APPLICATIONS OF ACCURACY CERTIFICATES FOR PROBLEMS WITH CONVEX STRUCTURE

A Thesis

Presented to

The Academic Faculty

by

Bruce Cox

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

School of Industrial and Systems Engineering

Georgia Institute of Technology

May 2011

APPLICATIONS OF ACCURACY CERTIFICATES FOR PROBLEMS WITH CONVEX STRUCTURE

Approved by:

Professor Arkadi Nemirovski, Advisor
H. Stewart Milton School of
Industrial and Systems Engineering
Georgia Institute of Technology

Professor Shabbir Ahmed
H. Stewart Milton School of
Industrial and Systems Engineering
Georgia Institute of Technology

Professor Santanu Dey
H. Stewart Milton School of
Industrial and Systems Engineering
Georgia Institute of Technology

Professor Alexander Shapiro
H. Stewart Milton School of
Industrial and Systems Engineering
Georgia Institute of Technology

Professor Eric Feron
The Daniel Guggenheim School of
Aerospace Engineering
Georgia Institute of Technology

Date Approved: 14 January 2011

.

For my wife and son

Acknowledgements

There are too many people I would like to thank and acknowledge, I shall pare it down and list them in reverse chronological order. First I would like to thank my Advisor Professor Arkadi Nemirovski, whose been more patient with me then I deserve and taught me how to *think* about optimization, as well as teaching me about optimization. I'd also like to thank all of my other professors at Georgia Tech, each of whom has helped shaped me in positive fashions. I like to particularly note Professors Shapiro, Ahmed, Feron, and Dey, the members of my committee. Finally I would like to thank Mr Wilkinson, my 5th grade teacher – the first teacher I had who believed in me and helped me believe in myself.

TABLE OF CONTENTS

Summary

This dissertation addresses the efficient generation and potential applications of accuracy certificates in the framework of "black-box-represented" convex optimization problems — convex problems where the objective and the constraints are represented by "black boxes" which, given on input a value x of the argument, somehow (perhaps in a fashion unknown to the user) provide on output the values and the derivatives of the objective and the constraints at x. The main body of the dissertation can be split into three parts. In the first part, we provide our background — state of the art of the theory of accuracy certificates for black-box-represented convex optimization. In the second part, we extend the toolbox of black-box-oriented convex optimization algorithms with accuracy certificates by equipping with these certificates a state-of-the-art algorithm for *large-scale* nonsmooth black-box-represented problems with convex structure, specifically, the Non-Euclidean Restricted Memory Level (NERML) method. In the third part, we present several novel academic applications of accuracy certificates.

The dissertation is organized as follows:

In Chapter 1, we motivate our research goals and present a detailed summary of our results.

In Chapter 2, we outline the relevant background, specifically, describe four generic black-box-represented generic problems with convex structure (Convex Minimization, Convex-Concave Saddle Point, Convex Nash Equilibrium, and Variational Inequality with Monotone Operator), and outline the existing theory of accuracy certificates for these problems.

In Chapter 3, we develop techniques for equipping with on-line accuracy certificates the state-of-the-art NERML algorithm for large-scale nonsmooth problems with convex structure, both in the cases when the domain of the problem is a simple solid and in the case when the domain is given by Separation oracle.

In Chapter 4, we develop several novel academic applications of accuracy certificates, primarily to (a) efficient certifying emptiness of the intersection of finitely many solids given by Separation oracles, and (b) building efficient algorithms for convex minimization over solids given by Linear Optimization oracles (both precise and approximate).

In Chapter 5, we apply accuracy certificates to efficient decomposition of "well structured" convex-concave saddle point problems, with applications to computationally attractive decomposition of a large-scale LP program with the constraint matrix which becomes block-diagonal after eliminating a relatively small number of possibly dense columns (corresponding to "linking variables") and possibly dense rows (corresponding to "linking constraints").

CHAPTER I

MOTIVATION, GOALS, OVERVIEW OF RESULTS

1.1 Motivation

Certificates, explicitly representable entities used to prove a hypothesis, are ubiquitous throughout mathematics and computer science. They range from the common, such as the existence of $\sqrt{2}$ which proves the hypothesis that the set of irrational numbers is nonempty, to field specific; such as the *Theorem of Alternative* which states that a system of strict and nonstrict linear inequalities has no solution if and only if one of two other systems of strict and nonstrict linear inequalities, explicitly given by the original system, has a solution. Certificates are quite prevalent throughout Optimization, the best known being the above mentioned Theorem of the Alternative, and Farkas's Lemma, though *accuracy certificates*, the focus of this thesis, are widely utilized.

