_0\| + \max_{\gamma_i(x_k)>1} \|z_0 - e_i\| \\
&\le \|x_0 - z_0\| + \max_{i\in[m]} \|z_0 - e_i\|.
\end{aligned}$$

□

The position of the initial iterate x_0 plays an important role in Lemma 5.2. To get some insight into how the choice of x_0 could affect the lower bounds in Lemma 5.2, let us define

$$D(x) := \max_{i\in[m]} \|x - e_i\|.$$

Then $D(x)$ measures the distance between x and the farthest reference point. So long as the reference points do not coincide, we have $D(x) > 0$ for all $x \in \mathcal{E}$. Also let

$$\bar{z} := \operatorname*{argmin}_{z\in S_0} D(z).$$

In some sense, $D(\bar{z})$ characterizes the distance between S_0 and the reference points $\{e_i\}_{i\in[m]}$, and higher values of $D(\bar{z})$ signify more difficult feasibility problems.

For any $w \in S_0$, define

$$\beta(w) := \frac{r_{S_0}(w)}{D(w)}.$$

The convex feasibility problem is trivial when any of the reference points lie in S_0. Hence we may assume

$$\beta^\star := \sup_{w\in S_0} \beta(w) < 1.$$

For points in S_0, the function β provides an upper bound on γ_0:

Lemma 5.3. *For any $w \in S_0$, we have $\gamma_0(w) \le \frac{1}{1+\beta(w)}$.*

Proof. For any $i \in [m]$, we have $\|w - e_i\| \le D(w)$ and

$$w + r_{S_0}(w)\frac{w - e_i}{D(w)} \in \overline{B(w, r_{S_0}(w))} \subseteq S_i.$$

Hence

$$\gamma_i(w) \le \frac{\|w - e_i\|}{\left\|\left(w + r_{S_0}(w)\frac{w-e_i}{D(w)}\right) - e_i\right\|} = \frac{1}{1 + \frac{r_{S_0}(w)}{D(w)}} = \frac{1}{1+\beta(w)}.$$

By the definition of γ_0, we see that the statement is true. □

The next lemma shows that there exist points that have relatively high β values while not being too far from the reference points:

Lemma 5.4. *There exists $w \in S_0$ such that $\beta(w) \geq \beta^\star/2$, while $D(w) \leq 2D(\bar{z})$.*

Proof. Given any $y \in S_0$ and $\lambda \in [2, \infty)$, we will show that there exists $\tilde{y} \in S_0$ such that $\beta(\tilde{y}) \geq \beta(y)/\lambda$ while $D(\tilde{y}) \leq \frac{\lambda}{\lambda-1} D(\bar{z})$, and the statement follows from taking $\lambda = 2$.

For any $t \in [0, 1]$, define $y(t) := \bar{z} + t(y - \bar{z})$. For any $i \in [m]$, consider the function $f_i(t) := \|y(t) - e_i\|$. When $t \in [0, 1]$, by the convexity of f_i and the triangle inequality, we have

$$\begin{aligned} \|y(t) - e_i\| = f_i(t) &\leq f(0) + t(f(1) - f_i(0)) \\ &= \|\bar{z} - e_i\| + t(\|y - e_i\| - \|\bar{z} - e_i\|) \\ &\leq D(\bar{z}) + tD(y). \end{aligned}$$

In particular, consider

$$t' = \frac{D(\bar{z})}{(\lambda - 1)D(y)} \leq \frac{D(\bar{z})}{D(y)} \leq 1.$$

then

$$D(y(t')) = \max_{i \in [m]} \|y(t') - e_i\| \leq D(\bar{z}) + \frac{D(\bar{z})D(y)}{(\lambda - 1)D(y)} = \frac{\lambda}{\lambda - 1} D(\bar{z}). \tag{5.8}$$

On the other hand, note that

$$\overline{B(y(t'), t' r_{S_0}(y))} = t' \overline{B(y, r_{S_0}(y))} + (1 - t')\bar{z} \subseteq S_0.$$

Hence $r_{S_0}(y(t')) \geq t' r_{S_0}(y)$, by (5.8), this yields

$$\beta(y(t')) = \frac{r_{S_0}(y(t'))}{D\big(y(t')\big)} \geq \frac{t'(\lambda - 1) r_{S_0}(y)}{\lambda D(\bar{z})} = \frac{r_{S_0}(y)}{\lambda D(y)} = \frac{\beta(y)}{\lambda}.$$

□

Let

$$\bar{e} := \frac{\sum_{i\in[m]} e_i}{m},$$

and consider $\bar{w} \in S_0$ satisfying the conditions in Lemma 5.4. Then we have $\|\bar{e} - \bar{z}\| \leq D(\bar{z})$ and

$$\frac{r_{S_0}(\bar{w})}{\|\bar{e} - \bar{z}\| + \max_{i\in[m]} \|\bar{z} - e_i\|} \geq \frac{r_{S_0}(\bar{w})}{2D(\bar{z})} \geq \frac{\beta^\star}{4}.$$

Also note that

$$\|\bar{e} - \bar{w}\| \leq \max_{i\in[m]} \|e_i - \bar{w}\| \leq 2D(\bar{z}).$$

Replace z_0 by $\bar{z}$ and let $w = \bar{w}$ in Lemma 5.2. By the remark following the lemma, we get

$$\mu_k \geq \frac{1}{\|\bar{e} - \bar{w}\|} \cdot \frac{r_{S_0}(\bar{w})}{\|\bar{e} - \bar{z}\| + \max_{i\in[m]} \|\bar{z} - e_i\|} \geq \frac{\beta^\star}{8D(\bar{z})}. \tag{5.9}$$

Consequently, when no further information is available, $\bar{e}$ is a good candidate for the initial iterate.

5.2 The cyclic scheme

In each iteration of Algorithm 1, in order to evaluate $\gamma_0(x_k)$ and get $g_k \in \partial\gamma_0(x_k)$, we need to compute the individual $\gamma_i(x_k)$ for all $i \in [m]$. By evaluating one $\gamma_i(x)$ at a time (and making subgradient updates accordingly) and iterating through all $i \in [m]$, we obtain the following cyclic algorithm:

Algorithm 2: Cyclic Subgradient Method with Polyak's Rule

input : an initial iterate $x_0 \in \mathcal{E}$

initialization: let $k = 0$ and $x_0^1 = x_0$

repeat

  let $i = 1$; `// loop through` $i \in [m]$

  while $i \leq m$ **do**

    compute $\gamma_i(x_k^i)$;

    if $\gamma_i(x_k^i) > 1$ **then**

      compute $g_k^i \in \partial\gamma_i(x_k^i)$;

      $x_k^{i+1} := x_k^i - \frac{\gamma_i(x_k^i)-1}{\|g_k^i\|^2} g_k^i$;

    else

      $x_k^{i+1} = x_k^i$;

    $i = i + 1$;

  $x_{k+1}^1 = x_k^{m+1}$;

  $k = k + 1$;

Remark. In Algorithm 2, one can evaluate $\gamma_0(x_{k+1}^1)$ at the end of each loop at the cost of doubling the total number of oracle calls of the individual γ functions per loop.

As we will see in the proof of Theorem 5.2, in a loop of Algorithm 2, it is crucial to go through all the functions $\{\gamma_i\}_{i\in[m]}$, but the order in which the functions are visited does not affect the analysis. Hence one could go through different orders in every iteration, and the rate in Theorem 5.2 would remain unchanged.

As its name suggests, during each loop in Algorithm 2, we apply subgradient updates (if necessary) according to $\{\gamma_i\}_{i\in[m]}$ in a cyclic fashion. This can be seen as a generalization of the idea behind the alternating projection algorithm by John

von Neumann [30], who considered the convex feasibility problem of two sets.

Theorem 5.2. *When* $\operatorname{int}(S_0) \neq \emptyset$*, for any* $k \in \mathbb{N}$ *in Algorithm 2, let* $z_k = \mathrm{P}_{S_0}(x_k^1)$ *and*

$$\nu_k := \frac{\operatorname{inrad} \mathcal{T}_{S_0}(z_k)}{\sqrt{2}} \min\left\{ \frac{1}{\sqrt{m-1}}, \min_{\gamma_i(x_k)>1} \left(\frac{r_i}{\|\pi_{S_i}(x_k^i) - e\|} \right) \right\}. \tag{5.10}$$

Then for any $\epsilon > 0$*, when*

$$k \geq \frac{1}{(\min_{k' \leq k} \nu_{k'})^2} \log_2 \left(\frac{\operatorname{dist}(x_0, S_0)}{\epsilon r_0} \right),$$

we must have $\min_{k' \leq k} \gamma_0(x_{k'}^1) \leq 1 + \epsilon$*.*

Remark. Similar to the remarks following Theorem 5.1, the statement remains true if we replace ν_k with

$$\frac{\operatorname{inrad} \mathcal{T}_{S_0}(z_k)}{\sqrt{2}} \cdot \min\left\{ \frac{1}{\sqrt{m-1}}, \frac{1}{\max_{i \in [m]} \tau_{(S_i, e_i)}} \right\}. \tag{5.11}$$

Proof. When $x_k^1 \notin S_0$, the vector z_k serves as the basis of our analysis. For any $i \in [m]$, we have

$$\|x_k^{i+1} - x_k^i\|^2 = \left(\frac{(\gamma_i(x_k^i) - 1)_+}{\|g_k\|} \right)^2.$$

Since $z_k \in S_0 \subseteq S_i$, we get $\gamma_i(z_k) \leq 1$. Hence substituting x_k^i and z_k into Lemma 5.1, we find

$$\|x_k^{i+1} - z_k\|^2 \leq \|x_k^i - z_k\|^2 - \left(\frac{(\gamma_i(x_k^i) - 1)_+}{\|g_k\|} \right)^2 = \|x_k^i - z_k\|^2 - \|x_k^{i+1} - x_k^i\|^2. \tag{5.12}$$

Let i_k be the index satisfying $\operatorname{dist}(x_k^1, S_{i_k}) = \max_{i \in [m]} \operatorname{dist}(x_k^1, S_i)$. Telescoping yields

$$\begin{aligned} \|x_{k+1}^1 - z_k\|^2 &\leq \|x_k^1 - z_k\|^2 - \sum_{i=1}^{m} \|x_k^{i+1} - x_k^i\|^2 \\ &\overset{(a)}{\leq} \|x_k^1 - z_k\|^2 - \sum_{i=1}^{i_k} \|x_k^{i+1} - x_k^i\|^2 \\ &= \operatorname{dist}(x_k, S_0)^2 - \left(\sum_{i=1}^{i_k - 1} \|x_k^{i+1} - x_k^i\|^2 + \|x_k^{i_k+1} - x_k^{i_k}\|^2 \right). \end{aligned} \tag{5.13}$$

By the Cauchy-Schwartz inequality and the triangle inequality, we get

$$\sum_{i=1}^{i_k-1}\|x_k^{i+1}-x_k^i\|^2 \geq \frac{\left(\sum_{i=1}^{i_k-1}\|x_k^{i+1}-x_k^i\|\right)^2}{i_k-1} \geq \frac{\|x_k^1-x_k^{i_k}\|^2}{i_k-1} \geq \frac{\|x_k^1-x_k^{i_k}\|^2}{m-1}. \quad (5.14)$$

Also note that by the sandwich inequality (2.18), we have

$$\|x_k^{i_k+1}-x_k^{i_k}\| = \frac{(\gamma_{i_k+1}(x_k^{i_k})-1)_+}{\|g_k^{i_k}\|} \geq \frac{r_{i_k}}{\|\pi_{S_i}(x_k^{i_k})-e\|}\mathrm{dist}(x_k^{i_k},S_{i_k}).$$

Combining the last two inequalities, again by Cauchy-Schwartz and the triangle inequality, we have

$$\begin{aligned}
&\sum_{i=1}^{i_k-1}\|x_k^{i+1}-x_k^i\|^2+\|x_k^{i_k+1}-x_k^{i_k}\|^2\\
&\geq \frac{1}{2}\min\left\{\frac{1}{m-1},\left(\frac{r_{i_k}}{\|\pi_{S_i}(x_k^{i_k})-e\|}\right)^2\right\}\left(\|x_k^1-x_k^{i_k}\|+\mathrm{dist}(x_k^{i_k},S_{i_k})\right)^2\\
&\geq \frac{1}{2}\min\left\{\frac{1}{m-1},\left(\frac{r_{i_k}}{\|\pi_{S_i}(x_k^{i_k})-e\|}\right)^2\right\}\mathrm{dist}(x_k^1,S_{i_k})^2.
\end{aligned}$$

Substituting into (5.13) yields

$$\|x_{k+1}^1-z_k\|^2 \leq \mathrm{dist}(x_k,S_0)^2-\frac{1}{2}\min\left\{\frac{1}{m-1},\left(\frac{r_{i_k}}{\|\pi_{S_i}(x_k^{i_k})-e\|}\right)^2\right\}\mathrm{dist}(x_k^1,S_{i_k})^2.$$

Recall that $\mathrm{dist}(x_k^1,S_{i_k})=\max_{i\in[m]}\mathrm{dist}(x_k^1,S_i)$. By Theorem 3.1, we have

$$\begin{aligned}
&\mathrm{dist}\left(x_{k+1}^1,S_0\right)^2\\
&\leq \mathrm{dist}(x_k^1,S_0)^2-\frac{\mathrm{inrad}\,\mathcal{T}_{S_0}(z_k)^2}{2}\min\left\{\frac{1}{m-1},\left(\frac{r_{i_k}}{\|\pi_{S_i}(x_k^{i_k})-e\|}\right)^2\right\}\mathrm{dist}(x_k^1,S_0)^2\\
&\leq (1-\nu_k^2)\mathrm{dist}\left(x_k^1,S_0\right)^2.
\end{aligned}$$

The rest of the proof is similar to that of Theorem 5.1. □

5.3 Polyak's rule vs. the cyclic scheme

In a loop of Algorithm 2, an oracle of each γ_i is called exactly once, which is also the case for one iteration of Algorithm 1. Consequently, it is reasonable to compare

the number of iterations required by Algorithm 1 and loops required by Algorithm 2 to reach the same accuracy. One should also bear in mind the while a loop of Algorithm 2 requires a first-order oracle call of all the γ_i, an iteration of Algorithm 1 only needs to call the the first-order oracle of the maximal γ_i and zeroth-order oracles of the individual γ functions.

As shown by (5.4) and (5.11), the comparison between Algorithm 1 and 2 boils down to

$$\tau_{\{(S_i,e_i)\}_{i\in[m]}} \text{ vs. } \max\left\{\sqrt{m-1}, \max_{i\in[m]} \tau_{(S_i,e_i)}\right\},$$

where a smaller quantity signifies a faster algorithm.[1]

To see why the convergence rates of the two algorithms involve different condition numbers, consider any $x \notin S_0$. Let $i' \in \operatorname{argmax}_{i\in[m]} \gamma_i(x)$ and $i'' \in \operatorname{argmax}_{i\in[m]} \operatorname{dist}(x, S_i)$. Then for any $g \in \partial\gamma_{i'}(x)$, by (2.23), we have

$$\begin{aligned}
\frac{\gamma_0(x)-1}{\|g\|} = \frac{\gamma_{i'}(x)-1}{\|g\|} &\geq g_{S_{i'}}(e_{i'})(\gamma_{i'}(x)-1) \\
&\geq g_{S_{i'}}(e_{i'})(\gamma_{i''}(x)-1) \\
&\geq \frac{g_{S_{i'}}(e_{i'})}{h_{S_{i''}}(e_{i''})}\operatorname{dist}(x, S_{i''}) \\
&= \frac{g_{S_{i'}}(e_{i'})}{h_{S_{i''}}(e_{i''})} \cdot \max_{i\in[m]} \operatorname{dist}(x, S_i).
\end{aligned}$$

Consequently, in the analysis of Algorithm 1, we need to compare the Lipschitz constant and growth rate of different γ functions, leading to $\tau_{\{(S_i,e_i)\}_{i\in[m]}}$.

In contrast, for Algorithm 2, our analysis only concerns an

$$i_k \in \operatorname*{argmax}_{i\in[m]} \operatorname{dist}(x_k^1, S_i).$$

Hence it is the condition numbers of the individual γ functions that matter.

[1]Recall that by (3.18), we have $\tau_{\{(S_i,e_i)\}_{i\in[m]}} \geq \max_{i\in[m]} \tau_{(S_i,e_i)}$.

Also note that when $\gamma_i(x_k^i) > 1$ for most of $i \in [m]$, the Cauchy-Schwartz inequality in (5.14) could be very loose. Part (a) of the inequality (5.13) is also very conservative, since in the loop studied, the progress made after encountering γ_{i_k} is not accounted for. Consequently, our analysis of Algorithm 2 could be very pessimistic in practice, as we will see in the experiments.

5.4 Intersection of ellipsoids

Computing orthogonal projections onto ellipsoids in $\mathbb{R}^n$ is non-trivial, and some papers were dedicated to this task [7, 16]. Methods to compute orthogonal projections onto the intersection of ellipsoids have also been studied by [14, 17]. However, the residuals of ADMM-based method in [14] does not converge at a linear convergence rate, while the algorithm proposed in [17] posits a feasible initial iterate, i.e., $x_0 \in S_0$, and applies the interior point method at each iteration.

Consider m ellipsoids in $\mathbb{R}^n$ of the form

$$S_i := \{x \mid \|B_i(x - c_i)\| \leq a_i\}, \tag{5.15}$$

where $B_i \in \mathbb{R}^{n\times n}$, $c_i \in \mathbb{R}^n$ and $a_i > 0$. For each S_i, we define the γ function for S_i by taking c_i as the reference point. One can see that

$$\tau_{(S_i,c_i)} = \kappa(B_i),$$

where $\kappa(B_i)$ denotes the condition number of B_i.

In our experiments, the directions of centers of the ellipsoids are generated by the uniform distribution over the unit sphere in $\mathbb{R}^n$. The eigenvectors of $\{B_i\}_{i\in[m]}$ are drawn from the Haar measure on orthogonal group $O(n)$. For any $r > 0$, we

set

$$a_i = \|B_i c_i\| + \|B_i\| r, \tag{5.16}$$

where $\|B_i\|$ denotes the operator norm of B_i. Hence $\vec{0} \in S_0$ and $r_{S_0}(\vec{0}) \geq r$. We can control the width of S_0 via r.

We apply Algorithm 1 and 2 to solve the feasibility problem of such ellipsoids. The convergence behavior is presented in Figure 5.1 and Figure 5.2. In these experiments, we let the initial iterate x_0 be a random vector in $\mathbb{R}^{100}$ satisfying $\|x_0\| = 100$.[2] Both algorithms exhibit linear convergence rates.

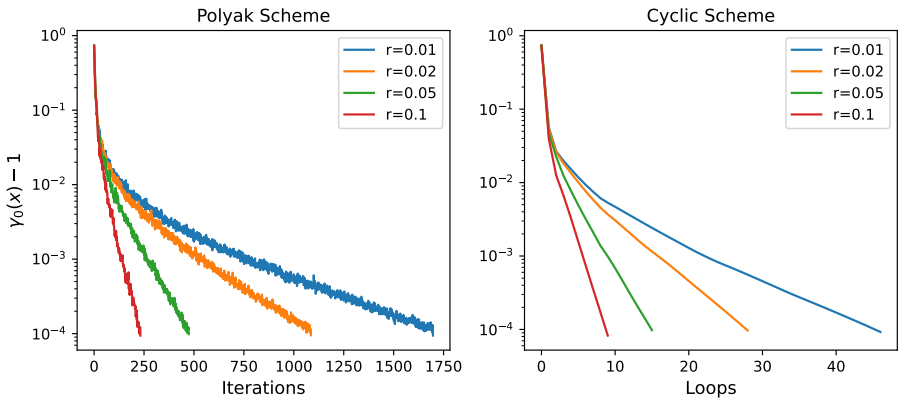

Figure 5.1: Algorithm 1 and 2 applied to the problem of the intersection of 100 ellipsoids in $\mathbb{R}^{100}$. For any i, we let $\|c_i\| = 100$ and $\|B_i\| = \kappa(B_i) = 10$.

In Figure 5.1, by only changing r in (5.16) ($\{c_i\}_{i\in[m]}$ and $\{B_i\}_{i\in[m]}$ remain unchanged across the experiment), we examine how the width of the intersection affects the convergence rate of the algorithms. As our theory predicts, when r increases, the intersections get bigger, and both algorithms converge faster.

In Figure 5.2, by including more ellipsoids in the problem, we examine how the

[2]Note that by the setup of our experiments, $\mathbb{E}c_i = \vec{0} \in S_0$. Hence we take a random initial iterate (instead of the average of the centers) to better present the convergence behavior of the algorithms. Check Appendix A.4.1 for an experiment initialized at the average of the centers. Relatively speaking, the setup here better demonstrates how the width of the intersection affects convergence rates.

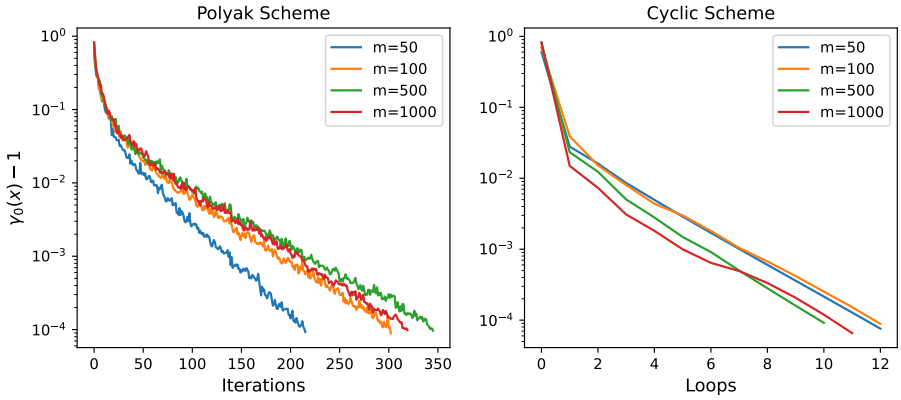

Figure 5.2: Algorithm 1 and 2 applied to the problem of the intersection of m ellipsoids in $\mathbb{R}^{100}$. For any i, we let $\|c_i\| = 100$ and $\|B_i\| = \kappa(B_i) = 10$.

number of ellipsoids involved affects the convergence rate. As our theory predicts, Algorithm 1 does not necessarily slow down (in terms of iteration counts) as the number of sets increases (compare the cases where $m = 500$ and $m = 1000$).

In both experiments, to reach the same accuracy, the number of loops taken by the cyclic scheme is much less than the number of iterations taken by Polyak's rule. As explained in Section 5.3, this could partially be explained by the fact that our analysis of the cyclic scheme could be very loose in practice.

More interestingly, although the convergence result in Theorem 5.2 includes the number of sets considered, in Figure 5.2, the number of sets does not have an obvious impact on the convergence rate of Algorithm 2. To understand this phenomenon intuitively, note that as the problem gets more complicated with the introduction of more ellipsoids, each loop of the cyclic scheme also makes more progress, as it gets to visit more sets (and make more progress) in a single loop.

CHAPTER 6

A STOCHASTIC ALGORITHM

Consider a probability triple $(\Omega, \mathcal{F}, P)$ and a system of convex inequality constraints

$$f_\omega(x) \leq 0, \quad \omega \in \Omega. \tag{6.1}$$

Beginning with Polyak [22] (and more recently, Necoara and Nedić [18]), under mild assumptions, the functional feasibility problem (6.1) has been tackled by considering

$$\min_{x \in \mathcal{E}} \mathbb{E}[f_\omega^+(x)],$$

where $f_\omega^+(x) := \max\{0, f_\omega(x)\}$.

When the intersection $S_0 = \cap_{i \in [m]} S_i$ is non-empty, finding a point $x \in S_0$ is equivalent to computing $x \in \mathcal{E}$ such that

$$\gamma_i(x) - 1 \leq 0, \ \forall i \in [m]. \tag{6.2}$$

Consequently, if $(\Omega, \mathcal{F}, P)$ is a distribution over $[m]$ such that $\mathbb{P}(\omega = i) > 0$ for all $i \in [m]$, then we can solve the convex feasibility problem (approximating a point $x \in S_0$) by considering the functional problem

$$\min_{x \in \mathcal{E}} \mathbb{E}[(\gamma_\omega(x) - 1)_+], \tag{6.3}$$

with 0 being its optimal value.

In this chapter, we solve the functional feasibility problem from a different perspective: Instead of considering the expected value of γ_ω as in (6.3), in each iteration of our algorithm, we sample γ_ω several times (more precisely, in Algorithm 3, we sample γ_ω without replacement) and update using a subgradient of the *maximal sampled function*. This approach leads to the guarantees in Theorem 6.1.

The methodology discussed in this chapter can be applied to the general context of convex functional constraints equipped with a probability triple. We refer the readers to [25] by Renegar and Zhou for the full discussion.

Algorithm 3: Stochastic Subgradient Method with Polyak's Rule

input: target accuracy $\epsilon > 0$, an initial iterate $x_0 \in \mathcal{E}$ and integer $L > 0$ (size of minibatch)

initialization: let $k = 0$

repeat

draw L indexes $\omega_{(k,1)}, \ldots, \omega_{(k,L)}$ from the uniform distribution over $[m]$ without replacement ;

select $i(k) \in \operatorname{argmax}_{i=\omega_{(k,1)},\ldots,\omega_{(k,L)}} \gamma_i(x_k)$;

if $\gamma_{i(k)}(x_k) > 1 + \epsilon$ **then**

compute $g_k \in \partial\gamma_{i(k)}(x_k)$;

let $x_{k+1} := x_k - \frac{\gamma_{i(k)}(x_k)-1}{\|g_k\|^2} g_k$

Remark. Algorithm 3 essentially becomes Algorithm 1 when $L = m$. Here we consider the uniform distribution over the indexes. Other distributions satisfying $\mathbb{P}(\omega = i) > 0$ for all $i \in [m]$ can also be used, and would lead to different convergence guarantees.