At its most basic an accuracy certificate is an "easy to represent entity," like a vector or a finite collection of vectors, which allows one to justify in advance and in a "fully algorithmic" fashion, a desired conclusion. As far as Optimization is concerned, the simplest examples of certificates are given by Linear Programming. Specifically,

1. The simplest way to certify feasibility of a Linear Programming problem

$$\mathrm{Opt} = \min_x \left\{ c^T x : Ax \geq b \right\} \tag{1}$$

   is to point out a vector $\bar{x}$ which satisfies the system of constraints $Ax \geq b$. Indeed, given a candidate certificate $\bar{x}$ of this type, it is easy to verify its validity (that is, to check whether $A\bar{x} \geq b$); given that the certificate is valid, we do know that (1) is feasible.

2. The simplest way to justify *infeasibility* of (1) is to point out a vector $\lambda \geq 0$ such that $A^T\lambda = 0$ and $\lambda^T b > 0$. Here again, given a candidate certificate, it is easy to check its validity, and if the latter does take place, (1) definitely is infeasible (since assuming that x is such that $Ax \geq b$, we would have $0 = \lambda^T Ax \geq \lambda^T b > 0$, which is a contradiction). Moreover, by General Theorem on Alternative, (1) is infeasible *if and only if* there exists an infeasibility certificate of the outlined type.

3. The simplest way to certify that $\bar{x}$ is a feasible and ϵ-optimal (i.e., with $c^T x - \mathrm{Opt} \leq \epsilon$)

solution to (1) is to augment x by a feasible solution $\bar{\lambda}$ to the dual problem

$$\max_{\lambda} \left\{ b^T \lambda : \lambda \geq 0, A^T \lambda = c \right\}$$

such that $c^T \bar{x} - b^T \bar{\lambda} \leq \epsilon$. By Linear Programming Duality Theorem, a feasible solution $\bar{x}$ to (1) is ϵ-optimal if and only if it admits such a certificate of ϵ-optimality.

The third example in the above list is closely related to the main topic of our research – this is an *accuracy certificate*, a certificate for the property of a feasible solution $\bar{x}$ to an LP program to be ϵ-optimal. As we see, in the case of LP such an accuracy certificate is given by a solution $\bar{\lambda}$ for the dual problem which makes a small ($\leq \epsilon$) duality gap $c^T \bar{x} - b^T \bar{\lambda}$ with the primal solution $\bar{x}$. Note that the standard accuracy certificates in Conic Programming (see, e.g., [2]), which is a natural extension of LP, have similar structure – they are feasible solutions $\bar{\lambda}$ to the conic dual of the conic program of interest such that the duality gap, as evaluated at $\bar{\lambda}$ and the primal feasible solution $\bar{x}$ ϵ-optimality of which we want to certify, is at most ϵ.

1.2 Summary of Thesis Goals and Results

In contrast to the just outlined "well structured" case, for a long time there was no well defined general notion of an "accuracy certificate" for a *black box represented* convex problem, that is, convex problem where the objective and the constraints are represented by a "black box" which, given on input a value x of the argument, somehow (perhaps in a fashion unknown to the user) provides on output the values and the derivatives of the objective and the constraints at x. Note that this "black box represented setting" is the most traditional and the most general way to pose a *nonlinear* convex optimization problem. The general notion of accuracy certificate for a "black box represented problem with convex structure" was worked out only recently [27]. The primary goal of this Thesis is *further development of the theory of accuracy certificates for black-box-represented problems with convex structure, with emphasis on*

- *developing computationally cheap techniques for building "good" accuracy certificates for large-scale problems with convex structure, and*
- *investigating a spectrum of novel applications of accuracy certificates.*

1.2.1 Preliminaries: Black Box Represented Problems with Convex Structure and Accuracy Certificates

In order to outline in more details the major goals and results of the Thesis, we start with a brief summary of the background, originating from [27], on problems with convex

structure and their accuracy certificates. To simplify our summary, we sometimes impose on the situation in question more restrictions than in the main body of Thesis; for detailed presentation of the background, see Chapter 2.