Given an initial iterate x_0 and a minibatch size L, it is straightforward to see that the lower bounds on the growth rate of γ_0 outside S_0 in Section 5.1.1 are also applicable to the iterates encountered in Algorithm 3 with minor adjustments. In the rest of this chapter, we let $\mu > 0$ denote the largest number such that

$$\gamma_0(x_k) - 1 \geq \mu \cdot \operatorname{dist}(x_k, S_0)$$

for all possible x_k encountered in Algorithm 3.

Given a target accuracy $\epsilon > 0$, define

$$K := 2\left(\frac{1}{r_0\mu}\right)^2 \log_2\left(\frac{\operatorname{dist}(x_0, S_0)}{\epsilon r_0}\right).$$

Algorithm 3 has the following convergence guarantee:

Theorem 6.1. *Algorithm 3 finds an iterate satisfying $\gamma_0(x) \leq 1+\epsilon$ within an expected number of iterations not exceeding $(\frac{m}{L})K$.*

Moreover, for any $k \geq (\frac{m}{L})K$, we have

$$\mathbb{P}\left(\min_{k'\leq k} x_k > 1+\epsilon\right) \leq \exp\left(-\frac{k(\frac{L}{m}-\frac{K}{k})^2}{2}\right).$$

Proof. In Algorithm 3, for any $k \in \mathbb{N}$, if $\gamma_{i(k)}(x_k) > 1+\epsilon$, then by substituting x_k for x and $\mathrm{P}_{S_0}(x_k)$ for z in (5.2), we get

$$\operatorname{dist}(x_{k+1}, S_0)^2 \leq \|x_k - \mathrm{P}_{S_0}(x_k)\|^2 - \left(\frac{\gamma_{i(k)}(x_k)-1}{\|g_k\|}\right)^2 \leq \operatorname{dist}(x_k, S_0)^2. \tag{6.4}$$

Moreover, if

$$\{\omega_{(k,1)}, \ldots, \omega_{(k,L)}\} \cap \operatorname*{argmax}_{i\in[m]} \gamma_i(x_k) \neq \emptyset, \tag{6.5}$$

then $i(k) \in \operatorname{argmax}_{i\in[m]} \gamma_i(x_k)$ and $\gamma_{i(k)}(x_k) = \gamma_0(x_k)$. Thus when $\gamma_0(x_k) > 1+\epsilon$ and (6.5) holds, similar to (5.7) in the analysis of Algorithm 1, we have

$$\operatorname{dist}(x_{k+1}, S_0)^2 = \operatorname{dist}(x_k, S_0)^2 - \left(\frac{\gamma_0(x_k)-1}{\|g_k\|}\right)^2 \leq \left(1-(r_0\mu)^2\right)\operatorname{dist}(x_k, S_0)^2. \tag{6.6}$$

Now for any $k \in \mathbb{N}$, define

$$X_k = \begin{cases} 1, & \text{when (6.5) holds;} \\ 0, & \text{otherwise.} \end{cases} \tag{6.7}$$

Since the indexes in Algorithm 3 are sampled from the uniform distribution over $[m]$ without replacement, we see that

$$\mathbb{P}\left(X_k = 1 \mid \gamma_0(x_k) > 1+\epsilon\right) \geq \frac{L}{m}.$$

Note that the indexes in different iterations are generated independently, we can lower bound $\{X_k\}_{k\in\mathbb{N}}$ by $\{Y_k\}_{k\in\mathbb{N}}$, a series of i.i.d. Bernoulli distribution which

takes the value 1 with probability $\frac{L}{m}$. By (6.4), the sequence $\{\text{dist}(x_k, S_0)\}_{k\in\mathbb{N}}$ is decreasing. The expected number of iterations required to reach the target accuracy follows from (6.6), the proof of Theorem 5.1, and a straightforward application of the Ward's equation on $\{Y_k\}_{k\in\mathbb{N}}$..

By Hoeffding's inequality, for any $k \in \mathbb{N}_+$ and $t > 0$, we have

$$\mathbb{P}\left(\frac{\sum_{k'<k} Y_{k'}}{k} \leq \frac{L}{m} - t\right) \leq \exp\left(-\frac{kt^2}{2}\right).$$

Thus for all $k \geq (\frac{m}{L})K$, we get

$$\mathbb{P}\left(\sum_{k'<k} Y_{k'} \leq K\right) = \mathbb{P}\left(\frac{\sum_{k'<k} Y_{k'}}{k} \leq \frac{K}{k}\right) \leq \exp\left(-\frac{k(\frac{L}{m} - \frac{K}{k})^2}{2}\right).$$

□

CHAPTER 7

A FINITE ALGORITHM FOR THE CONVEX FEASIBILITY PROBLEM

Recall that when $\text{int}(S_0) \neq \emptyset$, we have $\gamma_0^\star = \inf_{x \in \mathcal{E}} \gamma_0(x) < 1$. Now consider an iterative method applying Polyak's rule with a "target" value $t \in (\gamma_0^\star, 1)$, i.e.,

$$x_{k+1} := x_k - \frac{\gamma_0(x_k) - t}{\|g_k\|^2} g_k. \tag{7.1}$$

If its iterates converge to the t-sublevel set of γ_0, then we can find an point in S_0 when the accuracy $\epsilon = 1 - t$ is reached after finitely many iterations.

Recall that our analysis of the Polyak's rule in Section 5.1 relies on the growth rate of γ_0 with respect to S_0, the 1-sublevel set of γ_0. In order to propose a finite algorithm based on updates like (7.1), let us first study the growth rate of γ_0 with respect to generic sublevel sets which are nonempty:

7.1 Growth rate of γ_0 with respect to generic sublevel sets

Consider a single closed convex set S and the corresponding γ function.

For any $t > 0$, if $x \notin S(t)$, then $\gamma(x(1/t)) = \gamma(e + (x - e)/t) = \gamma(x)/t > 1$ and

$$x(1/t) \notin S. \tag{7.2}$$

By (2.5), we have

$$\frac{\gamma(x) - t}{t} = \frac{\gamma(x)}{t} - 1 = \gamma(x(1/t)) - 1 > 0.$$

Also note that (2.10) implies

$$\begin{aligned}\frac{\operatorname{dist}(x, S(t))}{t} &= \frac{\operatorname{dist}(e + (x-e), e + t(S-e))}{t} \\ &= \operatorname{dist}\left(e + \frac{x-e}{t}, e + (S-e)\right) \\ &= \operatorname{dist}(x(1/t), S).\end{aligned}$$

Consequently,

$$\frac{\gamma(x) - t}{\operatorname{dist}(x, S(t))} = \frac{\gamma(x(1/t)) - 1}{\operatorname{dist}(x(1/t), S)}. \tag{7.3}$$

Recall that (2.8) implies $\pi_S(x(1/t)) = \pi_S(x)$. By (7.2) and Lemma 2.4, we get

Lemma 7.1. *For any $t > 0$, when $x \notin S(t) \neq \emptyset$, we have* [1]

$$\frac{\gamma(x) - t}{\operatorname{dist}(x, S(t))} \geq \frac{1}{\|\pi_S(x) - e\|} \geq \frac{t}{\|\mathrm{P}_{S(t)}(x) - e\|} \geq \frac{t}{\|x - e\|}.$$

Consequently, for any $t > 0$ and $x \notin S(t)$, we get

$$\frac{1}{\|\pi_S(x) - e\|} \leq \frac{\gamma(x) - t}{\operatorname{dist}(x, S(t))} \leq \frac{1}{r_S(e)}. \tag{7.4}$$

Similarly, we can show that $\tau_{(S,e)}$ can also serve as a condition number of γ with respect to all t-sublevel set:

Lemma 7.2. *Assume $t > 0$ and $x \notin S(t)$. We have*

$$\frac{1}{h_S(e)} \leq \frac{\gamma(x) - t}{\operatorname{dist}(x, S(t))} \leq \frac{1}{g_S(e)}.$$

Proof. By (7.2), (7.3) and Corollary 2.1, we get

$$\frac{\gamma(x) - t}{\operatorname{dist}(x, S(t))} = \frac{\gamma(x(1/t)) - 1}{\operatorname{dist}(x(1/t), S)} = \max_{d \in \Omega_S} \left\{ \frac{\operatorname{dist}(x(1/t), H^-(S, d))}{\operatorname{dist}(e, H(S, d))} \right\}.$$

The rest of the proof follows from the definitions of $g_S(e)$ and $h_S(e)$ and substituting $x(1/t)$ for x in Lemma 2.5, i.e.,

$$\operatorname{dist}(x(1/t), S) = \max_{d \in \Omega_S} \operatorname{dist}\left(x(1/t), H^-(S, d)\right).$$

□

[1] The proof of the last two inequalities also relies on (2.10), and is deferred to Appendix A.5.

For any $t \geq 0$, let us denote the t-sublevel set of γ_0 by

$$S_0(t) := \{x \in \mathcal{E} \mid \gamma_0(x) \leq t\} = \cap_{i\in[m]} S_i(t) = \cap_{i\in[m]} \{e_i + t(S_i - e_i)\}. \tag{7.5}$$

Here the last equality follows from (2.10). In particular, we have $S_0 = S_0(1)$.

Proposition 7.1. *Assume $t > 0$, $x \notin S_0(t)$ and $\operatorname{int}(S_0(t)) \neq \emptyset$. Let $z = \mathrm{P}_{S_0(t)}(x)$. We have*

$$\begin{aligned} \frac{\gamma_0(x) - t}{\operatorname{dist}(x, S_0(t))} &\geq \frac{\operatorname{inrad}(\mathcal{T}_{S_0(t)}(z))}{\max_{\gamma_i(x)>t} \|\pi_{S_i}(x) - e\|} \\ &\geq \frac{t \cdot \operatorname{inrad}(\mathcal{T}_{S_0(t)}(z))}{\max_{\gamma_i(x)>t} \|\mathrm{P}_{S_i(t)}(x) - e\|} \\ &\geq \frac{t \cdot \operatorname{inrad}(\mathcal{T}_{S_0(t)}(z))}{\max_{\gamma_i(x)>t} \|x - e_i\|} \end{aligned} \tag{7.6}$$

and

$$\frac{\gamma_0(x) - t}{\operatorname{dist}(x, S_0(t))} \geq \frac{\operatorname{inrad}(\mathcal{T}_{S_0(t)}(z))}{\max_{i\in[m]} h_{S_i}(e_i)}. \tag{7.7}$$

Proof. The statements follows from Lemma 7.1 and Lemma 7.2, and the application of Theorem 3.1 to $S_0(t) = \cap_{i\in[m]} S_i(t)$. □

Consequently, a *condition number of* γ_0 with respect to non-empty $S_0(t)$ is

$$\operatorname{inrad}(\mathcal{T}_{S_0(t)}(z))^{-1} \tau_{\{(S_i,e_i)\}_{i\in[m]}}, \tag{7.8}$$

which is a generalization of (3.20), a condition number of γ_0 with respect to S_0.

We close this section by showing how the growth the growth rates of γ_0 with respect to different sublevel sets are connected:

Lemma 7.3. *Given any $x \in \mathcal{E}$ and $t, t' \in \mathbb{R}$ such that $\gamma_0^\star < t < t' < \gamma_0(x)$, we have*

$$\frac{\gamma_0(x) - t}{\operatorname{dist}(x, S_0(t))} \leq \frac{\gamma_0(x) - t'}{\operatorname{dist}(x, S_0(t'))}.$$

Hence γ_0 grows faster with respect to sublevel sets of higher values.

Proof. Let $z = \mathrm{P}_{S_0(t)}(x)$ and $w = x + \frac{\gamma_0(x)-t'}{\gamma_0(x)-t}(z-x)$. Then $z \in \mathrm{bdy}(S_0(t))$ and $\gamma_0(z) = t$. Due to the convexity of γ_0, we have

$$\gamma_0(w) \leq \frac{(\gamma_0(x)-t')\gamma_0(z) + (t'-t)\gamma_0(x)}{\gamma_0(x)-t} = \frac{(\gamma_0(x)-t')t + (t'-t)\gamma_0(x)}{\gamma_0(x)-t} = t'.$$

Consequently, $w \in S_0(t')$, and

$$\frac{\gamma_0(x)-t}{\mathrm{dist}(x, S_0(t))} = \frac{\gamma_0(x)-t}{\|x-z\|} = \frac{\gamma_0(x)-t'}{\|x-w\|} \leq \frac{\gamma_0(x)-t'}{\mathrm{dist}(x, S_0(t'))}.$$

□

7.2 A finite algorithm

In the Polyak's update rule (7.1), setting the target $t \in (\gamma_0^\star, 1)$ is crucial to obtaining a finite algorithm: If $t = 1$ (as in Algorithm 1), then by (5.3), we have $x_{k+1} = \mathrm{P}_{H^-(S_{i_k}, g_k)}(x_k)$, where the definitions of S_{i_k} and g_k follows from Algorithm 1. Hence $x_{k+1} \notin S_0$ unless

$$\mathrm{dist}(x_k, H^-(S_{i_k}, g_k)) = \mathrm{dist}(x_k, S_{i_k}) = \mathrm{dist}(x_k, S_0),$$

which is generally not true. Consequently, for generic convex feasibility problems, Algorithm 1 does not generate iterates in S_0. When $t > 1$, we even have $x_{k+1} \notin S_{i_k} \supseteq S_0$.

On the other hand, if we work with a target $t < \gamma_0^\star$, then the step-size is outside the interval provided in Lemma 5.1 for any $z \in S_0$, and the algorithm is not guaranteed to converge.

In practice, however, we do not know $\gamma_0^\star$ a priori. To tackle this, we present Algorithm 4, an algorithm which proceeds by making sequential estimates of $\gamma_0^\star$.

Algorithm 4: Finite Method with the Polyak's Rule

input : an initial iterate $x_0 \in \mathcal{E}$
output: a point $\bar{x} \in \mathcal{E}$ satisfying $\gamma_0(\bar{x}) \leq 1$
initialization: let $x_0^0 = x_0$, $J = 0$, $k_0 = 0$ and $\Delta_0 = 0$

repeat
  for $j \in [J]$ **do**
    if $\gamma_0(x_{k_j}^j) \leq 1$ **then**
      return $\bar{x} = x_{k_j}^j$
    else
      compute $g_{k_j}^j \in \partial\gamma_0(x_{k_j}^j)$;
      $x_{k_j+1}^j := x_{k_j}^j - \frac{\gamma_0(x_{k_j}^j)-(1-\Delta_j)}{\|g_{k_j}^j\|^2} g_{k_j}^j$;
      $k_j = k_j + 1$
  $j' = \operatorname{argmin}_{j\in[J]} \gamma_0(x_{k_j}^j)$, $\bar{x} = \operatorname{argmin}\left\{\gamma(x_{k_{j'}}^{j'}), \gamma(\bar{x})\right\}$;
  while $\gamma_0(x') \leq 1 + 2^{-(J+1)}$ **do**
    if $J = 0$ *or* $k_{\lceil\sqrt{J}\rceil} \geq 2^{\lceil\sqrt{J}\rceil}$ **then**
      $J = J + 1$;
      let $k_J = 0$, $\Delta_J = 2^{-J}$ and $x_0^J = \bar{x}$; `// initiate copy` J

In an iteration of Algorithm 4, we update $J+1$ different copies in parallel. For each integer j such that $1 \leq j \leq J$, we update the current iterate according to Polyak's rule, with the target being $1-\Delta_j = 1-2^{-j}$, while the copy 0 is essentially running Algorithm 1. Once initiated, different copies do not communicate with each other. According to our analysis in Section 5.1, the objective values in copy 0 converge to 1.

For $j \geq 1$, a copy with target $1 - \Delta_j$ is initiated when both of these criteria are met:

1. The algorithm has encountered an iterate satisfying $x_{k_j'}^{j'} \leq 1 + \Delta_j$. Since the objective values in copy 0 converge to 1, this condition will be satisfied after finitely many iterations.
2. Enough iterations of the algorithm has elapsed. This prevents us from work-

ing with too many copies simultaneously (and hence slowing down the algorithm).

In order to analyze Algorithm 4, for any $j \in \mathcal{N}$, let $\mu(j) > 0$ be the biggest number satisfying

$$\gamma_0(x_{k_j}^j) - (1 - \Delta_j) \geq \mu(j) \cdot \operatorname{dist}\left(x_{k_j}^j, S_0(1 - \Delta_j)\right)$$

for all the iterates generated by copy j in Algorithm 4. Then $\mu(0) = \min_{k \in \mathbb{N}} \mu_k$, where μ_k is defined as in Theorem 5.1, and the lower bounds in Section 5.1.1 are applicable to $\mu(0)$.

For any $j \in \mathbb{N}_+$ such that $1 - \Delta_j > \gamma_0^\star$, since γ_0 is Lipschitz-continuous, we see that $\operatorname{int}(S_0(1 - \Delta_j)) \neq \emptyset$. According to Proposition 7.1, we can lower bound $\mu(j)$ with the results in Section 5.1.1 by substituting $S_0(1 - \Delta_j)$ for S_0 when necessary.

We next discuss the time required for the initiation of a copy in Algorithm 4:

Lemma 7.4. *For any $j \geq 1$, if Algorithm 4 does note terminate after making*

$$2\left(\frac{1}{r_0\mu(0)}\right)^2 \log_2\left(\frac{\operatorname{dist}(x_0, S_0)}{r_0\Delta_j}\right) + 2^{\lceil\sqrt{j}\rceil+1}. \tag{7.9}$$

updates of copy 0, then copy j must have been initiated.

Proof. By Theorem 5.1, we see that $\min_{k' \leq k} \gamma_0(x_{k'}^0) \leq 1 + \Delta_j$ when

$$k \geq 2\left(\frac{1}{r_0\mu(0)}\right)^2 \log_2\left(\frac{\operatorname{dist}(x_0, S_0)}{r_0\Delta_j}\right).$$

Hence within additional

$$\sum_{1 \leq j' \leq \lceil\sqrt{j}\rceil} 2^{j'} = 2^{\lceil\sqrt{j}\rceil+1} - 1 \leq 2^{\lceil\sqrt{j}\rceil+1}$$

updates of copy 0, either the algorithm terminates, or copies 1 through j are initiated. □

Lemma 7.5. *For any $j \geq 1$ such that $1 - \Delta_j = 1 - 2^{-j} \in (\gamma_0^\star, 1)$, if copy j is initiated in Algorithm 4, then when*

$$k_j \geq \xi(k) := 2\left(\frac{1}{r_0\mu(j)}\right)^2 \log_2\left(\frac{2}{r_0\mu(j)}\right),$$

we must have $\min_{k' \leq k_j} \gamma_0(x_{k'}^j) \leq 1$.

Proof. By the construction of Algorithm 4, we have $\gamma_0(x_0^j) \leq 1 + \Delta_j$. Due to the definition of $\mu(j)$, we have

$$\operatorname{dist}\left(x_0^j, S_0(1-\Delta_j)\right) \leq \frac{\gamma_0(x_0^j) - (1-\Delta_j)}{\mu(j)} \leq \frac{2\Delta_j}{\mu(j)}.$$

Substitute $S_0(1-\Delta_j)$ for S_0, $\mu(j)$ for $\min_{k \leq k'} \mu_k$ and let $\epsilon = \Delta_j$ in Theorem 5.1, we get $\min_{k' \leq k_j} \gamma_0(x_{k'}^j) \leq 1$ when

$$k_j \geq 2\left(\frac{1}{r_0\mu(j)}\right)^2 \log_2\left(\frac{\operatorname{dist}\left(x_0^j, S_0(1-\Delta_j)\right)}{r_0\Delta_j}\right) = 2\left(\frac{1}{r_0\mu(j)}\right)^2 \log_2\left(\frac{2}{r_0\mu(j)}\right).$$

□

Let

$$\mathbf{j} := -\lceil \log_2(1-\gamma_0^\star) \rceil - 1. \tag{7.10}$$

Then $2^{\mathbf{j}} \in [\frac{1-\gamma_0^\star}{2}, 1-\gamma_0^\star)$. Combining Lemma 7.4 and Lemma 7.5 produces the following result:

Theorem 7.1. *Algorithm 4 terminates with copy 0 making no more than*

$$2\left(\frac{1}{r_0\mu(0)}\right)^2 \log_2\left(\frac{\operatorname{dist}(x_0, S_0)}{r_0\Delta_{\mathbf{j}}}\right) + 2\left(\frac{1}{r_0\mu(\mathbf{j})}\right)^2 \log_2\left(\frac{2}{r_0\mu(\mathbf{j})}\right) + 2^{\lceil\sqrt{\mathbf{j}}\rceil+1}$$

updates and making no more than

$$\left(\max\left\{\mathbf{j}, \log_2(\xi(\mathbf{j}))\right\}^2 + 1\right)\left(2\left(\frac{1}{r_0\mu(0)}\right)^2 \log_2\left(\frac{\operatorname{dist}(x_0, S_0)}{r_0\Delta_{\mathbf{j}}}\right) + 2\left(\frac{1}{r_0\mu(\mathbf{j})}\right)^2 \log_2\left(\frac{2}{r_0\mu(\mathbf{j})}\right) + 2^{\lceil\sqrt{\mathbf{j}}\rceil+1}\right)$$

first-order oracle calls of γ_0.

Remark. When the optimal value $\gamma_0^\star$ is known, we can remove the terms $(\max\{\mathbf{j}, \log_2(\xi(\mathbf{j}))\}^2 + 1)$ and $2^{\lceil\sqrt{\mathbf{j}}\rceil+1}$ in Theorem 7.1. The first terms is due to the number of copies run in parallel, and the second term follows from our second condition when initiating new copies. To get an rough idea of the magnitude of these two terms, we note that $\log_2(10^8)^2 \approx 706$ and $2^{-\sqrt{\log_2(10^{-8})}} \approx 36$.

Proof. The first statement follows immediately from Lemma 7.4 and Lemma 7.5.

To see the second statement, note that by Lemma 7.5, the algorithm terminates before copy j makes more than $\xi(\mathbf{j})$ updates. Thus when the algorithm terminates, copies with indexes greater than or equal to $\mathbf{j}$ have iteration counts less than or equal $\xi(\mathbf{j})$. By the second condition for the initiation of new copies in Algorithm 4, we get $J \leq \max\{\mathbf{j}, \log_2(\xi(\mathbf{j}))\}^2$. □

7.3 Demonstration on the intersection of ellipsoids

In Figure 7.1, we generate random ellipsoids as in Section 5.4, and apply Algorithm 4 to compute a point in the intersection of 100 ellipsoids.

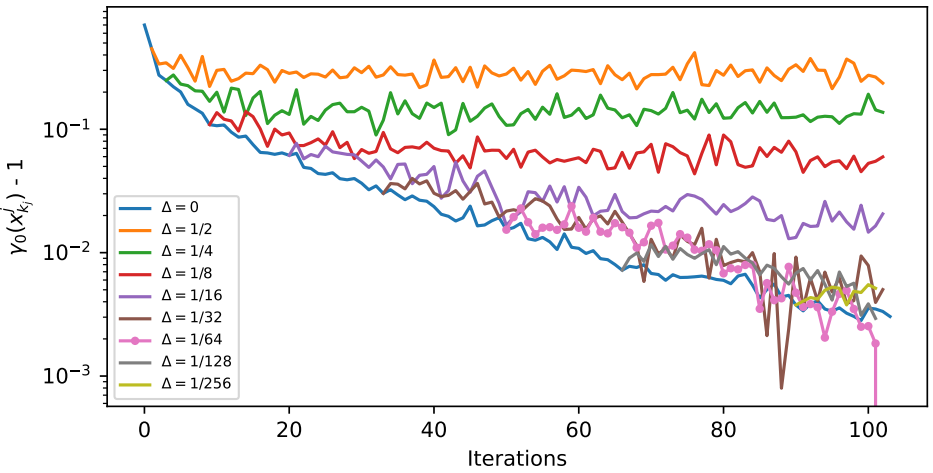

Figure 7.1: For any i, we let $\|c_i\| = 100$ and $\|B_i\| = \kappa(B_i) = 10$. In this experiment, the copy with target $1 - \frac{1}{64}$ (the dotted line) generates a point in S_0, while copies with target less than $1 - \frac{1}{16}$ do not converge, indicating $\gamma_0^\star > \frac{15}{16}$.

CHAPTER 8

THE γ FUNCTION FOR MULTIPLE SETS WHEN THE INTERSECTION HAS EMPTY INTERIOR

Recall that for a convex feasibility problem involving closed convex sets $\{S_i\}_{i\in[m]}$, its linear regularity concerns upper bounding the ratio $\frac{\operatorname{dist}(x,S_0)}{\max_{i\in[m]}\operatorname{dist}(x,S_i)}$ at $x \notin S_0 = \cap_{i\in[m]} S_i$.

Fix $z \in \operatorname{bdy}(S_0)$. When $\operatorname{int}(S_0) \neq \emptyset$, for any $x \notin S_0$ satisfying $\mathrm{P}_{S_0}(x) = z$, Theorem 3.1 upper bounds the linear regularity at x by a finite positive number $\operatorname{inrad}(\mathcal{T}_{S_0}(z))^{-1}$. In contrast, if $\operatorname{int}(S_0) = \emptyset$, then such an upper bound may not exist even for points in a bounded region:

Example 8.1. Consider two closed convex sets in $\mathbb{R}^2$:

$$S_1 = \{(x,y) \mid (x+1)^2 + y^2 \leq 1\} \cup \{(x,y) \mid x \in [-2,0] \text{ and } y \in [-1,0]\}$$

and

$$S_2 = \{(x,y) \mid (x-1)^2 + y^2 \leq 1\} \cup \{(x,y) \mid x \in [0,2] \text{ and } y \in [-1,0]\}.$$

Then $S_0 = \{(0,y) \mid y \in [-1,0]\}$, which is a set with empty interior. For any $y > 0$, we have

$$\mathrm{P}_{S_0}((0,y)) = (0,0)$$

and

$$\frac{\operatorname{dist}((0,y),S_0)}{\max\{\operatorname{dist}((0,y),S_1),\operatorname{dist}((0,y),S_2)\}} = \frac{y}{\sqrt{y^2+1}-1} = \frac{\sqrt{y^2+1}+1}{y},$$

which tends to ∞ when $y \searrow 0$.