1.2.1.1 Problems with Convex Structure

Problems with convex structure as defined in [27] are *Convex Minimization problems*, *Convex-Concave Saddle Point problems*, *Convex Nash Equilibrium problems*, and *Variational Inequalities with Monotone Operators*. These problems are posed as follows.

Convex Minimization problem: Given a solid (closed convex and bounded set with a nonempty interior) $X \subset \mathbf{R}^n$ and a continuous real-valued convex function f on X, solve the problem

$$\mathrm{Opt} = \min_{x \in X} f(x) \tag{2}$$

of minimizing f over X, that is, find a *minimizer* $x^* \in X$ of f on X such that $f(x) \geq f(x^*)$ for all $x \in X$.

A *black box representation* of (2) is given by

- a *Separation Oracle* for X – a routine which, given on input $x \in \mathbf{R}^n$, reports whether $x \in \mathrm{int}X$, or $x \in \partial X$, or $x \notin X$, and in the last two cases returns a *separator* $e \in \mathbf{R}^n$ of x and X such that
$$e \neq 0 \ \& \ e^T x \geq \max_{x' \in X} e^T x';$$
- a *First Order oracle* for f – a routine which, given on input a point $x \in \mathrm{int}X$, returns the value $f(x)$ and a subgradient $f'(x)$ of f at X.

The first Order oracle defines, in particular, the vector field

$$F(x) = f'(x) : \mathrm{int}X \to \mathbf{R}^n.$$

We measure the inaccuracy of a candidate solution $x \in X$ to problem (2) as

$$\varepsilon_{\mathrm{opt}}(x) = f(x) - \mathrm{Opt} = f(x) - \min_{x' \in X} f(x').$$

Convex-Concave Saddle Point problem : Given two solids $X_1 \subset \mathbf{R}^{n_1}$, $X_2 \subset \mathbf{R}^{n_2}$ and a continuous cost function $\phi(x_1, x_2) : X_1 \times X_2 \to \mathbf{R}$ which is convex in $x_1 \in X_1$, and concave in $x_2 \in X_2$, solve the saddle point problem

$$\mathrm{SadVal} = \min_{x_1 \in X_1} \max_{x_2 \in X_2} \phi(x_1, x_2), \tag{3}$$

that is, find a *saddle point* $x^* = (x_1^*, x_2^*) \in X = X_1 \times X_2$ of ϕ on $X_1 \times X_2$ such that $\phi(x_1, x_2^*) \geq \phi(x_1^*, x_2^*) \geq \phi(x_1^*, x_2)$ for all $(x_1, x_2) \in X_1 \times X_2$.

A convex-concave saddle point problem gives rise to a primal-dual pair of convex optimization problems

$$\begin{array}{rcl} \mathrm{Opt}(P) & = & \min_{x_1 \in X_1} \overline{f}(x_1) := \max\limits_{x_2 \in X_2} \phi(x_1, x_2) \\ \mathrm{Opt}(D) & = & \max\limits_{x_2 \in X_2} \underline{f}(x_2) := \min\limits_{x_1 \in X_!} \phi(x_1, x_2) \end{array} \tag{4}$$

with equal optimal values:

$$\mathrm{Opt}(P) = \mathrm{Opt}(D) = \mathrm{SadVal}.$$

Saddle points of ϕ on $X_1 \times X_2$ are exactly the pairs (x_1^*, x_2^*) where x_1^* solves the primal, and x_2^* solves the dual problem.

A *black box representation* of (4) is given by

- a Separation oracle for the domain $X = X_1 \times X_2 \subset \mathbf{R}^{n_1+n_2}$ of the problem
- a First Order oracle representing ϕ. Given on input $x = (x_1, x_2) \in \mathrm{int}X$, this oracle returns the value $\phi(x_1, x_2)$ and a vector $F(x) = [F_1(x); F_2(x)]$, where $F_1(x)$ is a subgradient of $\phi(\cdot, x_2)$ taken at x_1, and $F_2(x)$ is a subgradient of $-\phi(x_1, \cdot)$ taken at x_2.