Recall that the proof of Theorem 3.1 relies on a pair of polar cones, $\mathcal{T}_{S_0}(z)$ and $\mathcal{N}_{S_0}(z)$. When $\operatorname{int}(S_0) = \emptyset$, however, both cones exhibit different properties:

- Since $r_{S_0}(w) = 0$ for all $w \in S_0$, we have $\text{inrad}(\mathcal{T}_{S_0}(z)) = 0$ at all $z \in \text{bdy}(S_0)$.
- While the equality $\mathcal{N}_{S_1}(z) + \cdots + \mathcal{N}_{S_m}(z) \subseteq \mathcal{N}_{S_0}(z)$ always holds, without any additional assumptions, the equality

$$\mathcal{N}_{S_1}(z) + \cdots + \mathcal{N}_{S_m}(z) = \mathcal{N}_{S_0}(z)$$

could fail, preventing us from writing normal vectors to S_0 in terms of normal vectors to the individual sets.

For instance, in Example 8.1, let $z = (0, 0)$. Then $\mathcal{N}_{S_1}(z) = \mathbb{R}_+ \times \{0\}$, $\mathcal{N}_{S_2}(z) = \mathbb{R}_- \times \{0\}$ and $\mathcal{N}_{S_1}(z) + \mathcal{N}_{S_2}(z) = \mathbb{R} \times \{0\}$, while $\mathcal{N}_{S_0}(z) = \mathbb{R} \times \mathbb{R}_+$.

In order to establish a linear regularity result when the interior of S_0 is empty, let us introduce the following assumption:

Assumption 8.1. For closed convex sets $\{S_i\}_{i\in[m]}$, assume $S_0 = \cap_{i\in[m]} S_i \neq \emptyset$ and

$$\mathcal{N}_{S_1}(z) + \cdots + \mathcal{N}_{S_m}(z) = \mathcal{N}_{S_0}(z). \tag{8.1}$$

at all $z \in \text{bdy}(S_0)$.

In Chapter 9, we introduce Assumption 9.1, which can be seen as an analogy of Assumption 8.1 in the setting of constrained optimization.

Example 8.2 (Example 8.1 Continued). Let us consider an extra set in Example 8.1:

$$S_3 = \{(x, y) \mid x^2 + (y + 0.5)^2 \leq 0.25\}.$$

Obviously, we still have $S_0 = \cap_{i\in[3]} S_i = \{(0, y) \mid y \in [-1, 0]\}$. All three sets are active at $z = (0, 0)$. Since $\mathcal{N}_{S_3}(z) = \{0\} \times \mathbb{R}_+$, we get

$$\mathcal{N}_{S_1}(z) + \mathcal{N}_{S_2}(z) + \mathcal{N}_{S_3}(z) = \mathbb{R} \times \mathbb{R}_+ = \mathcal{N}_{S_0}(z).$$

Moreover, for any $y \geq 0$, we have

$$\max_{i\in[3]} \operatorname{dist}((0,y), S_i) = \operatorname{dist}((0,y), S_3) = y = \operatorname{dist}((0,y), S_0).$$

One can go on to check that Assumption 8.1 holds for this new problem.

For $z \in \operatorname{bdy}(S_0)$, let

$$\mathcal{I} := \{i \in [m] \mid z \in \operatorname{bdy}(S_i)\} \tag{8.2}$$

denote the sets active at z. Define the set

$$\begin{aligned} P_{\{S_i\}_{i\in[m]}}(z) :&= \operatorname{conv}\left(\{d \mid d \in \mathcal{N}_{S_i}(z),\ \|d\| = 1\}_{i\in\mathcal{I}} \cup \{\vec{0}\}\}\right) \\ &= \operatorname{conv}\left(\{d \mid d \in \mathcal{N}_{S_i}(z),\ \|d\| \leq 1\}_{i\in\mathcal{I}}\right). \end{aligned} \tag{8.3}$$

Then

$$P_{\{S_i\}_{i\in[m]}}(z) \subseteq \overline{B(\vec{0},1)} \cap (\mathcal{N}_{S_1}(z) + \cdots + \mathcal{N}_{S_m}(z)) \subseteq \overline{B(\vec{0},1)} \cap \mathcal{N}_{S_0}(z). \tag{8.4}$$

We also measure the *reach* of $P_{\{S_i\}_{i\in[m]}}(z)$ by

$$\operatorname{reach}_{\{S_i\}_{i\in[m]}}(z) := \max\left\{r \geq 0 \mid \left(\overline{B(\vec{0},r)} \cap \mathcal{N}_{S_0}(z)\right) \subseteq P_{\{S_i\}_{i\in[m]}}(z)\right\}.$$

By (8.4), $\operatorname{reach}_{\{S_i\}_{i\in[m]}}(z) \leq 1$. In particular, when S_0 is a singleton, we have $\mathcal{N}_{S_0}(z) = \mathcal{E}$ and[1]

$$\operatorname{reach}_{\{S_i\}_{i\in[m]}}(z) = r_{P_{\{S_i\}_{i\in[m]}}(z)}(\vec{0}). \tag{8.5}$$

Theorem 8.2. *Fix $z \in \operatorname{bdy}(S_0)$. If Assumption 8.1 holds, then*

$$\operatorname{reach}_{\{S_i\}_{i\in[m]}}(z) > 0$$

Moreover, for any $x \notin S_0$ satisfying $z = \mathrm{P}_{S_0}(x)$, we have

$$\operatorname{dist}(x, S_0) = \|x - z\| \leq \frac{\max_{i\in\mathcal{I}} \operatorname{dist}(x, S_i)}{\operatorname{reach}_{\{S_i\}_{i\in[m]}}(z)} \leq \frac{\max_{i\in[m]} \operatorname{dist}(x, S_i)}{\operatorname{reach}_{\{S_i\}_{i\in[m]}}(z)}. \tag{8.6}$$

[1]For the definition of $r_{P_{\{S_i\}_{i\in[m]}}(z)}(\vec{0})$, see (2.15).

Remark. Compare Theorem 8.2 and Theorem 3.1. We see that in Theorem 8.2, $\mathrm{reach}_{\{S_i\}_{i\in[m]}}(z)$ essentially plays the role of $\mathrm{inrad}(\mathcal{T}_{S_0})(z)$ in Theorem 3.1. When $\mathrm{int}(S_0) \neq \emptyset$, Assumption 8.1 holds automatically (c.f., Theorem 6.42 in [27]).

For any unit vector $d \in \mathcal{N}_{S_0}(z)$, we have $z + d \notin S_0$ and $\mathrm{P}_{S_0}(z + d) = z$. Consider $x = z + d$ in the proof of Theorem 3.1, then[2]

$$\max_{d_i\in\mathcal{N}_{S_i}(z),\ \|d_i\|=1,\ i\in\mathcal{I}} \langle d, d_i\rangle \geq \mathrm{inrad}(\mathcal{T}_{S_0})(z) \cdot \|(z + d) - z\| = \mathrm{inrad}(\mathcal{T}_{S_0})(z). \quad (8.7)$$

Now take the minimum over the unit vectors in $\mathcal{N}_{S_0}(z)$. Combined with (8.16) (in the proof of Theorem 8.2), we have[3]

$$\mathrm{reach}_{\{S_i\}_{i\in[m]}}(z) = \min_{d\in\mathcal{N}_{S_0}(z),\ \|d\|=1} \left\{ \max_{d_i\in\mathcal{N}_{S_i}(z),\ \|d_i\|=1,\ i\in\mathcal{I}} \langle d, d_i\rangle \right\} \geq \mathrm{inrad}(\mathcal{T}_{S_0})(z).$$

Hence Theorem 8.2 is both a generalization and an improvement of Theorem 3.1. See Example A.3 in Appendix A.6 for an example where $\mathrm{int}(S_0) \neq \emptyset$ and $\mathrm{reach}_{\{S_i\}_{i\in[m]}}(z) > \mathrm{inrad}(\mathcal{T}_{S_0})(z) > 0$.

Before proving Theorem 8.2, let us take a brief detour and study the connection between the functions g_S and r_S for generic closed convex sets in $\mathcal{E}$.[4] Due to (8.5), Lemma 8.1 can serve as a simplified special case of (8.14), a key step in the proof of Theorem 8.2:

Lemma 8.1. *For any closed convex set $S \subset \mathcal{E}$ and $e \in \mathrm{int}(S) \neq \emptyset$, we have*

$$g_S(e) = r_S(e). \quad (8.8)$$

Moreover, if $w \in \mathrm{bdy}(S)$ satisfies $\|w - e\| = r_S(e)$, then

$$\mathcal{N}_S(w) = \{t \cdot (w - e) \mid t \geq 0\}, \quad (8.9)$$

and $d = \frac{w-e}{\|w-e\|} \in \mathcal{N}_S(w)$ reaches the minimum in $g_S(e)$.

[2]Inequality (8.7) can be shown by using (3.13) instead of (3.14) in the last few steps of the proof of Theorem 3.1.

[3]Recall that f is the support function, defined in Section 2.2.

[4]The function g_S and the set Ω_S are defined in Section 2.2.

Proof. For any $w \in \mathrm{bdy}(S)$ and unit vector $d \in \mathcal{N}_{S_0}(w)$, since $w \in H(S,d)$, we get

$$\mathrm{dist}(e, H(S,d)) \le \|w - e\|. \tag{8.10}$$

Hence

$$\begin{aligned} r_S(e) &= \min_{w \in \mathrm{bdy}(S)} \|w - e\| \ge \min_{w \in \mathrm{bdy}(S)} \left\{ \min_{d \in \mathcal{N}_S(w), \|d\|=1} \mathrm{dist}(e, H(S,d)) \right\} \\ &= \min_{d \in \Omega_S} \mathrm{dist}(e, H(S,d)) \\ &= g_S(e). \end{aligned}$$

On the other hand, for any $d \in \Omega_S$, we have $S \subseteq H^-(S,d)$. Thus

$$r_S(e) \le \mathrm{dist}(e, \mathrm{bdy}(H^-(S,d))) = \mathrm{dist}(e, H(S,d)).$$

Taking the infimum over $d \in \Omega_S$, we get $r_S(e) \le g_S(e)$. Thus (8.8) holds.

Now assume that $w' \in \mathrm{bdy}(S)$ satisfies $\|w' - e\| = r_S(e)$. Then for any unit vector $d' \in \mathcal{N}_S(w')$, by (8.10), we have

$$g_S(e) \le \mathrm{dist}(e, H(S,d')) \le \|w' - e\| = r_S(e) = g_S(e).$$

Hence $w' = \mathrm{P}_{H(S,d')}(e)$, and $d' = \frac{w'-e}{\|w'-e\|}$. □

Proof of Theorem 8.2. Let us first study the support functions of $P_{\{S_i\}_{i\in[m]}}(z)$. Since $P_{\{S_i\}_{i\in[m]}}(z)$ is a compact set, for any unit vector $d \in \mathcal{E}$, there exists $w \in P_{\{S_i\}_{i\in[m]}}(z)$ such that

$$\langle w, d\rangle = \max_{w \in P_{\{S_i\}_{i\in[m]}}(z)} \langle w, d\rangle = f_{P_{\{S_i\}_{i\in[m]}}(z)}(d).$$

Hence $f_{P_{\{S_i\}_{i\in[m]}}(z)}$ is finite at any unit vector. Now consider any two vectors d and d'. Then for any $w \in P_{\{S_i\}_{i\in[m]}}(z)$, we have

$$\langle w, d\rangle - \langle w, d'\rangle \le \|w\|\|d - d'\| \le \|d - d'\|,$$

where the last inequality follows from the fact that $P_{\{S_i\}_{i\in[m]}}(z)$ is a subset of the unit ball. Consequently,

$$\begin{aligned} f_{P_{\{S_i\}_{i\in[m]}}(z)}(d) &= \max_{w\in P_{\{S_i\}_{i\in[m]}}(z)} \langle w, d\rangle \\ &\leq \max_{w\in P_{\{S_i\}_{i\in[m]}}(z)} \{\langle w, d'\rangle + \|d-d'\|\} \\ &= f_{P_{\{S_i\}_{i\in[m]}}(z)}(d') + \|d-d'\|. \end{aligned}$$

Similarly, $f_{P_{\{S_i\}_{i\in[m]}}(z)}(d') \leq f_{P_{\{S_i\}_{i\in[m]}}(z)}(d) + \|d-d'\|$, and we conclude that the support function of $P_{\{S_i\}_{i\in[m]}}(z)$ is 1-Lipschitz.

Thus there exists a unit vector $d' \in \mathcal{N}_{S_0}(z)$ such that d' is the minimizer of $f_{P_{\{S_i\}_{i\in[m]}}(z)}(d')$ over the compact set $\mathcal{N}_{S_0}(z) \cap \text{bdy}\left(\overline{B(\vec{0},1)}\right)$. By Assumption 8.1, we can write

$$d' = \sum_{i\in\mathcal{I}} \lambda_i d_i, \text{ where } \lambda_i \geq 0, \ \sum_{i\in\mathcal{I}} \lambda_i > 0, \ d_i \in \mathcal{N}_{S_i}(z), \ \|d_i\| = 1.$$

Hence by the definition of $P_{\{S_i\}_{i\in[m]}}(z)$,

$$\frac{d'}{\sum_{i\in\mathcal{I}}\lambda_i} = \sum_{i\in\mathcal{I}} \frac{\lambda_i}{\sum_{i\in\mathcal{I}}\lambda_i} d_i \in P_{\{S_i\}_{i\in[m]}}(z),$$

and

$$\min_{d\in\mathcal{N}_{S_0}(z),\ \|d\|=1} f_{P_{\{S_i\}_{i\in[m]}}(z)}(d) = f_{P_{\{S_i\}_{i\in\mathcal{I}}}}(d') \geq \left\langle d', \frac{d'}{\sum_{i\in\mathcal{I}}\lambda_i}\right\rangle = \frac{1}{\sum_{i\in\mathcal{I}}\lambda_i} > 0. \tag{8.11}$$

For any unit vector $w \in \mathcal{N}_{S_0}(z)$, define

$$\bar{w} := t\cdot w, \ t \geq 0 \text{ is the maximal number such that } t\cdot w \in P_{\{S_i\}_{i\in[m]}}(z).$$

Since $\vec{0} \in P_{\{S_i\}_{i\in[m]}}(z) \subseteq \overline{B(\vec{0},1)}$, we get $0 < \|\bar{w}\| \leq 1$. Note that $\bar{w} \in \text{bdy}\left(P_{\{S_i\}_{i\in[m]}}(z)\right)$. The definition of $\text{reach}_{\{S_i\}_{i\in[m]}}(z)$ implies

$$\text{reach}_{\{S_i\}_{i\in[m]}}(z) = \min_{w\in\mathcal{N}_{S_0}(z),\ \|w\|=1} \|\bar{w}\|. \tag{8.12}$$

Noting that

$$P_{\{S_i\}_{i\in[m]}}(z) \subset \mathcal{N}_{S_0}(z),$$

by (8.11), for any unit vector $d \in \mathcal{N}_{S_0}(z)$, one can easily show that there exists unit vector $w \in \mathcal{N}_{S_0}(z)$ such that

$$f_{P_{\{S_i\}_{i\in\mathcal{I}}}}(d) = \langle \bar{w}, d\rangle > 0. \tag{8.13}$$

With a proof similar to that of Lemma 8.1 (deferred to Appendix A.6) and (8.11), we have

$$\operatorname{reach}_{\{S_i\}_{i\in[m]}}(z) = \min_{d\in\mathcal{N}_{S_0}(z),\ \|d\|=1} f_{P_{\{S_i\}_{i\in[m]}}(z)}(d) > 0. \tag{8.14}$$

Recall that $P_{\{S_i\}_{i\in[m]}}(z)$ is the convex hull of the set $\{d \mid d \in \mathcal{N}_{S_i}(z),\ \|d\| = 1\}_{i\in\mathcal{I}} \cup \{\vec{0}\}\}$. For any unit vector $d \in \mathcal{N}_{S_0}(z)$, by (8.11) and the convexity of the function $\langle d, \cdot\rangle$, we see that

$$\langle d, \vec{0}\rangle < \max_{w\in P_{\{S_i\}_{i\in[m]}}(z)} \langle d, w\rangle = \max_{d_i\in\mathcal{N}_{S_i}(z),\ \|d_i\|=1,\ i\in\mathcal{I}} \langle d, d_i\rangle. \tag{8.15}$$

Hence by (8.14),

$$\begin{aligned}
\operatorname{reach}_{\{S_i\}_{i\in[m]}}(z) &= \min_{d\in\mathcal{N}_{S_0}(z),\ \|d\|=1} f_{P_{\{S_i\}_{i\in[m]}}(z)}(d) \\
&= \min_{d\in\mathcal{N}_{S_0}(z),\ \|d\|=1} \left\{ \max_{w\in P_{\{S_i\}_{i\in[m]}}(z)} \langle d, w\rangle \right\} \\
&= \min_{d\in\mathcal{N}_{S_0}(z),\ \|d\|=1} \left\{ \max_{d_i\in\mathcal{N}_{S_i}(z),\ \|d_i\|=1,\ i\in\mathcal{I}} \langle d, d_i\rangle \right\}.
\end{aligned} \tag{8.16}$$

Now consider any $x \notin S_0$ such that $\mathrm{P}_{S_0}(x) = z$. For any $i \in \mathcal{I}$ and $d_i \in \mathcal{N}_{S_i}(z)$ such that $\|d_i\| = 1$, by Lemma 2.5, we have

$$\langle x - z, d_i\rangle \le (\langle x - z, d_i\rangle)_+ = \operatorname{dist}(x, H^-(S_i, d_i)) \le \operatorname{dist}(x, S_i).$$

Hence

$$\begin{aligned}\max_{i\in\mathcal{I}} \operatorname{dist}(x, S_i) &\geq \max_{i\in\mathcal{I}} \left\{ \max_{\|d_i\|=1, d_i\in\mathcal{N}_{S_i}(z)} \langle x - z, d_i\rangle \right\} \\ &= \|x - z\| \left\{ \max_{i\in\mathcal{I}} \left\{ \max_{\|d_i\|=1, d_i\in\mathcal{N}_{S_i}(z)} \left\langle \frac{x-z}{\|x-z\|}, d_i \right\rangle \right\} \right\} \qquad (8.17) \\ &\geq \operatorname{reach}_{\{S_i\}_{i\in[m]}}(z)\|x - z\|,\end{aligned}$$

where the last inequality follows from $x - z \in \mathcal{N}_{S_0}(z)$ and (8.16). □

Proposition 8.1. *Fix $z \in \operatorname{bdy}(S_0)$. If Assumption 8.1 holds, then for any $x \notin S_0$ satisfying $z = \mathrm{P}_{S_0}(x)$, we have*

$$\gamma_0(x) - 1 \geq \frac{\operatorname{reach}_{\{S_i\}_{i\in[m]}}(z)}{\max_{i\in\mathcal{I}} \|z - e_i\|} \operatorname{dist}(x, S_0) \geq \frac{\operatorname{reach}_{\{S_i\}_{i\in[m]}}(z)}{\max_{i\in[m]} \|z - e_i\|} \operatorname{dist}(x, S_0).$$

Proof. For $i \in \mathcal{I}$, consider any unit vector $d \in \mathcal{N}_{S_i}(z)$. Then $\langle z, d\rangle = f_{S_i}(d)$ and

$$0 < f_{S_i}(d) - \langle e_i, d\rangle = \langle z - e_i, d\rangle \leq \|z - e_i\|.$$

By Proposition 2.2, we have

$$\begin{aligned}\gamma_i(x) - 1 &= \left(\max_{d\in\Omega_{S_i}} \left\{ \frac{\langle x - e_i, d\rangle}{f_{S_i}(d) - \langle e_i, d\rangle} \right\} \right)_+ - 1 \\ &\geq \max_{d\in\mathcal{N}_{S_i}(z), \|d\|=1} \left\{ \frac{\langle x - e_i, d\rangle}{f_{S_i}(d) - \langle e_i, d\rangle} - 1 \right\} \\ &= \max_{d\in\mathcal{N}_{S_i}(z), \|d\|=1} \left\{ \frac{\langle x - z, d\rangle}{f_{S_i}(d) - \langle e, d\rangle} \right\} \qquad (8.18) \\ &\geq \max_{d\in\mathcal{N}_{S_i}(z), \|d\|=1} \left\{ \frac{\langle x - z, d\rangle}{\|z - e_i\|} \right\},\end{aligned}$$

where the second equality follows from $f_{S_i}(d) = \langle z, d\rangle$.

Since $x - z \in \mathcal{N}_{S_0}(z)$, taking the maximum over $i \in \mathcal{I}$, we get

$$\begin{aligned}
\gamma_0(x) - 1 &\geq \max_{i \in \mathcal{I}} \{\gamma_i(x) - 1\} \\
&\geq \max_{i \in \mathcal{I}} \left\{ \max_{d \in \mathcal{N}_{S_i}(z), \|d\|=1} \left\{ \frac{\langle x - z, d \rangle}{\|z - e_i\|} \right\} \right\} \\
&\overset{(a)}{\geq} \frac{\max_{i \in \mathcal{I}} \{\max_{\|d_i\|=1, d_i \in \mathcal{N}_{S_i}(z)} \langle x - z, d_i \rangle\}}{\max_{i \in \mathcal{I}} \|z - e_i\|} \\
&\overset{(b)}{\geq} \frac{\operatorname{reach}_{\{S_i\}_{i \in [m]}}(z) \|x - z\|}{\max_{i \in \mathcal{I}} \|z - e_i\|},
\end{aligned}$$

where inequality (a) follows from the fact that $\max_{i \in \mathcal{I}} \{\max_{\|d_i\|=1, d_i \in \mathcal{N}_{S_i}(z)} \langle x - z, d_i \rangle\} \geq 0$, and inequality (b) is due to the last inequality in (8.17). □

Consequently, when Assumption 8.1 holds, γ_0 grows linearly outside S_0, and we can apply the algorithms discussed in Chapter 5 and 6 to solve the convex feasibility problem. To get the convergence rates of the algorithms, replace $\operatorname{inrad}(\mathcal{T}_{S_0}(z_k))$ by $\operatorname{reach}_{\{S_i\}_{i \in [m]}}(z_k)$ in Theorem 5.1, Theorem 5.2 and Theorem 6.1.

CHAPTER 9

SOLVING CONSTRAINED OPTIMIZATION PROBLEMS VIA RADIAL PROJECTIONS

Consider the constrained optimization problem

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & x \in S_i, \ \forall i \in [m], \end{aligned} \tag{C-Opt}$$

where $f : \mathcal{E} \to \mathbb{R}$ is an M-Lipschitz convex function, and $\{S_i\}_{i\in[m]} \subset \mathcal{E}$ are closed convex sets such that $S_0 = \cap_{i\in[m]} S_i \neq \emptyset$. We also assume that for each constraint set S_i, an interior point $e_i \in \text{int}(S_i)$ is available, and the γ functions are defined accordingly.

In [23], Renegar proposed radially projected subgradient methods for convex optimization problems with a single constraint. In contrast, (C-Opt) involves multiple constraints, and requires a different treatment.

Denote the set of optimal solutions to (C-Opt) by $X^\star$. Then $X^\star$ is a closed convex set. Let us introduce the following assumption:

Assumption 9.1. Given an constrained optimization problem (C-Opt), assume $X^\star \neq \emptyset$ and

$$\mathcal{N}_{S_1}(x^\star) + \cdots + \mathcal{N}_{S_m}(x^\star) + \text{cone}\left(\partial f(x^\star)\right) = \mathcal{N}_{X^\star}(x^\star). \tag{9.1}$$

at all $x^\star \in \text{bdy}(X^\star)$.

Remark. One can easily see that

$$\mathcal{N}_{S_1}(x^\star) + \cdots + \mathcal{N}_{S_m}(x^\star) + \text{cone}\left(\partial f(x^\star)\right) \subseteq \mathcal{N}_{X^\star}(x^\star)$$

always holds.

Fix $x^\star \in X^\star$. We immediately see that Assumption 9.1 is quite similar to Assumption 8.1, except that the new assumption also involves the subgradients of f at $x^\star$. Again, let $\mathcal{I} := \{i \in [m] \mid x^\star \in \mathrm{bdy}(S_i)\}$ denote the set of constraints active at $x^\star$. Define the set

$$Q_{(\mathrm{C-Opt})}(x^\star) := \mathrm{conv}\left(\left\{d \mid d \in \mathcal{N}_{S_i}(x^\star),\ \|d\| = \frac{1}{\|x^\star - e_i\|}\right\}_{i\in\mathcal{I}} \cup \partial f(x^\star) \cup \left\{\vec{0}\right\}\right) \tag{9.2}$$

and its reach

$$\mathrm{reach}_{(\mathrm{C-Opt})}(x^\star) := \max\left\{r \geq 0 \mid \left(\overline{B(\vec{0}, r)} \cap \mathcal{N}_{X^\star}(x^\star)\right) \subseteq Q_{(\mathrm{C-Opt})}(x^\star)\right\}. \tag{9.3}$$

Let $f^\star := f(x^\star)$ and

$$F(x) := \{\max\{\gamma_0(x) - 1, f(x) - f^\star\}\}. \tag{9.4}$$

Then we have the following result:[1]

Proposition 9.1. *Fix $x^\star \in \mathrm{bdy}(X^\star)$. If Assumption 9.1 holds, then*

$$0 < \mathrm{reach}_{(\mathrm{C-Opt})}(x^\star).$$

Moreover, for any $x \notin X^\star$ satisfying $x^\star = \mathrm{P}_{X^\star}(x)$, we have

$$F(x) \geq \mathrm{reach}_{(\mathrm{C-Opt})}(x^\star)\|x - x^\star\|. \tag{9.5}$$

Consider the constrained optimization problem

$$\begin{aligned} \min\ & 0 \\ s.t.\ & x \in S_i,\ \forall i \in [m]. \end{aligned} \tag{9.6}$$

Then $\partial f(x^\star) = \{\vec{0}\}$, and (9.6) is equivalent to the convex feasibility problem of finding $x \in S_0 = \cap_{i\in[m]} S_i$. Let $x^\star$ denote the feasible solution considered in

[1] Proposition 9.1 can be seen as a combination of Theorem 8.2 and Proposition 8.1 in the case of constrained optimization. Its derivation is also quite similar to the previous results, and is deferred to Appendix A.7.

Assumption 8.1. We immediately see that Assumption 8.1 and Assumption 9.1 are equivalent. Moreover, by the definition of $Q_{(\mathrm{C-Opt})}(x^\star)$, we have

$$\min_{i\in\mathcal{I}} \|x^\star - e_i\| \cdot \mathrm{reach}_{(\mathrm{C-Opt})}(x^\star) \leq \mathrm{reach}_{\{S_i\}_{i\in[m]}}(x^\star) \leq \max_{i\in\mathcal{I}} \|x^\star - e_i\| \cdot \mathrm{reach}_{(\mathrm{C-Opt})}(x^\star) \tag{9.7}$$

when the assumptions hold. Consequently, (9.5) can be seen as a strengthened (i.e., tighter) extension of Proposition 8.1.

9.1 Radial projection-based Polyak's rule when the optimal value is known

When the optimal value $f^\star$ is known, (C-Opt) is equivalent to

$$\min_{x\in\mathcal{E}} F(x) = \{\max\{\gamma_0(x) - 1, f(x) - f^\star\}\}. \tag{Radial-Opt}$$

Algorithm 5 applies Polyak's rule to solve (Radial-Opt).

Algorithm 5: Polyak's Rule via Radial Projections with Known Optimal Value

input : target accuracy $\epsilon > 0$, the optimal value $f^\star$ and an initial iterate $x_0 \in \mathcal{E}$
output: a point $\bar{x} \in \mathcal{E}$ satisfying $\gamma_0(\bar{x}) - 1 \leq \epsilon$ and $f(x) - f^\star \leq \epsilon$
initialization: let $k = 0$

while $F(x) > \epsilon$ **do**
  if $\gamma_0(x_k) - 1 > f(x_k) - f^\star$ **then**
    compute $g_k \in \partial\gamma_0(x_k)$;
    $x_{k+1} := x_k - \frac{\gamma_0(x_k)-1}{\|g_k\|^2} g_k$;
  else
    compute $g_k \in \partial f(x_k)$;
    $x_{k+1} := x_k - \frac{f(x_k)-f^\star}{\|g_k\|^2} g_k$;

return $\bar{x} := x_k$

When $S_0 \neq \emptyset$ (which implies $f^\star < \infty$), by applying the standard one-step analysis (5.2) repeatedly, one can show that Algorithm 5 terminates in

$$\left(\frac{\max\{1/r_0, M\} \cdot \operatorname{dist}(x_0, X^\star)}{\epsilon}\right)^2$$

iterations. Improved bounds can be obtained when Assumption 9.1 holds:

Theorem 9.2. *For* (C-Opt)*, if Assumption 9.1 holds and the optimal value $f^\star$ is known, then Algorithm 5 terminates in*

$$2\left(\frac{\max\{1/r_0, M\}}{\min_{x^\star \in X^\star} \operatorname{reach}_{(\mathrm{C-Opt})}(x^\star)}\right)^2 \log_2\left(\frac{\|x_0 - x^\star\| \max\{1/r_0, M\}}{\epsilon}\right)$$

iterations.

Remark. It is easy to see that in Theorem 9.2, one only needs to consider the reaches at the orthogonal projections of the iterates (instead of the entire boundary of $X^\star$). The same observation also applies to Theorem 9.2' and Theorem 9.3.

Proof. It is easy to see that F is $(\max\{1/r_0, M\})$-Lipschitz. Combined with its linear growth rate in (9.5), the proof follows from a standard subgradient method analysis similar to that of Theorem 5.1. □

In Algorithm 5, the objective function f and γ_0 share the same target accuracy ϵ. To allow different target accuracies for γ_0 and f, define

$$F_\eta(x) := \max\{\gamma_0(x) - 1, \eta(f(x) - f^\star)\}. \tag{9.8}$$

for $\eta > 0$. For any $\epsilon_0 > 0$ and $\epsilon > 0$, by replacing $F(x)$ with $F_{\left(\frac{\epsilon_0}{\epsilon}\right)}$ in Algorithm 5, we obtain an algorithm which allows users to set different target accuracies for γ_0 and f:

Algorithm 5': Polyak's Rule via Radial Projections with Known Optimal Value and Different Targets

input : target accuracies $\epsilon_0 > 0$, $\epsilon > 0$, the optimal value $f^\star$ and an initial iterate $x_0 \in \mathcal{E}$

output: a point $\bar{x} \in \mathcal{E}$ satisfying $\gamma_0(\bar{x}) - 1 \leq \epsilon_0$ and $f(x) - f^\star \leq \epsilon$

initialization: let $k = 0$

while $F_{\left(\frac{\epsilon_0}{\epsilon}\right)}(x_k) > \epsilon_0$ **do**

  if $\gamma_0(x_k) - 1 > (\frac{\epsilon_0}{\epsilon})(f(x_k) - f^\star)$ **then**

    $\cdots$

  else

    $\cdots$

return $\bar{x} := x_k$

In order to analyze the convergence rate of Algorithm 5', for any $\eta > 0$, define

$$Q_{(\mathrm{C-Opt},\eta)}(x^\star) := \operatorname{conv}\left(\left\{d \mid d \in \mathcal{N}_{S_i}(x^\star),\ \|d\| = \frac{1}{\|x^\star - e_i\|}\right\}_{i\in\mathcal{I}} \cup \{\eta \cdot g \mid g \in \partial f(x^\star)\} \cup \left\{\vec{0}\right\}\right)$$

and

$$\operatorname{reach}_{(\mathrm{C-Opt},\eta)}(x^\star) := \max\left\{r \geq 0 \mid \left(\overline{B(\vec{0},r)} \cap \mathcal{N}_{X^\star}(x^\star)\right) \subseteq Q_{(\mathrm{C-Opt},\eta)}(x^\star)\right\}. \tag{9.9}$$

When Assumption 9.1 holds, for any $0 < \eta_1 < \eta_2$, we have

$$\operatorname{reach}_{(\mathrm{C-Opt},\eta_1)}(x^\star) \leq \operatorname{reach}_{(\mathrm{C-Opt},\eta_2)}(x^\star) \leq \left(\frac{\eta_2}{\eta_1}\right) \operatorname{reach}_{(\mathrm{C-Opt},\eta_1)}(x^\star).$$

Similar to (9.5), we can show that for all $x \notin X^\star$ such that $\mathrm{P}_{X^\star}(x) = x^\star$,

$$F_\eta(x) \geq \operatorname{reach}_{(\mathrm{C-Opt},\eta)}(x^\star)\|x - x^\star\|. \tag{9.10}$$

Also note that by the definition of f_η, it is $\max\left\{\frac{1}{r_0}, \eta M\right\}$-Lipschitz. Like Theorem 9.2, we can show that

Theorem 9.2'. *For* (C-Opt)*, if Assumption 9.1 holds and the optimal value* $f^\star$ *is known, then Algorithm 5' terminates in*

$$2\left(\frac{\max\left\{\frac{1}{r_0},(\frac{\epsilon_0}{\epsilon})M\right\}}{\min_{x^\star\in X^\star}\operatorname{reach}_{(\mathrm{C-Opt},\frac{\epsilon_0}{\epsilon})}(x^\star)}\right)^2\log_2\left(\|x_0-x^\star\|\cdot\max\left\{\frac{1}{\epsilon_0 r_0},\frac{M}{\epsilon}\right\}\right)$$

iterations.

Remark. Note that for the function F_η, neither its growth rate (i.e., $\operatorname{reach}_{(\mathrm{C-Opt},\eta)}(x^\star)$) nor its Lipschitz constant (i.e., $\max\{\frac{1}{r_0},\eta M\}$) changes linearly in η. Hence Theorem 9.2' does not necessarily produce a larger iteration count if we fix ϵ_0 and decrease ϵ in Algorithm 5'.

9.2 Radial projection-based Polyak's rule when the optimal value is unknown

In practice, the optimal value $f^\star$ is usually unknown to the users. In this case, we can apply the idea of the parallel scheme introduced by Renegar and Grimmer in [24].[2] As its name suggests, the parallel scheme runs copies of subgradient methods with different step sizes in parallel.

For any $\eta>0$ and $c\in\mathbb{R}$, define

$$F_{(\eta,c)}(x):=\max\{\gamma_0(x)-1,\eta(f(x)-c)\}. \tag{9.11}$$

When $f^\star$ is unknown, we can only work with $F_{(\eta,c)}$ (instead of F_η), where c is an estimate of $f^\star$. Before introducing a paralleled version of Algorithm 5', let us first establish some properties of $F_{(\eta,c)}$:

[2]The analysis in [24] is not immediately applicable here, since the estimates of $f^\star$ enter our objective function (see (9.11)), which is not covered in [24].

Lemma 9.1. *When Assumption 9.1 holds, consider $\eta > 0$ and $c \in \mathbb{R}$ such that $c - f^\star \in [0, 3\eta^{-1}\epsilon_0)$. Then for any $x \in \mathcal{E}$ satisfying $F_{(\eta,c)}(x) > \epsilon_0$, we have*

$$F_{(\eta,c)}(x) > \frac{F_\eta(x)}{4} > 0. \tag{9.12}$$

Moreover, let $L^{-1}(\eta, c)$ denote the 0-sublevel set of $F_{(\eta,c)}$, then for $x \notin L^{-1}(\eta, c)$, we have

$$F_{(\eta,c)}(x) \geq \left(\frac{\min_{x^\star \in X^\star} \operatorname{reach}_{(\mathrm{C-Opt},\eta)}(x^\star)}{4} \right) \operatorname{dist}\left(x, L^{-1}(\eta, c)\right). \tag{9.13}$$

Proof. Since $c \geq f^\star$, we have

$$\begin{aligned}
\eta(f(x) - f^\star) &= \eta(f(x) - c) + \eta(c - f^\star) \\
&\leq \max\{\gamma_0(x) - 1, \eta(f(x) - c)\} + \eta(c - f^\star) \\
&= F_{(\eta,c)}(x) + \eta(c - f^\star).
\end{aligned}$$

Hence

$$\frac{F_\eta(x)}{F_{(\eta,c)}(x)} \leq 1 + \frac{\eta(c - f^\star)}{F_{(\eta,c)}(x)} < 1 + \frac{3\epsilon_0}{\epsilon_0} = 4.$$

Let $x^\star = \mathrm{P}_{X^\star}(x^\star)$. Note that the optimality of $x^\star$ implies $x^\star \in L^{-1}(\eta, c)$. By (9.12) and (9.10), we get

$$\frac{F_{(\eta,c)}(x)}{\operatorname{dist}\left(x, L^{-1}(\eta, c)\right)} \geq \frac{F_{(\eta,c)}(x)}{\|x - x^\star\|} \geq \frac{F_\eta(x)/4}{\|x - x^\star\|} \geq \frac{\min_{x^\star \in X^\star} \operatorname{reach}_{(\mathrm{C-Opt},\eta)}(x^\star)}{4}.$$

□

We refer the readers to Chapter 5 (if $\operatorname{int}(S_0) \neq \emptyset$) and Chapter 8 (if $\operatorname{int}(S_0) = \emptyset$) for the time required by Algorithm 6 to compute x_{init}. In the remainder of this section, we focus on the analysis of the algorithm after x_{init} has been obtained.

After computing x_{init}, Algorithm 6 essentially run $N + 1$ copies of Algorithm 5' with different estimates of $f^\star$ in parallel. Note that the copies in Algorithm 6

Algorithm 6: Parallel Polyak's Rule via Radial Projections

input : target accuracies $\epsilon_0 > 0$, $\epsilon > 0$, an initial iterate $x_0 \in \mathcal{E}$ and integer $N \geq 0$ (number of copies)

initialization: let $k_n = 0$ for all integer n such that $-1 \leq n \leq N$;

run Algorithm 1 with target accuracy ϵ_0 and initial iterate x_0 to get x_{init} ;
// $\gamma_0(x_{\text{init}}) \leq 1 + \epsilon_0$

let $\bar{x} = x_{\text{init}}$, $\tilde{f}^\star = f(x_{\text{init}})$; // the current candidate iterate and value

for $-1 \leq n \leq N$ **do**

  $x_0^n = x_{\text{init}}$, $f_n = f(x_{\text{init}}) - 2 \cdot 2^n \epsilon$; // initiate copy n with x_{init}

repeat

  for $-1 \leq n \leq N$ **do**

    if $F_{(\frac{\epsilon_0}{2^n \epsilon}, f_n)}(x) > \epsilon_0$ **then** // update the iterate

      if $\gamma_0(x_{k_n}^n) - 1 > (\frac{\epsilon_0}{2^n \epsilon})(f(x_{k_n}^n) - f_n)$ **then**

        compute $g_k \in \partial \gamma_0(x_{k_n}^n)$;

        $x_{k_n+1}^n = x_{k_n}^n - \frac{\gamma_0(x_{k_n}^n) - 1}{\|g_k\|^2} g_k$;

      else

        compute $g_k \in \partial f(x_{k_n}^n)$;

        $x_{k_n+1}^n = x_{k_n}^n - \frac{f(x_{k_n}^n) - f_n}{\|g_k\|^2} g_k$;

      $k_n = k_n + 1$;

    else

      $x_0^n = x_{k_n}^n$, $f_n = f(x_{k_n}^n) - 2 \cdot 2^n \epsilon$, $k_n = 0$; // restart copy n with $x_{k_n}^n$

      if $f(x_{k_n}^n) < \tilde{f}^\star$ **then**

        $\bar{x} = x_{k_n}^n$, $\tilde{f}^\star = f(x_{k_n}^n)$; // update the candidate iterate and value

  for $-1 \leq n \leq N$ **do**

    if $\tilde{f}^\star - f_n \leq 2^n \epsilon$ **then**

      $x_0^n = \bar{x}$, $f_n = \tilde{f}^\star - 2 \cdot 2^n \epsilon$, $k_n = 0$; // restart copy n with $\bar{x}$

are always restarted (or initiated) at $x \in \mathcal{E}$ satisfying $\gamma_0(x) - 1 \leq \epsilon_0$. Moreover, if copy n is restarted with an iterate x, then

$$f(x) \leq f_n + \left(\frac{2^n\epsilon}{\epsilon_0}\right)\epsilon_0 = f_n + 2^n\epsilon = f(x_0^n) - 2^n\epsilon.$$

Hence after each restart, copy n improves the objective value by at least $2^n\epsilon$ (which is also the task given to copy n in the parallel scheme in [24]), while also satisfying the requirement on feasibility imposed by γ_0.

In order to apply the analysis of the parallel scheme from [24], we need to provide an upper bound on the number of updates required by each copy to restart (c.f. Lemma 1 in [24]):

Lemma 9.2. *In Algorithm 6, if Assumption 9.1 holds and copy n is initiated or restarted at x_0^n satisfying $\gamma_0(x_0^n) - 1 \leq \epsilon_0$ and $f(x_0^n) - f^\star \in [2 \cdot 2^n\epsilon, 5 \cdot 2^n\epsilon)$, then copy n will be restarted after at most*

$$2\left(\frac{\max\left\{\frac{1}{r_0}, M(\frac{\epsilon_0}{2^n\epsilon})\right\}}{\min_{x^\star \in X^\star} \operatorname{reach}_{(\mathrm{C-Opt}, \frac{\epsilon_0}{2^n\epsilon})}(x^\star)}\right)^2 \cdot \log_2\left(\frac{8}{\min_{x^\star \in X^\star} \operatorname{reach}_{(\mathrm{C-Opt}, \frac{\epsilon_0}{2^n\epsilon})}(x^\star)} \cdot \max\left\{\frac{1}{\epsilon_0 r_0}, \frac{M}{2^n\epsilon}\right\}\right)$$

additional updates.

Remark. It should be noted that while Lemma 1 in [24] does not require

$$f(x_0^n) - f^\star < 5 \cdot 2^n\epsilon, \tag{9.14}$$

in the proof of Corollary 8 in [24], a copy only enters the analysis of the parallel scheme if it is restarted at $x_0^n \in \mathcal{E}$ satisfying (9.14). Hence Lemma 9.2 is sufficient for our purpose here.

Proof. When x_0^n satisfies the conditions, we get

$$f_n - f^\star = (f(x_0^n) - f^\star) - 2 \cdot 2^n\epsilon \in [0, 3 \cdot 2^n\epsilon). \tag{9.15}$$

Let $\eta = \frac{\epsilon_0}{2^n\epsilon}$ and $c = f_n$ in (9.13), we see that $F_{\left(\frac{\epsilon_0}{2^n\epsilon}, f_n\right)}$ has a growth rate of $\left(\min_{x^\star \in X^\star} \operatorname{reach}_{(\text{C}-\text{Opt}, \frac{\epsilon_0}{2^n\epsilon})}(x^\star)/4\right)$ outside its 0-sublevel set. Also note that $F_{\left(\frac{\epsilon_0}{2^n\epsilon}, f_n\right)}(x)$ is $\max\{\frac{1}{r_0}, M(\frac{\epsilon_0}{2^n\epsilon})\}$-Lipschitz. The proof of this result follows from a standard subgradient method analysis similar to that of Theorem 5.1. □

We next present Theorem 9.3, which is essentially an application of Theorem 2 in [24]. We refer the readers to [24] for the ideas and details of its derivation.

Theorem 9.3. *For* (C-Opt)*, if Assumption 9.1 holds and* $f(x_{\text{init}}) - f^\star < 5 \cdot 2^N \epsilon$*, then after* x_{init} *has been computed, Algorithm 6 produces* $\bar{x} \in \mathcal{E}$ *satisfying* $\gamma_0(\bar{x}) - 1 \leq \epsilon_0$ *and* $f(\bar{x}) - f^\star < 2 \cdot \epsilon$ *with at most*

$$(N_0+1)+3\sum_{n=-1}^{N_0} 2\left(\frac{\max\left\{\frac{1}{r_0}, M(\frac{\epsilon_0}{2^n\epsilon})\right\}}{\min_{x^\star \in X^\star} \operatorname{reach}_{(\text{C}-\text{Opt}, \frac{\epsilon_0}{2^n\epsilon})}(x^\star)}\right)^2 \cdot \log_2\left(\frac{8}{\min_{x^\star \in X^\star} \operatorname{reach}_{(\text{C}-\text{Opt}, \frac{\epsilon_0}{2^n\epsilon})}(x^\star)} \cdot \max\left\{\frac{1}{\epsilon_0 r_0}, \frac{M}{2^n\epsilon}\right\}\right)$$

first-order oracle calls of γ_0*, where* N_0 *is the smallest integer satisfying both* $f(\bar{x}) - f^\star < 5 \cdot 2^{N_0}\epsilon$ *and* $N_0 \geq -1$*.*

Similar to the remark following Theorem 9.2', without further knowledge regarding the structure of the problem, there is no straightforward comparison between the rate in Theorem 9.3 (which involves $F_{(\frac{\epsilon_0}{2^n\epsilon})}$, where $-1 \leq n \leq N_0$) and the rate in Theorem 9.2' (which only involves $F_{(\frac{\epsilon_0}{\epsilon})}$).

Part II

Margin Maximization of the Intersection of Convex Cones

CHAPTER 10

INTRODUCTION

Part II applies many results in Part I to the conic setting. For completeness, we prove everything from scratch in this part, which could lead to repetition in some cases, especially in the earlier chapters.

Given finitely many closed convex cones, let K_0 denote their intersection. The *margin maximization problem* concerns finding points with large margins in K_0. Consider the perceptron problem [28], which studies the intersection of finitely many half-spaces. Let $\{e_i\}_{i\in[m]}$ denote normal vectors of the half-spaces with unit-norm. Define the *margin function* $\omega_0(x) := \min_{i\in[m]}\langle x, e_i\rangle$. In the perceptron setting, the margin maximization problem is to solve

$$\max_{x\in\mathcal{E}} \frac{\omega_0(x)}{\|x\|}. \tag{10.1}$$

where $\|\cdot\|$ is the Euclidean norm. Note that given any $c > 0$, (10.1) is equivalent to

$$\begin{aligned} \min \quad & \tfrac{1}{2}\|x\|^2 \\ s.t. \quad & \langle x, e_i\rangle \geq c, \quad \forall i \in [m]. \end{aligned}$$

Hence in the perceptron setting, (10.1) can be seen as the support vector machine problem without intercepts. A plethora of (primal-)dual methods involving quadratic programs have been proposed for the support vector machine problem. See [29] for a review.

Primal first-order methods for (10.1) been studied by Gentile [9] and Kowalczyk [15]. For a relative accuracy ϵ, the Approximate Large Margin Algorithm (ALMA) proposed by Gentile terminates after at most

$$O\left(\left(\frac{1}{\epsilon \operatorname{inrad}(K_0)}\right)^2\right)$$

supgradient oracle calls of ω_0, where the inradius of K_0 measures the width of K_0. Kowalczyk [15] proposed another method with the same convergence rate. Both algorithms are *first-order methods*, and came out around 2000.

In the next two chapters, we first generalize the margin function to the the intersection of generic closed convex cones. Some useful properties of the margin function are introduced. Surprisingly, even for the intersection of generic closed convex cones, we have that the margin function decreases "sharply" with respect to its suplevel sets. After showing that the decerease rate becomes quadratic when restricted to the unit ball, we present a two-stage algorithm to solve the margin maximization problem. The first stage is a generalized modification of the ALMA algorithm in [9], while the second stage is an application of the parallel scheme proposed by Renegar and Grimmer [24]. This new algorithm finds an ϵ-optimal solution with at most

$$O\left(\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\left(\frac{1}{r_0\,\mathrm{inrad}(K_0)}\right)^2\right)$$

supgradient oracle calls of the margin function.

Our algorithm relies on reference points in the interior of the individual convex cones, and r_0 measures the centrality of these reference points with respect to the cones. In the perceptron case, by taking the normal vector with unit-norm as the reference point for each half-space considered, we get $r_0 = 1$, and the new algorithm improves the existing results.

It should be noted that while ALMA terminates with an output satisfying the desired relative accuracy, our new algorithm never terminates by itself.

We close this paper with two numerical examples. In the first example, we demonstrate how the new algorithm improves the convergence rate of the ALMA

in the perceptron setting. The second example studies the intersection of second-order cones, which is not covered by the existing literature.

CHAPTER 11

DEFINITION AND PROPERTIES OF ω FOR A SINGLE SET

Let $\mathcal{E}$ denote a finite-dimensional Euclidean space endowed with an inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|\cdot\|$. Consider a closed convex cone $K \subset \mathcal{E}$. We assume K has nonempty interior, and a vector $e \in \text{int}(K)$ is given. By rescaling, we may assume

$$\|e\| = 1.$$

For all $x \in \mathcal{E}$, define the margin function

$$\omega(x) := \max\{t \mid x \in K + t \cdot e\} = \max\{t \mid x - t \cdot e \in K\}.$$

The margin function ω can be evaluated with high precision using a line-search. For more details, see the remark following Proposition 11.5.

In [23], Renegar defined a function $\lambda_{\min}$, which is basically equivalent to ω. We adopt a different notation in this work, since later ω will be extended to multiple cones. While the Lipschitz continuity and characterization of its supgradients have been studied in [23], here we study the geometry of the suplevel sets of ω (Proposition 11.1) and discover the linear decrease of ω with respect to distances to those sets (Lemma 11.2). These results leads to the decrease rate of the ω function for multiple cones (Corollary 12.1) in the next chapter, which is an entirely novel addition to the literature. We include the study of the Lipschitzness and supgradient of ω for completeness.

The following result shows that the suplevel sets of ω are simply translations of K:

Proposition 11.1. *The margin function ω is finite for all $x \in \mathcal{E}$. For any $t \in \mathbb{R}$,*

we have

$$\omega(x) \geq t \iff x \in K(t) := K + t \cdot e.$$

Consequently, we get

$$x \in \mathrm{bdy}(K(\omega(x))). \tag{11.1}$$

for all $x \in \mathcal{E}$.

Remark. For general closed convex cones, the value of $\omega(x)$ cannot be computed exactly. However, similar to the function $\lambda_{\min}$ in [23], the value of $\omega(x)$ can be approximated accurately. All that is required is accurately approximating where the line $\{x - t \cdot e \mid t \in \mathbb{R}\}$ intersects the boundary of K. To do so, one can perform bisection on the half-lines $\{x - t \cdot e \mid t \leq 0\}$ and $\{x - t \cdot e \mid t \geq 0\}$ to find out (whether and) where they intersect the boundary of K. Since $\omega(x) \in \mathbb{R}$, at least one of the two half-lines will intersect the boundary of K, and the value of ω can be estimated accordingly. For most closed convex cones, this can be accomplished far more easily than traditional orthogonal projections.

As for special cones such as half-spaces and second-order cones, the margin function can be evaluated exactly by utilizing the properties of these cones. See the numerical chapter for some brief examples.

In what follows, we assume $\omega(x)$ can be computed exactly.

Proof. Since $e \in \mathrm{int}(K)$, we have $-e \in \mathrm{int}(K^\circ)$, where $K^\circ := \{y \in \mathcal{E} \mid \langle y, x\rangle \leq 0,\ \forall x \in \mathcal{E}\}$ denotes the polar cone of K. Consequently, for any $x \in \mathcal{E}$, we have

$$\lim_{t\to\infty} \frac{x - t \cdot e}{\|x - t \cdot e\|} \to -e \in \mathrm{int}(K^\circ),\ \lim_{t\to-\infty} \frac{x - t \cdot e}{\|x - t \cdot e\|} \to e \in \mathrm{int}(K).$$

Hence there exists $\infty > t_1 > t_2 > -\infty$ such that

$$x - t_1 \cdot e \in K^\circ,\quad x - t_2 \cdot e \in K,$$

and $\omega(x) \in (t_2, t_1) \subset \mathbb{R}$. The attainability of the maximization in the definition of $\omega(x)$ also follows from the closedness of K.

Thus $x - \omega(x) \cdot e \in K$ for all $x \in \mathcal{E}$. Since K is a convex cone and $x \in K$, for any $t \leq \omega(x)$, we get

$$x - t \cdot e = (x - \omega(x) \cdot e) + (\omega(x) - t) \cdot e \in K.$$

Hence $x \in K(t)$ for all $t \leq \omega(x)$. □

By Proposition 11.1, $K = K(0)$, and we have that $x \in K$ if and only if $\omega(x) \geq 0$. In general, we see that the higher the value of $\omega(x)$, the more internally x is placed in K with respect to e. The margin function ω is closely related to the Minkowski functional (gauge):

Proposition 11.2. *Let*

$$p_{(e,K)}(x) := \inf\{r > 0 \mid x \in e + r(K - e)\}$$

denote the Minkowski functional of K with e replacing the origin. For any x satisfying $\omega(x) \leq 1$, we have

$$\omega(x) = 1 - p_{(e,K)}(x).$$

Remark. Note that $p_{(e,K)}(x) = 0$ for all $x \in K(1)$, while $\omega(K(1)) = [1, \infty)$. We see that the margin function ω is in fact a refinement of the Minkowski functional.

Proof. When $\omega(x) = 1$, by Proposition 11.1, we have

$$x \in K(1) \subset K(1 - r) = K + (1 - r)e = e + r(K - e)$$

for all $r > 0$. Hence

$$p_{(e,K)}(x) = 0 = 1 - \omega(x).$$

When $\omega(x) < 1$, since K is a cone, we have

$$x \in e + r(K - e) \iff x \in K + (1 - r)e \iff x \in K(1 - r)$$

for all $r > 0$. The rest of the proof follows from Proposition 11.1 and the definition of ω and $p_{(e,K)}$. □

For any $z \in K$, define

$$r_K(z) := \max\left\{r \geq 0 \mid \overline{B(z, r)} \subset K\right\}.$$

Then $\omega(z)$ leads to a lower bound on $r_K(z)$ for all $z \in K$:

Proposition 11.3. *For all $z \in K$, we have*

$$r_K(z) \geq \omega(z) r_K(e).$$

Proof. For all $z \in K$, by Proposition 11.1, we have $\omega(x) \geq 0$ and $x \in K + \omega(x) \cdot e$. Hence for all $v \in \overline{B(\vec{0}, r_K(e))}$,

$$\begin{aligned} x + \omega(x) \cdot v &\in (K + \omega(x) \cdot e) + \omega(x) \cdot v \\ &= K + \omega(x)(e + v) \\ &\subseteq K + \omega(x) \cdot \overline{B(e, r_K(e))}. \end{aligned}$$

By the definition of r_K and the fact that K is a convex cone, we get

$$x + \omega(x) \cdot v \in K + \omega(x) \cdot \overline{B(e, r_K(e))} \subseteq K + \omega(x) \cdot K = K,$$

and $\overline{B(x, \omega(x) r_K(e))} \subseteq K$. □

Before discussing the first-order properties of ω, let us introduce another characterization of the margin function:

Proposition 11.4. *For any $x \in \mathcal{E}$, we have*

$$\omega(x) = \min\left\{\langle x, d\rangle \mid d \in K^\star,\ \langle e, d\rangle = 1\right\},$$

where $K^\star := \{y \in \mathcal{E} \mid \langle y, x\rangle \geq 0,\ \forall x \in \mathcal{E}\}$ denotes the dual cone of K. (Hence ω is a positively homogeneous concave function.)

Proof. For any non-zero vector d, since $e \in \operatorname{int}(K)$, we have

$$e - \frac{r_K(e)}{\|d\|} \cdot d \in K.$$

Thus if $d \in K^\star$,

$$\langle e, d\rangle = \left\langle e - \frac{r_K(e)}{\|d\|} \cdot d, d \right\rangle + \frac{r_K(e)}{\|d\|}\langle d, d\rangle \geq r_K(e)\|d\|. \tag{11.2}$$

Since $K^{\star\star} = K$, we have

$$\begin{aligned} x \in K &\iff \langle x, d\rangle \geq 0,\ \forall d \in K^\star \\ &\iff \left\langle x, \frac{d}{\langle e, d\rangle} \right\rangle \geq 0,\ \forall d \in K^\star,\ d \neq \vec{0} \\ &\iff \langle x, d\rangle \geq 0,\ \forall d \in K^\star,\ \langle e, d\rangle = 1. \end{aligned} \tag{11.3}$$

Note that for any $t \in \mathbb{R}$, $x \in \mathcal{E}$ and $d \in K^\star$ such that $\langle e, d\rangle = 1$, we have

$$\langle x - t \cdot e, d\rangle = \langle x, d\rangle - t\langle e, d\rangle = \langle x, d\rangle - t.$$

By (11.3), we have

$$x - t \cdot e \in K \iff \langle x, d\rangle \geq t,\ \forall d \in K^\star,\ \langle e, d\rangle = 1.$$

Taking the minimum over all eligible dual vectors on the right-hand side, the statement follows from the definition of ω. □

The supgradient sets of ω have the following characterizations:

Proposition 11.5. *For any $x \in \mathcal{E}$, let $\partial\omega(x)$ denote the set of supgradients of ω at x. Then we have*

$$\begin{aligned}\partial\omega(x) &= \{d \mid \langle x, d\rangle = \omega(x),\ d \in K^\star,\ \langle e, d\rangle = 1\} \\ &= \left\{d \mid d \in -N_{K(\omega(x))}(x),\ \langle e, d\rangle = 1\right\} \\ &= \{d \mid d \in -N_K(x - \omega(x)\cdot e),\ \langle e, d\rangle = 1\},\end{aligned}$$

where $N_{K(\omega(y))}(y)$ denotes the set of normal vectors of $K(\omega(y))$ at y.

Remark. To get a subgradient of $\gamma(x)$, it suffices to have an oracle which generates non-zero normal vectors to K at its boundary points.

Proof. Let $\mathcal{D} := \{d \mid d \in K^\star,\ \langle e, d\rangle = 1\}$ denote the set of vectors considered in the right-hand side of Proposition 11.4. Then by by (11.2), we get

$$\|d\| \le \frac{\langle e, d\rangle}{\|r_K(e)\|} = \frac{1}{r_K(e)} \tag{11.4}$$

for all $d \in \mathcal{D}$. Hence $\mathcal{D}$ is a bounded set. Moreover, since the Euclidean space $\mathcal{E}$ is finite-dimensional and $\mathcal{D} = K^\star \cap \{d \mid \langle e, d\rangle = 1\}$, we see that $\mathcal{D}$ is a compact convex set. Hence for any $x \in \mathcal{E}$, the set $\mathcal{D}_x := \{d \mid \langle x, d\rangle = \omega(x),\ d \in \mathcal{D}\}$ is nonempty.

To show the first equation in the statement, first consider any $d' \in \mathcal{D}_x$. By Proposition 11.4, for any $y \in \mathcal{E}$, we have

$$\omega(y) - \omega(x) = \min\{\langle y, d\rangle \mid d \in \mathcal{D}\} - \langle x, d'\rangle \le \langle y, d'\rangle - \langle x, d'\rangle = \langle y - x, d'\rangle,$$

and $d' \in \partial\omega(x)$. On the other hand, first note that by the convexity and compactness of $\mathcal{D}$ and the Hahn-Banach Theorem, for any $\tilde{d} \notin \mathcal{D}$, there exists $x' \in \mathcal{E}$ such that

$$\left\langle x', \tilde{d}\right\rangle < \min\{\langle x', d\rangle \mid d \in \mathcal{D}\} = \omega(x').$$

Also note that by the homogeneity and concavity of ω, we have

$$\omega(x+x') - \omega(x) = 2\omega\left(\frac{x+x'}{2}\right) - \omega(x) \geq (\omega(x) + \omega(x')) - \omega(x) = \omega(x').$$

Hence

$$\left\langle x+x', \tilde{d}\right\rangle - \left\langle x, \tilde{d}\right\rangle = \left\langle x', \tilde{d}\right\rangle < \omega(x+x') - \omega(x),$$

and we get $\partial\omega(x) \subseteq \mathcal{D}$. For any $d'' \in \mathcal{D} \setminus \mathcal{D}_x$, we have $\langle x, d''\rangle > \omega(x)$. By the homogeneity of ω, we get

$$\omega\left(\frac{x}{2}\right) - \omega(x) = -\frac{\omega(x)}{2} > \left\langle -\frac{x}{2}, d''\right\rangle = \left\langle \frac{x}{2} - x, d''\right\rangle,$$

and $d'' \notin \partial\omega(x)$. We conclude that $\partial\omega(x) \subseteq \mathcal{D}_x$.

To prove the second equation in the statement, let $g \in \partial\omega(x)$. Then for all $y \in K(\omega(x))$, we have

$$\langle g, y - x\rangle \geq \omega(y) - \omega(x) \geq 0.$$

Hence $g \in -N_{K(\omega(x))}$. On the other hand, since $K(\omega(x)) = K + \omega(x)\cdot e$, when $d' \in -N_{K(\omega(x))}(x) \subseteq K^\star$, we get

$$d' \in -N_{K(\omega(x))}(x) = -N_K(x - \omega(x)\cdot e).$$

Since K is a cone, we have[1]

$$\langle x - \omega(x)\cdot e, d'\rangle = 0.$$

If d' also satisfies $\langle e, d'\rangle = 1$, we get

$$\langle x, d'\rangle = \langle x - \omega(x)\cdot e, d'\rangle + \omega(x)\langle e, d'\rangle = \omega(x).$$

Hence the second equation in the statement also holds. The last equation in the statement follows from $K(\omega(x)) = K + \omega(x)\cdot e$. □

[1]To see this, first note that $d' \in K^\star$ implies $\langle x - \omega(x)\cdot e, d'\rangle \geq 0$. If $\langle x - \omega(x)\cdot e, d'\rangle > 0$, then $\langle x - \omega(x)\cdot e, d'\rangle > 0 = \langle \vec{0}, e\rangle$, contradicting our assumption $d' \in -N_K(x - \omega(x)\cdot e)$. Hence we conclude that $\langle x - \omega(x)\cdot e, d'\rangle = 0$.

Lemma 11.1. *The margin function ω is $\frac{1}{r_K(e)}$-Lipschitz.*

Proof. For any $x, y \in \mathcal{E}$, consider $g \in \partial\omega(x)$. Then by (11.4), we get $\|g\| \leq \frac{1}{r_K(e)}$. Hence

$$\omega(y) - \omega(x) \leq \langle y - x, g \rangle \leq \frac{\|y - x\|}{r_K(e)}.$$

One can also prove $\omega(x) - \omega(y) \leq \frac{\|y-x\|}{r_K(x)}$ with a similar argument. □

Lemma 11.2. *For any $x \in \mathcal{E}$ and $t \geq \omega(x)$, we have*

$$t - \omega(x) \geq \operatorname{dist}(x, K(t)).$$

Proof. By Proposition 11.1, we have

$$K(t) = K(\omega(x)) + (t - \omega(x)) \cdot e.$$

Since $x \in K(\omega(x))$, we get $x + (t - \omega(x)) \cdot e \in K(t)$. Since $\|e\| = 1$, we have

$$\operatorname{dist}(x, K(t)) \leq \|(t - \omega(x)) \cdot e\| = (t - \omega(x))\|e\| = t - \omega(x).$$

□

By assuming $\|e\| = 1$, we always have $r_K(e) \leq 1$. Combining Lemma 11.1 and Lemma 11.2, we see that for any $x \in \mathcal{E}$ and $t \geq \omega(x)$,

$$\frac{\operatorname{dist}(K(t), x)}{r_K(e)} \geq t - \omega(x) \geq \operatorname{dist}(K(t), x). \tag{11.5}$$

Hence $t - \omega(x)$ is an approximation of $\operatorname{dist}(K(t), x)$, and a bigger value of $r_K(e)$ leads to a better approximation. This motivates us to introduce the notion of *inradius* [11] (also called the width of K in the perceptron literature [5, 20]): Given any closed convex cone $K \subseteq \mathcal{E}$, define

$$\operatorname{inrad}(K) := \max_{z \in K \setminus \{\vec{0}\}} \frac{r_K(z)}{\|z\|} \tag{11.6}$$

$$= \max_{z \in K,\ \|z\| \leq 1} r_K(z). \tag{11.7}$$

Then

$$\text{inrad}(K) \leq 1$$

for all $K \subset \mathcal{E}$. The unique optimal solution to (11.7) is called the *center* of K. The center of K is the best reference point for defining the ω function, in order to make the left inequality in (11.5) as tight as possible.

As an example, for $d \in \mathcal{E} \setminus \{\vec{0}\}$, consider the half-space

$$\{x \mid \langle x, d \rangle \geq 0\}.$$

It is easy to see that the center of this half-space is the unit-vector $d/\|d\|$.

CHAPTER 12

DEFINITION AND PROPERTIES OF ω FOR MULTIPLE SETS

Now consider a finite number of sets $\{K_i\}_{i\in[m]}$, where $m > 1$ and $K_i \subset \mathcal{E}$. For $i \in [m]$, assume e_i is a known interior point of K_i satisfying $\|e_i\| = 1$. Let ω_i denote the ω function defined with respect to K_i and e_i. Define

$$K_0 := \cap_{i\in[m]} K_i$$

and the margin function

$$\begin{aligned}\omega_0(x) :&= \min_{i\in[m]} \omega_i(x) \\ &= \min\left\{\langle x, d\rangle \mid d \in K_i^\star,\ \langle e_i, d\rangle = 1,\ i \in [m]\right\}. \end{aligned} \tag{12.1}$$

Here the second inequality follows from Proposition 11.4. Clearly, ω_0 is a positively homogeneous concave function. Throughout the rest of this work, we assume

$$\operatorname{int}(K_0) \neq \emptyset. \tag{12.2}$$

Let

$$r_i := r_{K_i}(e_i),\ \forall i \in [m], \quad r_0 := \min_{i\in[m]} r_i.$$

By Proposition 11.3, we get

$$r_{K_0}(z) \geq \min_{i\in[m]} \omega_i(z) r_i \geq \omega(z) r_0 \tag{12.3}$$

for all $z \in K_0$. Since ω_i is $1/r_i$-Lipschitz continuous for all $i \in [m]$ (Lemma 11.1), by the definition of ω_0 and r_0, we have that ω_0 is $1/r_0$-Lipschitz continuous. By (12.1), we also have

$$\partial\omega_i(x) \subseteq \partial\omega_0(x) \tag{12.4}$$

for all $x \in \mathcal{E}$ and $i \in [m]$ such that $\omega_i(x) = \omega_0(x)$.

When

$$K_i = \{x \mid \langle x, e_i \rangle \geq 0\}, \quad \|e_i\| = 1, \quad \forall i \in [m],$$

K_0 is the intersection of a collection of half-spaces, and we recover the perceptron setting. In this case, $K_i^\star = \{te_i \mid t \geq 0\}$ and

$$\omega_0(x) = \min\{\langle x, e_i \rangle \mid i \in [m]\} = r_{K_0}(x) \tag{12.5}$$

for all $x \in K_0$. Hence in the perceptron setting, ω_0 reflects the "centrality" of a point in K_0 by measuring its distance to the boundary of K_0. For the intersection of generic closed convex cones, ω_0 can be thought of as the margin function of K_0 defined with respect to reference points $\{e_i\}_{i \in [m]}$.

For any $t \in \mathbb{R}$ and $i \in [m]$, we let $K_i(t) := K_i + t \cdot e_i$. Also define

$$K_0(t) := \cap_{i \in [m]} K_i(t) = \{x \mid \omega_0(x) \geq t\},$$

where the second equation follows from Proposition 11.1. Then $K_0 = K_0(0)$.

Proposition 12.1. *When* $\mathrm{int}(K_0) \neq \emptyset$*, we have* $K_0(t) \neq \emptyset$ *for all* $t \in \mathbb{R}$*.*

Proof. Let $z \in \mathrm{int}(K_0)$. For any $i \in [m]$, since $z \in \mathrm{int}(K_i)$ we get $\omega_i(z) > 0$ by Proposition 11.1. Hence $\omega_0(z) = \min_{i \in [m]} \omega_i(z) > 0$. Note that ω_0 is positively homogeneous, we have $\omega_0(s \cdot z) = s\omega(z)$ for all $s \geq 0$. Consequently,

$$\lim_{s \to \infty} \omega_0(s \cdot z) = \infty,$$

and $K_0(t) \neq \emptyset$ for all $t \in \mathbb{R}$. □

Lemma 11.2 establishes lower bounds on the decrease rate of the ω function for a single cone with respect to its suplevel sets. In order to derive a similar result for the ω_0 function for multiple cones, we first present the following result:

Proposition 12.2. *For any $x \in \mathcal{E}$ and $t \in \mathbb{R}$ satisfying $t \geq \omega_0(x)$, we have*

$$\operatorname{dist}(x, K_0(t)) \leq \frac{\max_{i\in[m]} \operatorname{dist}(x, K_i(t))}{\operatorname{inrad}(K_0)}.$$

Proof. Given $x \in \mathcal{E}$ and $t \in \mathbb{R}$ satisfying $t \geq \omega_0(x)$, let $z = \mathrm{P}_{K_0(t)}(x)$ denote the orthogonal projection of x onto $K_0(t)$. Then by Corollary 23.8.1 in [26], we have

$$x - z \in N_{K_0(t)}(z) = \sum_{i\in[m]} N_{K_i(t)}(z).$$

Hence we can write

$$x - z = \textstyle\sum_{i\in[m]} \lambda_i d_i, \quad \text{where } \lambda_i \geq 0, \quad d_i \in N_{K_i(t)}(z), \quad \|d_i\| = 1.$$

Consequently,

$$\|x - z\|^2 = \left\langle x - z, \sum_{i\in[m]} \lambda_i d_i \right\rangle = \sum_{i\in[m]} \lambda_i \langle x - z, d_i\rangle. \tag{12.6}$$

For any $i \in [m]$, since $d_i \in N_{K_i(t)}(z)$, we get

$$K_i(t) \subseteq H^-(z, d_i) := \{w \mid \langle w, d_i\rangle \leq \langle z, d_i\rangle\}.$$

Since $\|d_i\| = 1$ and d_i is the normal vector of the half-space $H^-(z, d_i)$, we have

$$\langle x - z, d_i\rangle = \operatorname{dist}(x, H^-(z, d)) \geq \operatorname{dist}(x, K_i(t)). \tag{12.7}$$

Now note that $\lambda_i \geq 0$ for all $i \in [m]$, we get

$$\begin{aligned}
\frac{\operatorname{dist}(x, K_0(t))}{\max_{i\in[m]} \operatorname{dist}(x, K_i(t))} &\leq \frac{\sum_{i\in[m]} \lambda_i \operatorname{dist}(x, K_0(t))}{\sum_{i\in[m]} \lambda_i \operatorname{dist}(x, K_i(t))} \\
&\overset{(12.7)}{\leq} \frac{\sum_{i\in[m]} \lambda_i \|x - z\|}{\sum_{i\in[m]} \lambda_i \langle x - z, d_i\rangle} \\
&\overset{(12.6)}{=} \frac{\|x - z\| \sum_{i\in[m]} \lambda_i}{\|x - z\|^2} \\
&= \frac{\sum_{i\in[m]} \lambda_i}{\|x - z\|}.
\end{aligned} \tag{12.8}$$

On the other hand, let $w^\star$ denote the center of K_0. Similar to (12.6), we can show

$$\langle -w^\star, x - z\rangle = \left\langle -w^\star, \sum_{i\in[m]} \lambda_i d_i \right\rangle = \sum_{i\in[m]} \lambda_i \langle -w^\star, d_i\rangle. \tag{12.9}$$

For any $i \in [m]$, since $d_i \in N_{K_i(t)}(z) \subseteq -K_i^\star$ and $\omega^\star$ is the center of K_0, we have

$$\langle -w^\star, d_i\rangle = \langle w^\star, -d_i\rangle \geq r_{K_i}(w^\star) \geq r_{K_0}(w^\star) = \mathrm{inrad}(K_0). \tag{12.10}$$

Note that $\|w^\star\| = 1$, we get

$$\begin{aligned} \|x - z\| &= \|w^\star\|\|x - z\| \\ &\geq \langle -w^\star, x - z\rangle \overset{(12.9)}{=} \sum_{i\in[m]} \lambda_i \langle -w^\star, d_i\rangle \\ &\overset{(12.10)}{\geq} \mathrm{inrad}(K_0) \sum_{i\in[m]} \lambda_i. \end{aligned} \tag{12.11}$$

Combining (12.8) and (12.11), we see that

$$\frac{\mathrm{dist}(x, K_0(t))}{\max_{i\in[m]} \mathrm{dist}(x, K_i(t))} \leq \frac{\sum_{i\in[m]} \lambda_i}{\|x - z\|} \leq \frac{1}{\mathrm{inrad}(K_0)}.$$

□

Corollary 12.1. *For any $x \in \mathcal{E}$ and $t \geq \omega_0(x)$, we have*

$$t - \omega_0(x) \geq \mathrm{inrad}(K_0)\mathrm{dist}(x, K_0(t)).$$

Proof. By Lemma 11.2, we have $t - \omega_i(x) \geq \mathrm{dist}(x, K_i(t)$ for all $i \in [m]$. Combined with Proposition 12.2, we get

$$\begin{aligned} t - \omega_0(x) = t - \min_{i\in[m]} \omega_i(x) &= \max_{i\in[m]} \{t - \omega_i(x)\} \\ &\geq \max_{i\in[m]} \mathrm{dist}(x, K_i(t)) \geq \mathrm{inrad}(K_0)\mathrm{dist}(x, K_0(t)). \end{aligned}$$

□

Since ω_0 is $1/r_0$-Lipschitz continuous, by Corollary 12.1, we get

$$\frac{\mathrm{dist}(x, K_0(t))}{r_0} \geq t - \omega_0(x) \geq \mathrm{inrad}(K_0)\mathrm{dist}(x, K_0(t)). \tag{12.12}$$

CHAPTER 13

THE APPROXIMATE MARGIN MAXIMIZATION ALGORITHM

By (12.1), ω_0 is a positively homogeneous function. We can find rays with large margin with respect to K_0 by solving

$$\begin{aligned} \max \quad & \omega_0(x) \\ s.t. \quad & \|x\| \leq 1. \end{aligned} \tag{13.1}$$

Let $\omega_0^\star$ and $z^\star$ denote the optimal value and optimal solution to (13.1). The existence of $z^\star$ follows from the continuity of ω_0 and compactness of the unit ball in $\mathcal{E}$. In the perceptron setting, when the centers of the half-spaces are used as reference points, we have $\omega_0^\star = \operatorname{inrad} K_0$ and $z^\star$ is the center of K_0.

Since ω_0 is concave, (13.1) is a convex optimization problem. For any $\epsilon \in (0, 1)$, we call x an ϵ-relative optimal solution to (13.1) when x satisfies

$$\|x\| \leq 1, \quad \omega_0(x) \geq (1 - \epsilon)\omega_0^\star.$$

We have the following results for $\omega_0^\star$ and $z^\star$:

Lemma 13.1. *The optimal value and optimal solution to* (13.1) *satisfy*

$$\omega_0^\star \geq \operatorname{inrad}(K_0), \quad \|z^\star\| = 1, \quad z^\star = \mathrm{P}_{K_0(\omega_0^\star)}(\vec{0}),$$

and the optimal solution $z^\star$ is unique.

Proof. Let $x = \vec{0}$ and $t = \operatorname{inrad}(K_0)$ in Corollary 12.1, we get

$$\operatorname{dist}(\vec{0}, K_0(\operatorname{inrad}(K_0))) \leq \frac{\operatorname{inrad}(K_0)}{\operatorname{inrad}(K_0)} = 1.$$

Hence there exists $x \in \overline{B(\vec{0}, 1)}$ such that $\omega_0(x) = \operatorname{inrad}(K_0)$, and thus,

$$\omega_0^\star \geq \operatorname{inrad}(K_0).$$

By the positive homogeneity of ω_0, we have

$$\omega_0\left(\frac{z^\star}{\|z^\star\|}\right) = \frac{\omega_0(z^\star)}{\|z^\star\|}.$$

Due to the optimality of $z^\star$ over $\overline{B(\vec{0},1)}$, we get $\|z^\star\| = 1$. To further characterize $z^\star$, let $z' := \mathrm{P}_{K_0(\omega_0^\star)}(\vec{0})$. Since $\omega_0(z^\star) = \omega_0^\star$, we have $\|z'\| \leq \|z^\star\| = 1$. If $z' \neq z^\star$, let $u = (z^\star + z')/2$. Then

$$\|u\| = \left\|\frac{z^\star + z'}{2}\right\| < \frac{\|z^\star\| + \|z'\|}{2} \leq 1$$

and $u \in \overline{B(\vec{0},1)}$. Also note that the concavity of ω_0 implies $\omega_0(u) \geq \omega_0^\star$. Hence

$$\omega_0\left(\frac{u}{\|u\|}\right) = \frac{\omega_0(u)}{\|u\|} > \omega_0^\star,$$

contradicting the optimality of $z^\star$ over the unit ball. Thus we get $z^\star = z' = \mathrm{P}_{K_0(\omega_0^\star)}(\vec{0})$. □

The following lemma is crucial to the analysis of our algorithms:

Proposition 13.1. *For all $x \in \overline{B(\vec{0},1)}$, we have*

$$\omega_0^\star - \omega_0(x) \geq \frac{\mathrm{inrad}(K_0)\|x - z^\star\|^2}{2}.$$

Proof. By Lemma 13.1, $z^\star$ is the closest point in $K_0(\omega^\star)$ to $\vec{0}$. Hence the half-space $\{z \mid \langle z - z^\star, z^\star\rangle \geq 0\}$ contains $K_0(\omega^\star)$. Hence by Lemma 13.1, for any $z \in K_0(\omega^\star)$, we have

$$1 = \langle z^\star, z^\star\rangle \leq \langle z, z^\star\rangle. \tag{13.2}$$

Now consider any $x \in \overline{B(\vec{0},1)}$. We have

$$\|x - z^\star\|^2 = (\|x\|^2 + \|z^\star\|^2) - 2\langle x, z^\star\rangle \leq 2(1 - \langle x, z^\star\rangle). \tag{13.3}$$

Now let $z = \mathrm{P}_{K_0(\omega_0^\star)}(x)$ in (13.2) and substitute for 1 in (13.3), we get

$$\begin{aligned}\|x - z^\star\|^2 &\leq 2\left(\left\langle \mathrm{P}_{K_0(\omega_0^\star)}(x), z^\star\right\rangle - \langle x, z^\star\rangle\right)\\ &= 2\left\langle \mathrm{P}_{K_0(\omega_0^\star)}(x) - x, z^\star\right\rangle\\ &\leq 2\left\|\mathrm{P}_{K_0(\omega_0^\star)}(x) - x\right\|\\ &= 2\mathrm{dist}\left(x, K_0(\omega_0^\star)\right).\end{aligned}$$

Consequently, by Corollary 12.1, for all $x \in \overline{B(\vec{0},1)}$, we have

$$\omega_0^\star - \omega_0(x) \geq \mathrm{inrad}(K_0)\mathrm{dist}\left(x, K_0(\omega_0^\star)\right) \geq \frac{\mathrm{inrad}(K_0)\|x - z^\star\|^2}{2}.$$

□

The next result extends the $\{z^\star\}$ in Proposition 13.1 to generic suplevel sets of ω_0 over the unit ball:

Corollary 13.1. *For any $x \in \overline{B(\vec{0},1)}$ and $t \in [\omega_0(x), \omega_0^\star]$, we have*

$$t - \omega_0(x) \geq \frac{\mathrm{inrad}(K_0)\mathrm{dist}\left(x, K_0(t) \cap \overline{B(\vec{0},1)}\right)^2}{2}.$$

Proof. For any $x \in \overline{B(\vec{0},1)}$ and $s \in [0,1]$, define

$$x(s) = z^\star + s(x - z^\star), \quad f(s) = \omega_0^\star - \omega_0(x(s)).$$

Then f is a convex function, and

$$f(0) = \omega_0^\star - \omega_0(z^\star) = 0, \quad f(1) = \omega_0^\star - \omega_0(x).$$

Since

$$\omega_0^\star - (t - \omega_0(x)) \in [\omega_0(x), \omega_0^\star],$$

due to the continuity of ω_0, there exists $s \in [0,1]$ such that

$$\omega_0(x(s)) = \omega_0^\star - (t - \omega_0(x)), \quad f(s) = \omega_0^\star - \omega_0(x(s)) = t - \omega_0(x).$$

Using Proposition 13.1, we get

$$s\|x - z^\star\| = \|x(s) - z^\star\| \leq \sqrt{\frac{2(\omega_0^\star - \omega_0(x(s)))}{\text{inrad}(K_0)}} = \sqrt{\frac{2(t - \omega_0(x))}{\text{inrad}(K_0)}}.$$

By the convexity of f, we have

$$\frac{f(s)}{s} = \frac{f(s) - f(0)}{s} \leq \frac{f(1) - f(0)}{1} \leq \frac{f(1) - f(1-s)}{s}.$$

Thus

$$f(1-s) \leq f(1) - f(s) = (\omega_0^\star - \omega_0(x)) - (t - \omega_0(x)) = \omega_0^\star - t,$$

and

$$\omega_0(x(1-s)) = \omega_0^\star - f(1-s) \geq t.$$

By the convexity of the unit ball, we conclude that

$$x(1-s) \in K_0(t) \cap \overline{B(\vec{0}, 1)},$$

and

$$\text{dist}\left(x, K_0(t) \cap \overline{B(\vec{0}, 1)}\right) \leq \|x - x(1-s)\| = s\|x - z^\star\| \leq \sqrt{\frac{2(t - \omega_0(x))}{\text{inrad}(K_0)}}.$$

□

By Corollary 12.1 and Proposition 13.1, we see that when restricted to the unit ball, ω_0 is a Lipschitz continuous function with quadratic decrease rate with respect to its maximizer $z^\star$. In [24], the authors proposed an scheme which produces algorithms with near-optimal (with respect to the growth rate of the convex objective) convergence rates when a good estimate of the optimal value of the Lipschitz-continuous convex objective function is available. We next present the Approximate Margin Maximization Algorithm (AMMA), a two-stage algorithm, for solving (13.1). The first stage of AMMA, a modified extension of the Approximate Large Margin Algorithm in [9] designed for maximal margin perceptrons, produces an estimate of $\omega_0^\star$, which is then taken as an input to the second stage, an application of the parallel scheme introduced in [24].

13.1 First stage

We now present the first stage algorithm as Algorithm 7, a modified version of the Approximate Large Margin Algorithm [9] generalized from the perceptron setting to the intersection of generic closed convex cones:

Algorithm 7: Approximate Margin Maximization Algorithm-First Stage

input : target tolerance $\epsilon \in \left(\frac{1}{2}, 1\right)$
output: a point $\bar{x} \in \mathcal{E}$ satisfying $\|\bar{x}\| = 1$ and $\omega_0(\bar{x}) \geq (1-\epsilon)\omega_0^\star$
initialization: let $k = 0$, $\alpha = \frac{1-\epsilon}{2\epsilon-1}$, $\beta = \frac{1}{2\epsilon-1}$, $R_0 = 0$ and $x_0 = \vec{0} \in \mathcal{E}$

repeat
  compute $\omega_i(x_k)$ for all $i \in [m]$ and let $i_k \in \operatorname{argmin}_{i\in[m]} \omega_i(x_k)$;
  let $\omega_0(x_k) = \omega_{i_k}(x_k)$;
  if $k > 0$ *and* $\omega_0(x_k) \geq \frac{\alpha}{R_k}$ **then**
    return $\bar{x} := x_k$
  else
    compute $g_k \in \partial\omega_{i_k}(x_k)$;
    if $k = 0$ **then**
      $x_{k+1} := \frac{1}{\sqrt{\beta}\|g_k\|} g_k$;
    else
      $\hat{x}_{k+1} := x_k + \frac{1}{R_k\|g_k\|^2} g_k$;
      $x_{k+1} := \mathrm{P}_{\overline{B\left(\vec{0},1\right)}}(\hat{x}_{k+1})$;
    $R_{k+1} := \sqrt{R_k^2 + \frac{\beta}{\|g_k\|^2}}$;
    $k = k + 1$;

Remark. In the analysis of Algorithm 7 (Proposition 13.2), we need

$$\alpha = (1-\epsilon)\beta, \beta = 2\alpha + 1. \tag{13.4}$$

Solving this system gives $\alpha = \frac{1-\epsilon}{2\epsilon-1}$ and $\beta = \frac{1}{2\epsilon-1}$. Since the algorithm terminates when $\omega_0(x_k) \geq \frac{\alpha}{R_k}$, to have a reasonable stopping criterion, we have to have $\alpha > 0$, which limits our choice of ϵ to $(\frac{1}{2}, 1)$. It should be noted that when $\alpha > 0$, we have $\beta > 1$ and $\|x_1\| = 1/\sqrt{\beta} < 1$, and x_1 is also in the unit ball.

Remark. In the perceptron case, we always have $\|g_k\| = 1$. One recovers the

Approximate Large Margin Algorithm (ALMA) in [9] by setting

$$\alpha' = \frac{2-2\epsilon}{\epsilon},\ \beta' = \frac{1}{2}.$$

in Algorithm 7. By choosing different α and β in Algorithm 7, we are able to simplify the convergence analysis of the algorithm. One significant drawback of Algorithm 7, however, is a much more restricted range of the relative accuracy ϵ ($(\frac{1}{2}, 1)$) v.s. $(0,\ 1)$). We will extend the range of target accuracy to $(0,\ 1)$ in Algorithm 8.

For notational brevity, let $\bar{z} = z^\star/\omega_0^\star$. Then by Lemma 13.1 and the positive homogeneity of ω, we have

$$\omega_0(\bar{z}) = 1,\ \|\bar{z}\| = \frac{1}{\omega_0^\star},\ \bar{z} = \mathrm{P}_{K_0(1)}(\vec{0}). \tag{13.5}$$

Moreover, note that for any g_k in Algorithm 7, by Proposition 11.5, we have

$$g_k \in \left\{ d \mid d \in K_{i_k}^\star,\ \langle e_{i_k}, d\rangle = 1 \right\}.$$

Let $x = \bar{z}$ in (12.1), we get

$$\langle g_k, \bar{z}\rangle \geq \min\left\{ \langle \bar{z}, d\rangle \mid d \in K_i^\star,\ \langle e_i, d\rangle = 1,\ i \in [m] \right\} = \omega_0(\bar{z}) = 1. \tag{13.6}$$

Proposition 13.2. *For any $\epsilon \in \left(\frac{1}{2}, 1\right)$, if Algorithm 7 does not terminate at the kth iteration, then*

$$\langle x_{k+1}, \bar{z}\rangle \geq \frac{R_{k+1}}{\beta}. \tag{13.7}$$

Consequently, by the construction of Algorithm 7 and (13.5)*, we have*

$$R_{k+1} \leq \beta \|x_{k+1}\| \|\bar{z}\| \leq \frac{\beta}{\omega_0^\star} = \frac{\alpha}{(1-\epsilon)\omega_0^\star}. \tag{13.8}$$

Proof. For any $k \in \mathcal{N}$, let $s_k = \|g_k\|^{-1}$. Since $R_1 = \sqrt{\beta} s_0$, by (13.6), we have

$$\langle x_1, \bar{z}\rangle = \frac{s_0}{\sqrt{\beta}} \langle g_0, \bar{z}\rangle \geq \frac{s_0}{\sqrt{\beta}} = \frac{R_1}{\beta} = \frac{R_{k+1}}{\beta},$$

and (13.7) holds for $k = 0$. We now proceed by induction.

Now suppose (13.7) holds for $k = k_0 - 1 \in \mathcal{N}$ and Algorithm 7 does not terminate at the k_0th iteration. Then

$$\langle x_{k_0}, g_{k_0} \rangle = \omega_{i_{k_0}}(x_{k_0}) = \omega_0(x_{k_0}) < \frac{\alpha}{R_{k_0}}.$$

Consequently, we have

$$\|\hat{x}_{k_0+1}\|^2 = \left\| x_{k_0} + \frac{s_{k_0}^2}{R_{k_0}} g_{k_0} \right\|^2 = \|x_{k_0}\|^2 + \frac{s_{k_0}^2(1 + 2R_{k_0}\langle x_{k_0}, g_{k_0}\rangle)}{R_{k_0}^2} < 1 + \frac{s_{k_0}^2(1+2\alpha)}{R_{k_0}^2}.$$

Since $\beta = 1 + 2\alpha$ (see (13.4)), by the definition of R_{k_0+1}, we get

$$\|\hat{x}_{k_0+1}\|^2 < 1 + \frac{s_{k_0}^2(1+2\alpha)}{R_{k_0}^2} = \frac{R_{k_0}^2 + \beta s_{k_0}^2}{R_{k_0}^2} = \frac{R_{k_0+1}^2}{R_{k_0}^2}. \tag{13.9}$$

On the other hand, by (13.6), we have

$$\langle \hat{x}_{k_0+1}, \bar{z} \rangle = \langle x_{k_0}, \bar{z} \rangle + \frac{s_{k_0}^2}{R_{k_0}} \langle g_{k_0}, \bar{z} \rangle \geq \langle x_{k_0}, \bar{z} \rangle + \frac{s_{k_0}^2}{R_{k_0}}. \tag{13.10}$$

Note that (13.9) implies $\|\hat{x}_{k_0+1}\| < R_{k_0+1}/R_{k_0}$, by the definition of x_{k_0+1}, we have

$$\langle x_{k_0+1}, \bar{z} \rangle = \frac{\langle \hat{x}_{k_0+1}, \bar{z} \rangle}{\min\{\|\hat{x}_{k_0+1}\|, 1\}} \geq \frac{\langle \hat{x}_{k_0+1}, \bar{z} \rangle}{\|\hat{x}_{k_0+1}\|} \geq \frac{R_{k_0}\langle \hat{x}_{k_0+1}, \bar{z} \rangle}{R_{k_0+1}}.$$

Combined with (13.10) and our induction assumption $\langle x_{k_0}, \bar{z} \rangle \geq R_{k_0}/\beta$, we get

$$\langle x_{k_0+1}, \bar{z} \rangle > \frac{R_{k_0}\langle x_{k_0}, \bar{z} \rangle + s_{k_0}^2}{R_{k_0+1}} \geq \frac{R_{k_0}^2 + \beta s_{k_0}^2}{\beta R_{k_0+1}} = \frac{R_{k_0+1}^2}{\beta R_{k_0+1}} = \frac{R_{k_0+1}}{\beta}. \tag{13.11}$$

We conclude that (13.7) holds for all $k \in \mathcal{N}$.

□

Theorem 13.1. *For any $\epsilon \in \left(\frac{1}{2}, 1\right)$, Algorithm 7 terminates in at most*

$$\left(\frac{1}{2\epsilon - 1}\right)\left(\frac{1}{r_0 \omega_0^\star}\right)^2$$

iterations, and its output $\bar{x}$ satisfies

$$\omega_0(\bar{x}) \geq (1 - \epsilon)\omega_0^\star.$$

Proof. By Lemma 11.1, for any $k \in \mathcal{N}$, the margin function ω_{i_k} is $(1/r_{i_k})$-Lipschitz continuous. Consequently, assuming the algorithm has not terminated before iteration $k+1$, we have

$$R_{k+1}^2 = \sum_{j=0}^{k} \frac{\beta}{\|g_j\|^2} \geq \beta \sum_{j=0}^{k} r_{i_k}^2 \geq (k+1)\beta r_0^2.$$

Hence by (13.8), we get

$$k+1 \leq \frac{R_{k+1}^2}{\beta r_0^2} = \frac{\beta}{r_0^2}\left(\frac{R_{k+1}}{\beta}\right)^2 \leq \left(\frac{1}{2\epsilon - 1}\right)\left(\frac{1}{r_0 \omega_0^\star}\right)^2.$$

By our construction of Algorithm 7 and Proposition 13.2, $\bar{x}$, the output of Algorithm 7, satisfies

$$\omega_0(\bar{x}) \geq \frac{\alpha}{R_k} \geq (1-\epsilon)\omega_0^\star.$$

□

13.2 Second stage

In Algorithm 7, the range of relative accuracy $(\frac{1}{2}, 1)$ is very restrictive. However, unlike many first-order methods, Algorithm 7 terminates with an output $\bar{x}$ satisfying the required relative accuracy $\epsilon_0 \in (\frac{1}{2}, 1)$, i.e.,

$$\omega_0^\star \geq \omega_0(\bar{x}) \geq (1-\epsilon)\omega_0^\star.$$

This provides a good estimate of $\omega_0^\star$. Based on the output from Algorithm 7, we next present Algorithm 8, the second stage algorithm that extends the range of tolerance to $(0, 1)$.

Algorithm 8 can be thought of as an application of the parallel scheme proposed by [24]. Based on its input ϵ_0 and x_0, Algorithm 8 first generates a series of step

Algorithm 8: Approximate Margin Maximization Algorithm-Second Stage

input : initial tolerance $\epsilon_0 \in (0,1)$, a point x_0 satisfying $\omega_0(x_0) \geq (1-\epsilon_0)\omega_0^\star$ and the target tolerance $\epsilon \in (0, \epsilon_0)$

initialization: let $k = 0$ and $J = \left\lceil \log_{\frac{3}{2}} \left(\frac{\epsilon_0}{\epsilon(1-\epsilon_0)} \right) \right\rceil - 1$, compute $g_0 \in \partial\omega_0(x_0)$ and for all $j \in [J]$, let $x_0^j = x_0$, $g_0^j = g_0$, $\Delta_j = \frac{\epsilon_0 \cdot \omega_0(x_0)}{3(3/2)^j(1-\epsilon_0)}$ and $c_0^j = \omega_0(x_0) + \Delta_j$;

repeat

  for $0 \leq j \leq J$ **do**

    $\hat{x}_{k+1}^j = x_k^j + \frac{\Delta_j}{\|g_k^j\|^2} g_k^j$;

    $x_{k+1}^j = \mathrm{P}_{\overline{B(\vec{0},1)}}\left(\hat{x}_{k+1}^j\right)$;

    compute $\omega_i\left(x_{k+1}^j\right)$ for all $i \in [m]$ and let $i_{k+1}^j \in \mathrm{argmin}_{i\in[m]}\, \omega_i\left(x_{k+1}^j\right)$;

    let $\omega_0\left(x_{k+1}^j\right) = \omega_{i_{k+1}^j}\left(x_{k+1}^j\right)$ and compute $g_{k+1}^j \in \partial\omega_{i_{k+1}^j}\left(x_{k+1}^j\right)$;

  let $\mathbf{j}(k+1) = \mathrm{argmax}_{j\in[J]}\, \omega_0\left(x_{k+1}^j\right)$;

  for $0 \leq j \leq J$ **do**

    if $\omega_0\left(x^{\mathbf{j}}(k+1)_{k+1}\right) > c_k^j$ **then**

      let $x_{k+1}^j = x_{k+1}^{\mathbf{j}(k+1)}$, $g_{k+1}^j = g_{k+1}^{\mathbf{j}(k+1)}$ and $c_{k+1}^j = \omega_0\left(x_{k+1}^{\mathbf{j}(k+1)}\right) + \Delta_j$;

      `// restart copy` j

    else

      let $c_{k+1}^j = c_k^j$;

  $k = k + 1$;

sizes $\{\Delta_j\}_{j\in[J]}$ and targets $\{c_j\}_{j\in[J]}$. Then the algorithm executes the projected subgradient method on the unit ball with different step sizes and targets in parallel. In every iteration, we find the copies whose targets are reached, and restart these copies with new targets. It should be noted that Algorithm 8 never terminates by itself.

We next conduct the convergence analysis of Algorithm 8. The result can be seen as a corollary to Theorem 2 in [24] up to a constant factor. In the generic parallel scheme proposed in [24], a given copy of the parallel scheme can be relevant for multiple times. As for Algorithm 8, since the input $\omega_0(x_0)$ of Algorithm 8

provides an upper bound on $\omega_0^\star$, we do not need to rely on assumptions on the suboptimality of the initial iterate as in Theorem 2 in [24]. Utilizing this upper bound, we will identify the "relevant" copies in Algorithm 8, where the notion of relevance will be defined in (13.14).

As mentioned previously, in Algorithm 8, our choice of step sizes are based on estimates of $\omega_0^\star$ and the suboptimality of x_0:

Lemma 13.2. *In Algorithm 8, we have*

$$\omega_0^\star - \omega_0(x_0) \leq 3\Delta_0 \tag{13.12}$$

and

$$\epsilon\omega_0^\star \in [2\Delta_J, 3\Delta_0). \tag{13.13}$$

Proof. Since $\omega_0(x_0) \geq (1-\epsilon_0)\omega_0^\star$, we have

$$\omega_0^\star - \omega_0(x_0) \leq \epsilon_0\omega_0^\star \leq \frac{\epsilon_0\omega_0(x_0)}{1-\epsilon_0} = 3\Delta_0.$$

We also get

$$\epsilon\omega_0^\star < \epsilon_0\omega_0^\star \leq 3\Delta_0.$$

On the other hand, since $\omega_0(x) \leq \omega_0^\star$, our definition of J gives

$$2\Delta_J = \frac{2\Delta_0}{(3/2)^J} \leq \frac{\epsilon_0\omega_0(x_0)}{(3/2)^{J+1}(1-\epsilon_0)} \leq \epsilon\omega_0(x_0) \leq \epsilon\omega_0^\star.$$

□

By the construction of Algorithm 8, we see that the targets $\{c_k^j\}_{k\in\mathcal{N}}$ are increasing for all $j \in [J]$. The next result indicates how restarts and the objective values are connected:

Lemma 13.3. *For any $j \in [J]$ and $k \in \mathcal{N}$, if copy j is restarted in the kth iteration of Algorithm 8, then*

$$\omega_0\left(x_{k+1}^j\right) > \max_{j' \in [J], k' < k} \omega_0\left(x_{k'}^{j'}\right).$$

Proof. If copy j is restarted in iteration k, then $\omega_0\left(x_{k+1}^j\right) > c_k^j$. Suppose for contradiction that there exists $j' \leq J$ and $k' \leq k$ such that $\omega_0\left(x_{k'}^{j'}\right) \geq \omega_0\left(x_k^j\right)$. Then by the definition of $\mathbf{j}(k')$, we have

$$\omega_0\left(x_{k'}^{\mathbf{j}(k')}\right) \geq \omega_0\left(x_{k'}^{j'}\right) \geq \omega_0\left(x_{k+1}^j\right) > c_k^j \geq c_{k'}^j.$$

However, by the construction of Algorithm 8, we immediately have

$$\omega_0\left(x_{k'}^{\mathbf{j}(k')}\right) = \max_{j \in [J]} \omega_0\left(x_{k'}^j\right) \leq c_{k'}^j.$$

Thus we arrive at a contradiction and conclude that the statement is true. □

For any iteration k of Algorithm 8, we let $\tilde{j}(k)$ denote the "relevant" copy, which has the largest index j satisfying

$$c_k^j \in [\omega_0^\star - 3\Delta_j, \omega_0^\star - \Delta_j). \tag{13.14}$$

The next result shows that the step size of the relevant copy is closely related to the largest objective value encountered in Algorithm 8 so far:

Lemma 13.4. *For any $k \in \mathcal{N}$, if there exists a relevant copy $\tilde{j}(k)$, then we have*

$$\max_{j \in [J], k' \in [k]} \omega_0\left(x_{k'}^j\right) \in \left[\omega_0^\star - 4\Delta_{\tilde{j}(k)}, \omega_0^\star - \Delta_{\tilde{j}(k)}\right).$$

Proof. Note that by the construction of Algorithm 8, we have $\omega_0\left(x_{k'}^j\right) < c_{k'}^{\tilde{j}(k)}$ for all $k' \leq k$ and $j \in [J]$. Hence by the definition of relevance in (13.14), we get

$$\max_{j \in [J], k' \in [k]} \omega_0\left(x_{k'}^j\right) < \max_{k' \in [k]} c_{k'}^{\tilde{j}(k)} \leq c_k^{\tilde{j}(k)} < \omega_0^\star - \Delta_{\tilde{j}(k)}.$$

On the other hand, by the construction of Algorithm 8, there exists $\dot{k} \leq k$ such that

$$c_k^{\tilde{j}(k)} = \omega_0\left(x_{\dot{k}}^{\tilde{j}(k)}\right) + \Delta_{\tilde{j}(k)}.$$

Hence

$$\max_{j\in[J],k'\in[k]} \omega_0\left(x_{k'}^{j}\right) \geq \omega_0\left(x_{\dot{k}}^{\tilde{j}(k)}\right) \geq c_k^{\tilde{j}(k)} - \Delta_{\tilde{j}(k)} \geq \omega_0^\star - 4\Delta_{\tilde{j}(k)}.$$

□

Relevant copies are crucial to our analysis of Algorithm 8, and the next result characterizes some important properties of the relevant copies:

Proposition 13.3. *In Algorithm 8, for any $k \in \mathcal{N}$, if*

$$\max_{j\in[J],k'\in[k]} \omega_0\left(x_{k'}^{j}\right) < (1-\epsilon)\omega_0^\star, \tag{13.15}$$

then there exists a relevant copy $\tilde{j}(k)$. Moreover, when (13.15) *holds, the sequence of relevant copies $\left\{\tilde{j}(k')\right\}_{k'\in[k]}$ is increasing.*

Proof. We first consider the base case where $k = 0$. For any $j \in [J]$, since $c_0^j = \omega_0(x_0) + \Delta_j$, we have c_0^j satisfies (13.14) if and only if

$$\omega_0^\star - \omega_0(x_0) = \omega_0^\star - \left(c_0^j - \Delta_j\right) \in (2\Delta_j, 4\Delta_j]. \tag{13.16}$$

If there is no relevant copy $\tilde{j}(0)$, then $\omega_0^\star - \omega_0(x_0) \notin \cup_{j\in[J]}(2\Delta_j, 4\Delta_j]$. Thus

$$\omega_0^\star - \omega_0(x_0) \notin \cup_{j\in[J]}(2\Delta_j, 3\Delta_j] = (2\Delta_J, 3\Delta_0].$$

Here the last equation follows from our construction of $\{\Delta_j\}_{j\in[J]}$. However, (13.12) implies $\omega_0^\star - \omega_0(x_0) \leq 3\Delta_0$. Thus by (13.13), we have

$$\omega_0^\star - \omega_0(x_0) \leq 2\Delta_J \leq \epsilon\omega_0^\star,$$

and (13.15) does not hold. Consequently, when (13.15) holds, there exists a relevant copy $\tilde{j}(0)$ such that

$$\omega_0^\star - \omega_0\left(x_{k_0+1}^{\mathbf{j}(k_0+1)}\right) \in (2\Delta_{\tilde{j}(0)}, 3\Delta_{\tilde{j}(0)}], \tag{13.17}$$

and the statements hold when $k = 0$. We now proceed by induction.

Suppose the statements hold for $k = k_0$. We will show that they remain true for $k = k_0 + 1$:

Case 1. If there is no relevant copy in iteration k_0, then we have encountered an iterate satisfying the desired accuracy, and (13.15) does not hold for $k = k_0 + 1$.

Case 2. If the relevant copy $\tilde{j}(k_0)$ is not restarted in iteration k_0, then we have

$$c_{k_0+1}^{\tilde{j}(k_0)} = c_{k_0}^{\tilde{j}(k_0)} \in [\omega_0^\star - 3\Delta_{\tilde{j}(k_0)}, \omega_0^\star - \Delta_{\tilde{j}(k_0)}).$$

Hence $c_{k_0+1}^{\tilde{j}(k_0)}$ satisfies (13.14), and $\tilde{j}(k_0+1) \geq \tilde{j}(k_0)$.

Case 3. If the relevant copy $\tilde{j}(k_0)$ is restarted in iteration k_0, then by the construction of Algorithm 8,

$$\omega_0\left(x_{k_0+1}^{\mathbf{j}(k_0+1)}\right) = \omega_0\left(x_{k_0+1}^{\tilde{j}(k_0)}\right) > c_{k_0}^{\tilde{j}(k_0)} \geq \omega_0^\star - 3\Delta_{\tilde{j}(k_0)}. \tag{13.18}$$

(a) If

$$\omega_0\left(x_{k_0+1}^{\mathbf{j}(k_0+1)}\right) \in [\omega_0^\star - 3\Delta_{\tilde{j}(k_0)}, \omega_0^\star - 2\Delta_{\tilde{j}(k_0)}),$$

then

$$c_{k_0+1}^{\tilde{j}(k_0)} = \omega_0\left(x_{k_0+1}^{\mathbf{j}(k_0+1)}\right) + \Delta_{\tilde{j}(k_0)} \in [\omega_0^\star - 2\Delta_{\tilde{j}(k_0)}, \omega_0^\star - \Delta_{\tilde{j}(k_0)}).$$

Thus $c_{k_0+1}^{\tilde{j}(k_0)}$ also satisfies (13.14). There exists a relevant copy $\tilde{j}(k_0+1)$, and $\tilde{j}(k_0+1) \geq \tilde{j}(k_0)$.

(b) Consider the case where

$$\omega_0\left(x_{k_0+1}^{\mathbf{j}(k_0+1)}\right) \geq \omega_0^\star - 2\Delta_{\tilde{j}(k_0)}. \tag{13.19}$$

i. If $\tilde{j}(k_0) = J$, then by (13.19) and Lemma 13.2, we have

$$\omega_0\left(x_{k_0+1}^{\mathbf{j}(k_0+1)}\right) \geq \omega_0^\star - 2\Delta_{\tilde{j}(k_0)} = \omega_0^\star - 2\Delta_J \geq (1-\epsilon)\omega_0^\star,$$

and (13.15) does not hold when $k = k_0 + 1$.

ii. If $\tilde{j}(k_0) < J$, then by (13.19), we get

$$\omega_0^\star - \omega_0\left(x_{k_0+1}^{\mathbf{j}(k_0+1)}\right) \leq 2\Delta_{\tilde{j}(k_0)} \leq 3\Delta_{\tilde{k_0}+1}. \tag{13.20}$$

Replacing x_0 with $x_{k_0+1}^{\mathbf{j}(k_0+1)}$, similar to the discussion for the case where $k = 0$, by (13.20), (13.13) and the construction of Algorithm 8, we can show that either $x_{k_0+1}^{\mathbf{j}(k_0+1)}$ satisfies the desired accuracy and (13.15) does not hold for $k = k_0+1$, or there exists $\tilde{j}(k_0)+1 \leq j \leq J$ such that

$$\omega_0^\star - \omega_0\left(x_{k_0+1}^{\mathbf{j}(k_0+1)}\right) \in (2\Delta_j, 3\Delta_j]. \tag{13.21}$$

A. If copy j is restarted in iteration k_0, then

$$c_{k_0+1}^j = \omega_0\left(x_{k_0+1}^{\mathbf{j}(k_0+1)}\right) + \Delta_j \in [\omega_0^\star - 2\Delta_j, \omega_0^\star - \Delta_j).$$

B. Otherwise, we have

$$c_{k_0+1}^j = c_{k_0}^j \geq \omega_0\left(x_{k_0+1}^{\mathbf{j}(k_0+1)}\right) \geq \omega_0^\star - 3\Delta_{\tilde{j}(k_0)}.$$

Also note that by the construction of Algorithm 8, there exists $k' < k_0$ such that

$$c_{k_0}^j = \omega_0\left(x_{k'}^j\right) + \Delta_j.$$

Due to Lemma 13.3, we get

$$c_{k_0+1}^j = \omega_0\left(x_{k'}^j\right) + \Delta_j < \omega_0\left(x_{k_0+1}^{\mathbf{j}(k_0+1)}\right) + \Delta_j < \omega_0^\star - \Delta_j.$$

In both cases, we can conclude that copy j satisfies (13.14). Hence $\tilde{j}(k_0+1)$ exists, and $\tilde{j}(k_0+1) \geq j > \tilde{j}(k_0)$.

Hence the statements hold for $k = k_0 + 1$, and we conclude that they are true for all $k \in \mathcal{N}$. □

Now we are ready to analyze the convergence rate of Algorithm 8:

Proposition 13.4. *In Algorithm 8, there exists $j \in [J]$ and*

$$k \leq \frac{36}{\epsilon}\left(\frac{1}{r_0^2 \operatorname{inrad}(K_0)\omega_0(x_0)}\right)$$

such that

$$\omega_0\left(x_k^j\right) \geq (1-\epsilon)\omega_0^\star.$$

Remark. By Lemma 13.1, we have $\operatorname{inrad}(K_0) \leq \omega_0^\star$ and

$$\begin{aligned}\frac{36}{\epsilon}\left(\frac{1}{r_0^2 \operatorname{inrad}(K_0)\omega_0(x_0)}\right) &\leq \frac{36}{\epsilon}\left(\frac{1}{(1-\epsilon_0) r_0^2 \operatorname{inrad}(K_0)\omega_0^\star}\right) \\ &\leq \frac{36}{\epsilon}\left(\frac{1}{1-\epsilon_0}\right)\left(\frac{1}{r_0 \operatorname{inrad}(K_0)}\right)^2.\end{aligned} \tag{13.22}$$

Proof. By Proposition 13.3, if in an iteration of Algorithm 8, none of the copies are relevant, then we must have encountered an iterate satisfying the desired target. Hence we proceed by showing an upper bound on the number of iterations during which relevant copies exist.

We next bound the number of iterations it takes for a relevant copy to restart. Assume copy j is relevant in iteration k. Then there exists $k' \leq k$ such that

$$c_k^j = \omega_0(x_{k'}^j) + \Delta_j.$$

By the quadratic decrease rate of ω_0 over the unit ball with respect to its sublevel sets (Corollary 13.1), we get

$$\operatorname{dist}\left(x, K_0\left(c_k^j\right) \cap \overline{B(\vec{0},1)}\right)^2 \leq \frac{2\left(c_k^j - \omega_0\left(x_{k'}^j\right)\right)}{\operatorname{inrad}(K_0)} \leq \frac{2\Delta_j}{\operatorname{inrad}(K_0)}. \tag{13.23}$$

If copy j is not restarted in iteration k, then we have $\omega_0(x_k^j) \leq c_k^j$. By the definition of relevance, we get

$$\omega_0^\star - \omega_0\left(x_k^j\right) \geq \omega_0^\star - c_k^j > \omega_0^\star - (\omega_0^\star - \Delta_j) = \Delta_j. \tag{13.24}$$

Since $g_k^j \in \partial\omega_0\left(x_k^j\right)$, we have

$$\langle x_k^j - z^\star, g_k^j \rangle \leq \omega_0\left(x_k^j\right) - \omega_0^\star < \Delta_j, \tag{13.25}$$

and

$$\begin{aligned}
\left\|x_{k+1}^j - z^\star\right\|^2 &\leq \left\|\hat{x}_{k+1}^j - z^\star\right\|^2 \\
&= \left\|x_k^j + \frac{\Delta_j}{\|g_k^j\|^2} g_k^j - z^\star\right\|^2 \\
&= \left\|x_k^j - z^\star\right\|^2 + \frac{\Delta_j}{\|g_k^j\|^2}\left(\frac{\Delta_j}{\|g_k^j\|^2}\|g_k^j\|^2 + 2\langle x_k^j - z^\star, g\rangle\right) \\
&\overset{(13.25)}{\leq} \left\|x_k^j - z^\star\right\|^2 + \frac{\Delta_j}{\|g_k^j\|^2}\left(\frac{\Delta_j}{\|g_k^j\|^2}\|g_k^j\|^2 + 2\left(\omega_0\left(x_k^j\right) - \omega_0^\star\right)\right) \\
&\overset{(13.24)}{<} \left\|x_k^j - z^\star\right\|^2 - \left(\frac{\Delta_j}{\|g_k^j\|}\right)^2 \\
&\leq \left\|x_k^j - z^\star\right\|^2 - (r_0\Delta_j)^2 .
\end{aligned} \tag{13.26}$$

Here the first inequality follows from the definition of x_{k+1}^j, and the last inequality is due to (11.4) and the definition of r_0. Consequently, by (13.23) and (13.26), after iteration k, copy j must be restarted within

$$\begin{aligned}
\left(\frac{2\Delta_j}{\operatorname{inrad}(K_0)}\right)\left(\frac{1}{r_0^2\Delta_j^2}\right) &= \left(\frac{2}{r_0^2\operatorname{inrad}(K_0)}\right)\left(\frac{1}{\Delta_j}\right) \\
&= \left(\frac{3}{2}\right)^j \left(\frac{6}{r_0^2\operatorname{inrad}(K_0)}\right)\left(\frac{1-\epsilon_0}{\epsilon_0\omega_0(x_0)}\right)
\end{aligned} \tag{13.27}$$

iterations.

For any $j \in [J]$, assume that $\tilde{j}(k') = j$, then

$$c^j_{k'} \geq \omega^\star_0 - 3\Delta_j.$$

After each restart, the target c^j_k is increased by at least Δ_j. Hence copy j can be relevant and go through up to two restarts before $c^j_{k'} \geq \omega^\star_0 - \Delta_j$ and it becomes irrelevant forever. Using (13.27), we see that after

$$\begin{aligned}
& 2\sum_{j\in[J]} \left(\frac{3}{2}\right)^j \left(\frac{6}{r_0^2 \operatorname{inrad}(K_0)}\right)\left(\frac{1-\epsilon_0}{\epsilon_0\omega_0(x_0)}\right) \\
= & \left(\left(\frac{3}{2}\right)^{J+1} - 1\right)\left(\frac{24}{r_0^2 \operatorname{inrad}(K_0)}\right)\left(\frac{1-\epsilon_0}{\epsilon_0\omega_0(x_0)}\right) \\
< & \left(\frac{3\epsilon_0}{2\epsilon(1-\epsilon_0)} - 1\right)\left(\frac{24}{r_0^2 \operatorname{inrad}(K_0)}\right)\left(\frac{1-\epsilon_0}{\epsilon_0\omega_0(x_0)}\right) \\
= & \frac{36}{\epsilon}\left(\frac{1}{r_0^2 \operatorname{inrad}(K_0)\omega_0(x_0)}\right)
\end{aligned}$$

iterations, the relevant copy can no longer be restarted. Since we have seen that relevant copies must be restarted within finitely many iterations, this implies that there is no relevant copy. By Proposition 13.3, we conclude that the algorithm must have generated an iterate satisfying the target relative accuracy of ϵ. □

Theorem 13.2. *Since Algorithm 8 makes at most* $\left\lceil \log_{\frac{3}{2}}\left(\frac{\epsilon_0}{\epsilon(1-\epsilon_0)}\right)\right\rceil + 1$ *supgradient oracle calls per iteration, it produces an* ϵ*-relatively optimal solution after making at most*

$$\frac{36}{\epsilon}\left(\log_{\frac{3}{2}}\left(\frac{\epsilon_0}{\epsilon(1-\epsilon_0)}\right) + 1\right)\left(\frac{1}{r_0^2 \operatorname{inrad}(K_0)\omega_0(x_0)}\right)$$

supgradient oracle calls.

13.3 Approximate Margin Maximization Algorithm

Combining Algorithm 7 and 8, we get the Approximate Maximization Algorithm:

Algorithm 9: Approximate Margin Maximization Algorithm

input : target tolerance $\epsilon \in (0,1)$
output: a point $\bar{x} \in \mathcal{E}$ satisfying $\|\bar{x}\| = 1$ and $\omega_0(\bar{x}) \geq (1-\epsilon)\omega_0^\star$

run Algorithm 7 with $\epsilon = \frac{2}{3}$ to obtain x_0 ;
run Algorithm 8 with x_0, $\epsilon_0 = \frac{2}{3}$ and ϵ to obtain $\bar{x}$;

Remark. It should be noted that here our choice of ϵ_0 is somewhat arbitrary. According to Theorem 13.1, any choice of $\epsilon \in (\frac{1}{2}, 1)$ that is not too close to $\frac{1}{2}$ should suffice.

Combining Theorem 13.1 and (13.22), we obtain the follow convergence rate for Algorithm 9:

Theorem 13.3. *For any $\epsilon \in (0,1)$, Algorithm 9 terminates after making at most*

$$\frac{36}{\epsilon}\left(\log_{\frac{3}{2}}\left(\frac{1}{\epsilon}\right)+3\right)\left(\frac{1}{r_0 \operatorname{inrad}(K_0)}\right)^2$$

first-order oracles of ω_0, and its output $\bar{x}$ satisfies

$$\omega_0(\bar{x}) \geq (1-\epsilon)\omega_0^\star.$$

CHAPTER 14

NUMERICAL EXPERIMENTS

In this chapter, we present numerical experiments to demonstrate the results in Theorem 13.3.

14.1 Intersection of half-Spaces

We consider a perceptron problem with m centers in $\mathbb{R}^n$ such that

$$e_i = \left(e_i^1, \ldots, e_i^{n-1}, \rho\right), \quad \|e_i\| = 1. \tag{14.1}$$

and

$$\vec{0} \in \operatorname{conv}\left(\{(e_1^1, \ldots, e_1^{n-1}), \cdots, (e_m^1, \ldots, e_m^{n-1})\}\right) \subset \mathbb{R}^{n-1}. \tag{14.2}$$

Lemma 14.1. [1] *Given* $\{e_i\}_{i\in[m]}$ *satisfying* (14.1) *and* (14.2), K_0, *the intersection of the half-spaces* $K_i := \{x \mid \langle x, e_i\rangle \geq 0\}$, *has* $(0, \ldots, 0, 1) \in \mathbb{R}^n$ *as its center and satisfies* $\operatorname{inrad}(K_0) = \rho$.

Inspired by Lemma 14.1, we construct m centers for a given inradius $\operatorname{inrad}(K_0) \in (0, 1)$ with Algorithm 11 in Appendix B. The reference points $\{e_i\}_{i\in[m]}$ satisfy (14.1) and (14.2), with the first $n-1$ entries of the first $m-1$ reference points generated from the standard normal distribution in $\mathbb{R}^{n-1}$.

14.2 Intersection of second-order cones

Note that in the perceptron setting, we always have $r_0 = 1$. In order to demonstrate the effect of r_0 on the convergence rate of Algorithm 9, we study the intersection

[1] See the proof in Appendix B.

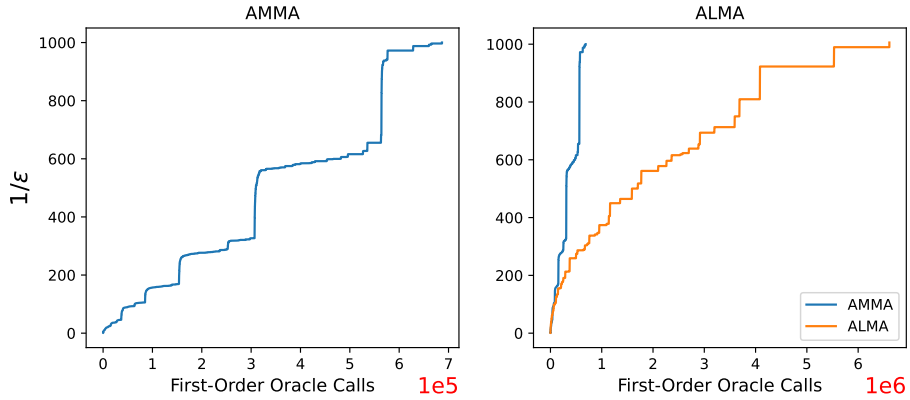

Figure 14.1: Convergence of the proposed Approximate Margin Maximization Algorithm and the Approximate Large Margin Algorithm from [9] when applied to a perceptron problem in $\mathbb{R}^{100}$ with 100 centers and $\operatorname{inrad}(K_0) = 0.1$. The target relative accuracy is $\epsilon = 0.001$. The y-axes in the plots show the reciprocal of the relative suboptimality of the iterates. One sees that the AMMA needs much fewer first-order oracle calls to reach the target relative accuracy. Moreover, the ALMA algorithm clearly demonstrates a $O(\frac{1}{\epsilon^2})$ convergence rate, while the convergence rate of AMMA is closer to a linear one in $O(\frac{1}{\epsilon})$.

of second-order cones of the form

$$K_i := \left\{ x \middle| \|x - \langle x, e_i\rangle e_i\| \leq \frac{r_0}{\sqrt{1-r_0^2}} \langle x, e_i\rangle \right\}. \tag{14.3}$$

Lemma 14.2. [2] *For second-order cones defined by* (14.3)*, we have*

$$\omega_i(x) = \langle x, e_i\rangle - \frac{\sqrt{1-r_0^2}}{r_0}\|x - \langle x, e_i\rangle e_i\|, \ \partial\omega_i(x) = e_i - \frac{\sqrt{1-r_0^2}}{r_0}\frac{x - \langle x, e_i\rangle e_i}{\|x - \langle x, e_i\rangle e_i\|}$$

and

$$r_i(x) = r_0\langle x, e_i\rangle - \sqrt{1-r_0^2}\|x - \langle x, e_i\rangle e_i\|.$$

We generate second-order cone examples with the following procedure, which relies on Algorithm 11 in Appendix B:

Corollary 14.1. [3] *For* $\{e_i\}_{i\in[m]}$ *generated by Algorithm 10 and* K_i *defined by* (14.3)*, we have*

$$r_i(e_i) = r_0, \quad \forall i \in [m].$$

[2] See the proof in Appendix B.
[3] See the proof in Appendix B.

Algorithm 10: Second-Order Cone Generation for Margin Maximization

input : target number of centers m, dimension n, inradius of the individual cones $r_0 \in (0, 1)$ and inradius of the intersection cone $\rho \in (0, r_0)$.

output: a set of centers $\{e_i\}_{i\in[m]}$ such that $\|e_i\| = 1$ for all $i \in [m]$.

compute $\zeta := \sqrt{(1 - r_0^2)(1 - \rho^2)} + r_0\rho$;
run Algorithm 11 with m, n and $\rho = \zeta$ to obtain $\{e_i\}_{i\in[m]}$.

Moreover, the intersection cone $K_0 := \cap_{i\in[m]} K_i$ *has* $(0, \ldots, 0, 1) \in \mathbb{R}^n$ *as its center, and*

$$\mathrm{inrad}(K_0) = \rho.$$

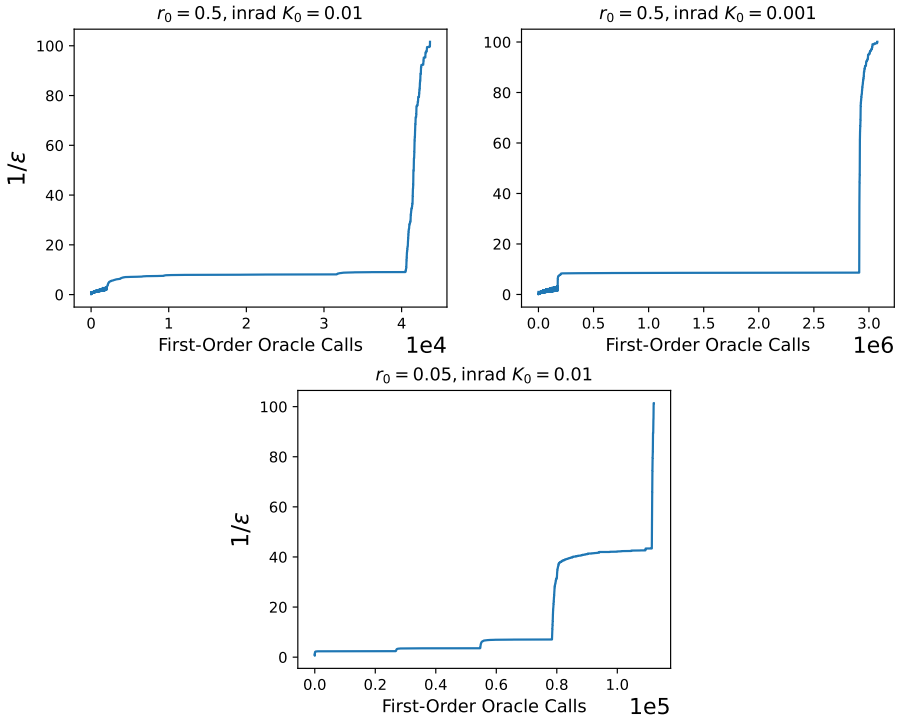

Figure 14.2: Convergence of the AMMA when applied to second-order cone problems in $\mathbb{R}^{100}$ with different r_0 and inrad(K_0). Each problem has 100 second-order cones. The target relative accuracy $\epsilon = 0.01$. Similar to Figure 14.1, the y-axes in Figure 14.2 show the reciprocal of the relative suboptimality of the iterates. One can see that the AMMA demonstrates an $\tilde{O}(\frac{1}{\epsilon})$ rate, which is affected by both r_0 (compare the plots on the top) and inrad(K_0) (compare the top left plot and the bottom plot).

APPENDIX A

ADDITIONAL PROOFS IN PART I

A.1 Additional proofs in Chapter 2

Lemma A.1. *Given a closed convex set $S \subset \mathcal{E}$, for any $x, y \in S$, we have* $\operatorname{recc}(S - x) = \operatorname{recc}(S - y)$.

Proof. Consider any $d \in \operatorname{recc}(S - x)$. We wish to show $y + \lambda d \in S$ for all $\lambda > 0$. Due to the convexity of S, we have $x + n\lambda d \in S$ for all $n \in \mathbb{N}$. Hence by the convexity of S, for all $n \in \mathbb{N}$, we see that

$$y(n) := \frac{ny}{n+1} + \frac{x + n\lambda d}{n+1} = \frac{n(y + \lambda d)}{n+1} + \frac{x}{n+1} \in S.$$

Since $y(n) \to y + \lambda d$ as $n \to \infty$, by the closedness of S, we conclude that $y + \lambda d \in S$. Hence $\operatorname{recc}(S - x) \subseteq \operatorname{recc}(S - y)$. The reversed inclusion can be shown similarly. □

A.2 Additional proofs in Chapter 3

Proof of (3.8). Assume T is a compact convex subset of $\operatorname{int}(\mathcal{T}_{S_0}(z))$. Since for any compact convex set T and $\epsilon > 0$, there exists a polytope P satisfying

$$T \subseteq P \subseteq \{x \mid \operatorname{dist}(x, T) \leq \epsilon\},$$

we may assume

$$T \subseteq P \subset \operatorname{int}(\mathcal{T}_{S_0}(z))$$

for a polytope P with vertices v_j for j in a finite index set J. By definition of the tangent cone, there exists scalars $t_j > 0$ such that $z + tv_j \in \operatorname{int}(S_0)$ for all

$0 < t \leq t_j$. Thus, letting $t' = \min_{j \in J} t_j$, we have

$$z + t' \cdot P \subset S_0,$$

and consequently,

$$z + t' \cdot T \subset S_0.$$

□

Example A.1. To see why the sup on the right-hand side of (3.6) might not be obtainable, consider two circles in $\mathbb{R}^2$:

$$S_1 := \{(x, y) \mid (x-1)^2 + y^2 \leq 1\},\ S_2 := \{(x, y) \mid x^2 + (y-1)^2 \leq 1\}.$$

Then $S_0 = S_1 \cap S_2$ has nonempty interior. Now let $z = (0, 0) \in S_0$. We can show graphically that

$$\mathcal{T}_{S_0}(z) = \{(d_1, d_2) \mid d_1 \leq 0,\ d_2 \leq 0\}.$$

Then $\text{inrad}(\mathcal{T}_{S_0}(z)) = \frac{1}{\sqrt{2}}$, which is obtained by considering $v = (\frac{1}{2}, \frac{1}{2})$ in the definition of $\text{inrad}(\mathcal{T}_{S_0}(z))$.

Now consider any $w = (w_1, w_2) \in S_0$ such that $w \neq z$. Then we have $w_1 > 0$ and $w_2 > 0$. Thus

$$(0-1)^2 + w_2^2 = 1 + w_2^2 > 1 \Longrightarrow (0, w_2) \notin S_1 \subset S.$$

Similarly, We can show $(w_1, 0) \notin S_2$. Hence $r_{S_0}(w) < \min\{w_1, w_2\}$, and

$$\frac{r_{S_0}(w)}{\|w - z\|} < \frac{\min\{w_1, w_2\}}{\sqrt{w_1^2 + w_2^2}} \leq \frac{1}{\sqrt{2}} = \text{inrad}(\mathcal{T}_{S_0}(z)).$$

Example A.2. To see why (3.10) cannot be strengthened in general, consider two sets in $\mathbb{R}^2$:

$$S_1 := \{(x, y) \mid x \geq 0,\ x^2 + y^2 \leq 1\} \cup \{(x, y) \mid x \leq 0,\ |x| \leq 1,\ |y| \leq 1\},$$

$$S_2 := \{(x,y) \mid x \le 0,\ x^2+y^2 \le 1\} \cup \{(x,y) \mid x \ge 0,\ |x| \le 1,\ |y| \le 1\}.$$

Then

$$S_0 = S_1 \cap S_2 = \{(x,y) \mid x^2+y^2 \le 1\}.$$

By considering $w = (0,0)$ on the right-hand side of (3.7), we see that $\mathrm{inrad}(\mathcal{T}_{S_0}(z)) = 1$ for all $z \in \mathrm{bdy}(S_0)$. For any $x = (x_1, x_2) \notin S$, when $x_1 \ge 0$, we see that

$$\mathrm{dist}(x, S_0) = \|x\| - 1 = \mathrm{dist}(x, S_1).$$

Similarly, when $x_1 \le 0$, we have $\mathrm{dist}(x, S_0) = \mathrm{dist}(x, S_2)$. Hence for any $x \notin S_0$,

$$\begin{aligned}\mathrm{dist}(x, S_0) &= \max\{\mathrm{dist}(x, S_1), \mathrm{dist}(x, S_2)\} \\ &= \mathrm{inrad}(\mathcal{T}_{S_0}(\mathrm{P}_{S_0}(x)))\max\{\mathrm{dist}(x, S_1), \mathrm{dist}(x, S_2)\}.\end{aligned}$$

A.3 Additional proofs in Chapter 4

Proof of Lemma 4.3. By Lemma 4.1, we have

$$\begin{aligned}\gamma_0(x) &= \max_{i\in[m]} \gamma_i(x) \\ &\le \max_{i\in[m]} \left\{ \gamma_i(y) + \frac{\|x-y\|}{r_{(S_i,\mathcal{L})}(e_i)} \right\} \\ &\le \left(\max_{i\in[m]} \gamma_i(y) \right) + \frac{\|x-y\|}{\min_{i\in[m]} r_{(S_i,\mathcal{L})}(e_i)} \\ &= \gamma_0(y) + \frac{\|x-y\|}{\min_{i\in[m]} r_{(S_i,\mathcal{L})}(e_i)}.\end{aligned}$$

The reversed inequality can be shown similarly. □

Proof of Lemma 4.4. By Lemma 2.4 and Theorem 3.1, we can write

$$\begin{aligned}\frac{\gamma_0(x)-1}{\operatorname{dist}_{\mathcal{L}}(x,S_0)} &= \frac{\max_{\gamma_i(x)>1}(\gamma_i(x)-1)}{\operatorname{dist}_{\mathcal{L}}(x,S_0)}\\ &\geq \frac{\max_{\gamma_i(x)>1}\left(\operatorname{dist}(x,S_i)/\|\pi_{S_i}(x)-e_i\|\right)}{\operatorname{dist}_{\mathcal{L}}(x,S_0)}\\ &\geq \max_{i\in[m]}\left\{\frac{\operatorname{dist}(x,S_i)}{\operatorname{dist}_{\mathcal{L}}(x,S_0)}\right\}\frac{1}{\max_{\gamma_i(x)>1}\|\pi_{S_i}(x)-e_i\|}\\ &\geq \frac{\operatorname{inrad}_{\mathcal{L}}(\mathcal{T}_{S_0}(z))}{\max_{\gamma_i(x)>1}\|\pi_{S_i}(x)-e_i\|}.\end{aligned}$$

The rest of the proof follows from (2.17). □

A.4 Additional proofs in Chapter 5

Finishing the proof of Theorem 5.1. For $u\in\mathbb{R}$, define

$$f(u) := 2^u-(1+u).$$

Then f is convex. Note that $f(0)=f(1)=0$, for any $u\in[0,1]$, we have $f(u)\leq 0$ and

$$1+u\geq 2^u.$$

Let $\mu := \min_{k'\leq k}\mu_{k'}$. Note that γ_0 is $1/r_0$-Lipshcitz, we get $r_0\mu\in(0,1)$ and $1+(r_0\mu)^2\geq 2^{(r_0\mu)^2}$.

Consequently, when

$$k\geq 2\left(\frac{1}{r_0\mu}\right)^2\log_2\left(\frac{\operatorname{dist}(x_0,S_0)}{\epsilon r_0}\right),$$

we get

$$\begin{aligned}\gamma_0(x_k)-1\leq(1-(r_0\mu)^2)^{\frac{k}{2}}\frac{\operatorname{dist}(x_0,S_0)}{r_0} &\leq \frac{1}{(1+(r_0\mu)^2)^{\frac{k}{2}}}\frac{\operatorname{dist}(x_0,S_0)}{r_0}\\ &\leq \frac{1}{2^{k(r_0\mu)2}}\frac{\operatorname{dist}(x_0,S_0)}{r_0}\\ &\leq \epsilon.\end{aligned}$$

□

A.4.1 Ellipsoid experiments initialized at the mean of the centers

In the experiments presented in Figure A.1, we let $c_i \sim 1000(X/\|X\|) + 100(X_i/\|X_i\|)$, where $X \sim N(\vec{0}, 1)$ and $X_i \sim N(\vec{0}, 1)$. Consequently, we have $\mathbb{E}c_i = 1000X$. The average of the center is away from S_0 and can be used to initialize the algorithms. As can be seen from Figure A.1, the convergence rates of the algorithms are not affected too much by the value of r. This is because $S_0 \neq \emptyset$ even when r is set at 0, and increasing r does not make the intersection much bigger.

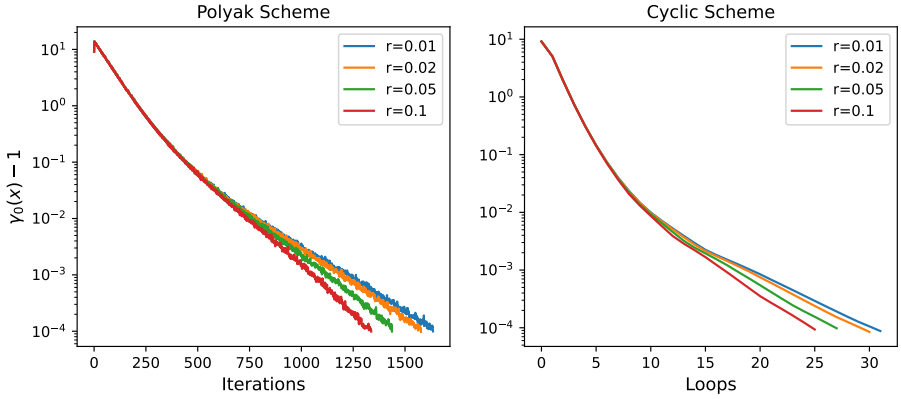

Figure A.1: Algorithm 1 and 2 applied to the problem of the intersection of 100 ellipsoids in $\mathbb{R}^{100}$. For any i, we let $\|B_i\| = \kappa(B_i) = 10$.

A.5 Additional proofs in Chapter 7

Proof of Lemma 7.1. By (2.10), we get

$$\begin{aligned}\frac{\|\mathrm{P}_{S(t)}(x)-e\|}{t} &= \frac{\|\mathrm{P}_{e+t(S-e)}(e+(x-e))-e\|}{t}\\ &= \left\|\mathrm{P}_{e+(S-e)}\left(e+\frac{x-e}{t}\right)-e\right\|\\ &= \|\mathrm{P}_S(x(1/t))-e\|.\end{aligned}$$

Substitute $x(1/t)$ into (2.17), we get

$$\frac{\gamma(x)-t}{\mathrm{dist}(x,S(t))} \geq \frac{1}{\|\pi_S(x(1/t))-e\|} \geq \frac{1}{\|\mathrm{P}_S(x(1/t))-e\|} = \frac{t}{\|\mathrm{P}_{S(t)}(x)-e\|}.$$

□

A.6 Additional proofs in Chapter 8

Example A.3. Consider three sets in $\mathbb{R}^2$: $S_1 = \{(x,y) \mid (x+1)^2+(y+1)^2 \leq 2\}$, $S_2 = \{(x,y) \mid (x-1)^2+(y+1)^2 \leq 2\}$ and $S_3 = \{(x,y) \mid x^2+(y+1)^2 \leq 1\}$. Then $\mathrm{int}(S_0) \neq \emptyset$ and $z = (0,0) \in \mathrm{bdy}(S_0)$. We have

$$\mathcal{N}_{S_1}(z) = \{(t,t) \mid t \geq 0\},\ \mathcal{N}_{S_2}(z) = \{(-t,t) \mid t \geq 0\},\ \mathcal{N}_{S_3}(z) = \{0\} \times \mathbb{R}_+.$$

One can show that

$$\mathrm{reach}_{\{S_i\}_{i\in[m]}}(z) = \cos\left(\frac{\pi}{8}\right) > \sin\left(\frac{\pi}{4}\right) = \mathrm{inrad}(\mathcal{T}_{S_0})(z).$$

Proof of (8.14). For unit vector $w \in \mathcal{N}_{S_0}(z)$ and unit vector $d \in \mathcal{N}_{P_{\{S_i\}_{i\in[m]}}(z)}(\bar{w})$, since

$$\bar{w} \in H(P_{\{S_i\}_{i\in[m]}}(z), d),$$

we get

$$f_{P_{\{S_i\}_{i\in[m]}}(z)}(d) = \operatorname{dist}\left(\vec{0}, H\left(P_{\{S_i\}_{i\in[m]}}(z), d\right)\right) \leq \|\bar{w}\|. \tag{A.1}$$

Hence by (8.12),

$$\begin{aligned}\operatorname{reach}_{\{S_i\}_{i\in[m]}}(z) &= \min_{w\in\mathcal{N}_{S_0},\ \|w\|=1} \|\bar{w}\| \\ &\geq \min_{\substack{w\in\mathcal{N}_{S_0}(z),\ \|w\|=1,\\ d\in\mathcal{N}_{P_{\{S_i\}_{i\in[m]}}(z)}(\bar{w}),\ \|d\|=1}} f_{P_{\{S_i\}_{i\in[m]}}(z)}(d).\end{aligned}$$

On the other hand, note that

$$P_{\{S_i\}_{i\in[m]}}(z) \subseteq H^-(P_{\{S_i\}_{i\in[m]}}(z), d).$$

We have

$$\operatorname{reach}_{\{S_i\}_{i\in[m]}}(z) \leq \operatorname{dist}\left(\vec{0}, \operatorname{bdy}\left(H^-\left(P_{\{S_i\}_{i\in[m]}}(z), d\right)\right)\right) = f_{P_{\{S_i\}_{i\in[m]}}(z)}(d).$$

Taking the infimum over w and d, we get

$$\operatorname{reach}_{\{S_i\}_{i\in[m]}}(z) \leq \min_{\substack{w\in\mathcal{N}_{S_0}(z),\ \|w\|=1,\\ d\in\mathcal{N}_{P_{\{S_i\}_{i\in[m]}}(z)}(\bar{w}),\ \|d\|=1}} f_{P_{\{S_i\}_{i\in[m]}}(z)}(d),$$

and conclude that

$$\operatorname{reach}_{\{S_i\}_{i\in[m]}}(z) = \min_{\substack{w\in\mathcal{N}_{S_0}(z),\ \|w\|=1,\\ d\in\mathcal{N}_{P_{\{S_i\}_{i\in[m]}}(z)}(\bar{w}),\ \|d\|=1}} f_{P_{\{S_i\}_{i\in[m]}}(z)}(d).$$

Now assume the unit vector $w^\star \in \mathcal{N}_{S_0}(z)$ satisfies $\|\bar{w}^\star\| = \operatorname{reach}_{\{S_i\}_{i\in[m]}}(z)$. Then for any unit vector

$$d' \in \mathcal{N}_{P_{\{S_i\}_{i\in[m]}}(z)}(\bar{w}^\star),$$

by (A.1), we have

$$\begin{aligned}\min_{\substack{w\in\mathcal{N}_{S_0}(z),\ \|w\|=1,\\ d\in\mathcal{N}_{P_{\{S_i\}_{i\in[m]}}(z)}(\bar{w}),\ \|d\|=1}} f_{P_{\{S_i\}_{i\in[m]}}(z)}(d) &\le f_{P_{\{S_i\}_{i\in[m]}}(z)}(d')\\ &\le \|\bar{w}^\star\|\\ &= \mathrm{reach}_{\{S_i\}_{i\in[m]}}(z)\\ &= \min_{\substack{w\in\mathcal{N}_{S_0}(z),\ \|w\|=1,\\ d\in\mathcal{N}_{P_{\{S_i\}_{i\in[m]}}(z)}(\bar{w}),\ \|d\|=1}} f_{P_{\{S_i\}_{i\in[m]}}(z)}(d).\end{aligned}$$

Hence $f_{P_{\{S_i\}_{i\in[m]}}(z)}(d') = \|\bar{w}^\star\|$, which implies $\bar{w}^\star = \mathrm{P}_{H(P_{\{S_i\}_{i\in[m]}}(z),d')}(\vec{0})$ and

$$d' = w^\star \in \mathcal{N}_{S_0}(z). \tag{A.2}$$

Note that for any unit vector $d \in \mathcal{N}_{S_0}(z)$, by (8.13), there exists a unit vector $w \in \mathcal{N}_{S_0}(z)$ such that $d \in \mathcal{N}_{P_{\{S_i\}_{i\in[m]}}(z)}(\bar{w})$. Consequently,

$$\begin{aligned}\mathrm{reach}_{\{S_i\}_{i\in[m]}}(z) &= \min_{\substack{w\in\mathcal{N}_{S_0}(z),\ \|w\|=1,\\ d\in\mathcal{N}_{P_{\{S_i\}_{i\in[m]}}(z)}(\bar{w}),\ \|d\|=1}} f_{P_{\{S_i\}_{i\in[m]}}(z)}(d)\\ &\le \min_{d\in\mathcal{N}_{S_0}(z),\ \|d\|=1} f_{P_{\{S_i\}_{i\in[m]}}(z)}(d)\\ &\le f_{P_{\{S_i\}_{i\in[m]}}(z)}(d')\\ &= \mathrm{reach}_{\{S_i\}_{i\in[m]}}(z),\end{aligned}$$

and (8.14) holds. □

A.7 Additional proofs in Chapter 9

Proof of Proposition 9.1. By the definition of $Q_{(\mathrm{C-Opt})}(x^\star)$, we have

$$Q_{(\mathrm{C-Opt})}(x^\star) \subseteq \max\left\{\frac{1}{\min_{i\in\mathcal{I}}\|x^\star - e_i\|}, M\right\}\cdot\overline{B(\vec{0},1)}.$$

Hence following the proof of Theorem 8.2, we can show that $f_{Q_{(\mathrm{C-Opt})}(x^\star)}$ is finite at all unit vectors in $\mathcal{N}_{X^\star}(x^\star)$ and

$$\min_{d\in\mathcal{N}_{X^\star}(x^\star),\ \|d\|=1} f_{Q_{(\mathrm{C-Opt})}(x^\star)}(d) > 0.$$

Similar to (8.14), we can show that

$$\mathrm{reach}_{(\mathrm{C-Opt})}(x^\star) = \min_{d\in\mathcal{N}_{X^\star}(x^\star),\ \|d\|=1} f_{Q_{(\mathrm{C-Opt})}(x^\star)}(d). \tag{A.3}$$

For any $d \in \mathcal{E}$, like (8.15), we have

$$f_{Q_{(\mathrm{C-Opt})}(x^\star)}(d) = \max\left\{\max_{i\in\mathcal{I}}\left\{\langle d, d_i\rangle \mid d_i \in \mathcal{N}_{S_i}(x^\star), \|d_i\| = \frac{1}{\|x^\star - e_i\|}\right\}, \max\{\langle d, g\rangle \mid g \in \partial f(x^\star)\}\right\} \tag{A.4}$$

Consider $x \neq x^\star$, then by (8.18),

$$\gamma_0(x) - 1 \geq \max_{i\in\mathcal{I}}\{\gamma_i(x) - 1\} \geq \max_{i\in\mathcal{I}}\left\{\max_{d\in\mathcal{N}_{S_i}(x^\star),\|d\|=1}\left\{\frac{\langle x - x^\star, d\rangle}{\|x^\star - e_i\|}\right\}\right\}.$$

Also note that

$$f(x) - f(x^\star) \geq \max\{\langle x - x^\star, g\rangle \mid g \in \partial f(x^\star)\}.$$

Let $d = \frac{x-x^\star}{x-x^\star}$ in (A.4), then $x - x^\star \in \mathcal{N}_{X^\star}(x^\star)$, and we see that

$$\begin{aligned}\max\{\gamma_0(x) - 1, f(x) - f(x^\star)\} &\geq f_{Q_{(\mathrm{C-Opt})}(x^\star)}(x - x^\star) \\ &= \|x - x^\star\| \cdot f_{Q_{(\mathrm{C-Opt})}(x^\star)}\left(\frac{x - x^\star}{\|x - x^\star\|}\right) \\ &\overset{(A.3)}{\geq} \mathrm{reach}_{(\mathrm{C-Opt})}(x^\star)\|x - x^\star\|.\end{aligned}$$

□

APPENDIX B

ADDITIONAL PROOFS AND ALGORITHMS IN PART II

Proof of Lemma 14.1. Since $\vec{0} \in \operatorname{conv}\left(\{(e_1^1, \ldots, e_1^{n-1}), \cdots, (e_m^1, \ldots, e_m^{n-1})\}\right)$, there exists convex multipliers $\{\lambda_i\}_{i\in[m]}$ such that

$$\sum_{i\in[m]} \lambda_i \left(e_i^1, \ldots, e_i^{n-1}\right) = \vec{0} \in \mathbb{R}^{n-1}.$$

Consequently, for any $x \in \mathbb{R}^n$, we get

$$\begin{aligned}
\min_{i\in[m]} \langle x, e_i\rangle &\leq \sum_{i\in[m]} \lambda_i \langle x, e_i\rangle \\
&= \left\langle \left(x^1, \ldots, x^{n-1}\right), \sum_{i\in[m]} \lambda_i \left(e_i^1, \ldots, e_i^{n-1}\right) \right\rangle + \rho x^n \\
&= \rho x^n.
\end{aligned}$$

Hence $(0, \ldots, 0, 1)$ is the center of K_0, and $\operatorname{inrad}(K_0) = \rho$. □

Remark. In our experiments, we use the Gurobi package [12] to compute the projection of $\vec{0}$ onto $\operatorname{conv}\left(\{e_i'\}_{i\in[m-1]}\right)$. One can see that the centers generated by Algorithm 11 satisfy (14.1) and (14.2).

Proof of Lemma 14.2. In order to study the properties of the second-order cones whose centers are generated by Algorithm 10, first note that the dual cones of the second-order cones defined in (14.3) take the form

$$K_i^\star := \left\{ x \middle| \|x - \langle x, e_i\rangle e_i\| \leq \frac{\sqrt{1-r_0^2}}{r_0} \langle x, e_i\rangle \right\}. \tag{B.1}$$

Consequently, for any $y \in K^\star$ such that $\langle y, e_i\rangle = 1$, we can write $y = e_i + u$, where

$$\langle u, e_i\rangle = 0, \ \|u\| \leq \frac{\sqrt{1-r_0^2}}{r_0}.$$

Algorithm 11: Half-Space Generation for Margin Maximization

input : target number of centers m, dimension n and the shared inradius of all the intersection cone $\rho \in (0, 1)$.

output: a set of centers $\{e_i\}_{i\in[m]}$ such that $\|e_i\| = 1$ for all $i \in [m]$.

for $i \in [m-1]$ **do**

  generate vector $\tilde{e}_i$ following the standard normal distribution in $\mathbb{R}^{n-1}$;

  compute $e'_i = \frac{\tilde{e}_i}{\|\tilde{e}_i\|}$;

compute $u = \mathrm{P}_{\mathrm{conv}\left(\{e'_i\}_{i\in[m-1]}\right)}\left(\vec{0}\right)$;

if $u = \vec{0}$ **then**

  generate vector $\tilde{e}_m$ following the standard normal distribution in $\mathbb{R}^{n-1}$;

  compute $e'_m = \frac{\tilde{e}_m}{\|\tilde{e}_m\|}$;

else

  let $e'_m = -\frac{u}{\|u\|}$;

for $i \in [m]$ **do**

  let $e_i = \left(\sqrt{1-\rho^2}e'^1_i, \ldots, \sqrt{1-\rho^2}e'^{n-1}_i, \rho\right)$.

Now for any $x \in \mathbb{R}^n$, define

$$v(x) := \langle x, e_i\rangle e_i, \quad w(x) := x - v(x).$$

Then $\langle w(x), e_i\rangle = 0$, and

$$\begin{aligned}
&\min\{\langle x, y\rangle \mid y \in K^\star, \langle y, e_i\rangle = 1\} \\
=\ &\min\left\{\langle v(x) + w(x), e_i + u\rangle \middle| \langle u, e_i\rangle = 0, \|u\| \le \frac{\sqrt{1-r_0^2}}{r_0}\right\} \\
=\ &\min\{\langle x, e_i\rangle + \langle w(x), e_i\rangle + \langle x, e_i\rangle\langle e_i, u\rangle + \langle w(x), u\rangle \mid \\
&\qquad\qquad \langle u, e_i\rangle = 0, \|u\| \le \frac{\sqrt{1-r_0^2}}{r_0}\Bigg\} \\
=\ &\langle x, e_i\rangle - \frac{\sqrt{1-r_0^2}}{r_0}\|w(x)\|,
\end{aligned}$$

and the minimum is obtained when

$$u = \frac{\sqrt{1-r_0^2}}{r_0}\frac{w(x)}{\|w(x)\|}.$$

Since K_i is a cone, we have

$$\begin{aligned} r_i(x) &= \min\{\langle x, y\rangle \mid y \in K_i^\star, \langle y, e_i\rangle = 1\} \\ &= \min\{\langle v(x) + w(x), v(y) + w(y)\rangle \mid y \in K_i^\star, \langle y, e_i\rangle = 1\} \\ &= \min\{\langle v(x), v(y)\rangle + \langle w(x), w(y)\rangle \mid y \in K_i^\star, \langle y, e_i\rangle = 1\}. \end{aligned} \tag{B.2}$$

When

$$v(y) = r_0 e_i,\ w(y) = -\sqrt{1 - r_0^2}\frac{w(x)}{\|w(x)\|},$$

the right-hand side of Theorem 13.3 obtains its minimum

$$r_0\|v(x)\| - \sqrt{1 - r_0^2}\|w(x)\|.$$

for all $x \in K_i$. □

Proof of Corollary 14.1. The first claim follows immediately from Lemma 14.2. To show the second claim, let $a \in [0, \pi/2)$ satisfy $\sin(a) = r_0$. For any x such that $\|x\| = 1$, let $b_i(x) \in [0, \pi)$ satisfy

$$\cos(b_i(x)) = \langle x, e_i\rangle,\quad \sin(b_i(x)) = \|x - \langle x, e_i\rangle e_i\|.$$

Then we get

$$r_i(x) = \sin(a)\cos(b_i(x)) - \cos(a)\sin(b_i(x)) = \sin(a - b_i(x)).$$

Similar to Lemma 14.1, we have

$$\min_{i\in[m]} \cos(b_i(x)) \le \zeta x^n$$

for any $x \in \mathbb{R}^n$ such that $\|x\| = 1$. Note that

$$\zeta = \cos(a - \arcsin(\rho)),$$

we have

$$\begin{aligned}\operatorname{inrad}(K_0) &= \max_{x\in\mathbb{R}^n,\|x\|=1} \min_{i\in[m]} r_i(x) \\ &\leq \sin(a - \arccos(\zeta)) \\ &= \sin(a - (a - \arcsin(\rho))) \\ &= \sin(\arcsin(\rho)) \\ &= \rho,\end{aligned}$$

and the maximum is obtained when $x = (0, \ldots, 0, 1) \in \mathbb{R}^n$. □

BIBLIOGRAPHY

[1] Shmuel Agmon. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6:382–392, 1954.

[2] Heinz H Bauschke and Jonathan M Borwein. On projection algorithms for solving convex feasibility problems. *SIAM review*, 38(3):367–426, 1996.

[3] Heinz H Bauschke, Jonathan M Borwein, and Wu Li. Strong conical hull intersection property, bounded linear regularity, jameson's property (g), and error bounds in convex optimization. *Mathematical Programming*, 86(1):135–160, 1999.

[4] Amir Beck and Marc Teboulle. Convergence rate analysis and error bounds for projection algorithms in convex feasibility problems. *Optimization Methods and Software*, 18(4):377–394, 2003.

[5] Alexandre Belloni, Robert M Freund, and Santosh Vempala. An efficient rescaled perceptron algorithm for conic systems. *Mathematics of Operations Research*, 34(3):621–641, 2009.

[6] 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.

[7] Yu-Hong Dai. Fast algorithms for projection on an ellipsoid. *SIAM Journal on Optimization*, 16(4):986–1006, 2006.

[8] Robert M Freund and Jorge R Vera. Condition-based complexity of convex optimization in conic linear form via the ellipsoid algorithm. *SIAM Journal on Optimization*, 10(1):155–176, 1999.

[9] Claudio Gentile. A new approximate maximal margin classification algorithm. *Journal of Machine Learning Research*, 2(Dec):213–242, 2001.

[10] Jean-Louis Goffin. The relaxation method for solving systems of linear inequalities. *Mathematics of Operations Research*, 5(3):388–414, 1980.

[11] LG Gubin, Boris T Polyak, and EV Raik. The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, 7(6):1–24, 1967.

[12] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2022.

[13] Alan J Hoffman. On approximate solutions of systems of linear inequalities'. *Journal of Research of the National Bureau of Standards*, 49(4):263, 1952.

[14] Zehui Jia, Xingju Cai, and Deren Han. Comparison of several fast algorithms for projection onto an ellipsoid. *Journal of Computational and Applied Mathematics*, 319:320–337, 2017.

[15] Adam Kowalczyk. Maximal margin perceptron. *Advances in Large Margin Classifiers*, pages 75–113, 2000.

[16] A Lin and SP Han. Projection on an ellipsoid. Technical report, Research report, Department of Mathematical Sciences, The Johns Hopkins ..., 2001.

[17] Anhua Lin and Shih-Ping Han. A class of methods for projection on the intersection of several ellipsoids. *SIAM Journal on Optimization*, 15(1):129–138, 2004.

[18] Ion Necoara and Angelia Nedić. Minibatch stochastic subgradient-based projection algorithms for feasibility problems with convex inequalities. *Computational Optimization and Applications*, 80(1):121–152, 2021.

[19] Angelia Nedić. Random projection algorithms for convex set intersection problems. In *49th IEEE Conference on Decision and Control (CDC)*, pages 7655–7660. IEEE, 2010.

[20] Javier Peña and Negar Soheili. A deterministic rescaled perceptron algorithm. *Mathematical Programming*, 155(1-2):497–510, 2016.

[21] Boris T Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1:32, 1987.

[22] Boris T Polyak. Random algorithms for solving convex inequalities. In *Studies in Computational Mathematics*, volume 8, pages 409–422. Elsevier, 2001.

[23] James Renegar. "efficient" subgradient methods for general convex optimization. *SIAM Journal on Optimization*, 26(4):2649–2676, 2016.

[24] James Renegar and Benjamin Grimmer. A simple nearly-optimal restart scheme for speeding-up first order methods. *arXiv preprint arXiv:1803.00151*, 2018.

[25] James Renegar and Song Zhou. A different perspective on the stochastic convex feasibility problem. *arXiv preprint arXiv:2108.12029*, 2021.

[26] R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1970.

[27] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[28] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[29] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[30] John von Neumann. The geometry of orthogonal spaces, functional operators-vol. ii. *Annals of Math. Studies*, 22, 1950. Reprint of mimeographed lecture notes first distributed in 1933.