The First Order Oracle defines, in particular, the vector field

$$F(x) = F(x_1, x_2) : \mathrm{int}X = \mathrm{int}X_1 \times \mathrm{int}X_2 \to \mathbf{R}^{n_1+n_2}.$$

We measure the inaccuracy of a candidate solution $x = (x_1, x_2) \in X = X_1 \times X_2$ to problem (4) as

$$\begin{array}{rcl} \varepsilon_{\mathrm{sad}}(x) & = & \max\limits_{x_2' \in X_2} f(x_1, x_2') - \min\limits_{x_1' \in X} f(x_1', x_2) = \overline{f}(x_1) - \underline{f}(x_2) \\ & = & [\overline{f}(x_1) - \mathrm{Opt}(P)] + [\mathrm{Opt}(D) - \underline{f}(x_2)]. \end{array}$$

Convex Nash Equilibrium problem is to find, given k solids $X_i \subset \mathbf{R}^{n_i}$, $1 \leq i \leq n$, and k continuous functions $\phi_i(x_1, ..., x_k) : X_1 \times ... \times X_k \to \mathbf{R}$ such that

- ϕ_i is convex in $x_i \in X_i$ and concave in $x^i = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_k) \in X^i = X_1 \times ... \times X_{i-1} \times X_{i+1} \times ... \times X_k$, $1 \leq i \leq k$, and
- $\sum_i \phi_i(x)$ is convex on $X = X_1 \times ... \times X_k$,

a *Nash Equilibrium*, that is, a point $x^* = (x_1^*, ..., x_k^*)$ such that for every i, x_i^* is a minimizer, over $x_i \in X_i$, of the function $\phi_i(x_1^*, ..., x_{i-1}^*, x_i, x_{i+1}^*, ..., x_k^*)$.

A *black box representation* of Nash Equilibrium problem is given by

- a Separation oracle for the domain $X = X_1 \times ... \times X_k \subset \mathbf{R}^{n_1+n_2+...+n_k}$ of the problem, and

- a First Order oracle representing ϕ.
  Given on input $x = (x_1, ..., x_k) \in \mathrm{int}X$, this oracle returns the values $\phi_i(x)$ and subgradients $[\phi_i]'_i(x)$ of the functions $\phi_i(x_1, ..., x_{i-1}, \cdot, x_{i+1}, ..., x_k)$ taken at x_i, $1 \le i \le k$.

The First Order Oracle defines, in particular, the vector field

$$F(x) = F(x_1, x_2, ..., x_k) = [\phi'_1(x); ...; \phi'_k(x)] : \mathrm{int}X = \mathrm{int}X_1 \times ... \times \mathrm{int}X_k \to \mathbf{R}^{n_1+n_2+...+n_k}.$$

We measure the inaccuracy of a candidate solution $x = (x_1, ..., x_k) \in X = X_1 \times ... \times X_k$ to the Nash Equilibrium problem as

$$\varepsilon_{\mathrm{Nash}}(x) = \sum_{i=1}^{k} \left[\phi_i(x) - \min_{x'_i \in X_i} \phi(x^1, ..., x^{i-1}, x'_i, x^{i+1}, ..., x^k) \right].$$

Note: The standard interpretation of Nash equilibrium problem is that there are k players choosing simultaneously their actions x_i in solids X_i; when the actions of all k players form a vector $x = (x_1, ..., x_k) \in X = X_1 \times ... \times X_k$, i-th player incurs loss $\phi_i(x)$. Nash Equilibrium is a choice $x^* \in X$ of the players where every single player has no incentive to change unilaterally his choice, and $\varepsilon_{\mathrm{Nash}}(x)$ is the sum, over players, of their incentives to change their choices from x_i to $\mathrm{argmin}_{x'_i \in X_i} \phi(x_1, ..., x_{i-1}, x'_i, x_{i+1}, ..., x_k)$.

Remark 1.2.1. *Note that a Convex Minimization problem can be considered as a Nash Equilibrium problem with single player, while convex-concave saddle point problem can be considered as a 2-player Nash Equilibrium problem with* $\phi_1(x_i, x_2) = -\phi_2(x_1, x_2) = \phi(x_1, x_2)$. *Note that the entities we have associated with the three problems in question – Separation and First Order oracles, the vector fields and the accuracy measures – are "compatible" with these identifications; e.g., the vector field and the accuracy measure associated with a convex-concave saddle point problem (3) remain intact when we treat this problem as a two-player Nash equilibrium problem with* $\phi_1 = -\phi_2 = \phi$.

Variational Inequalities with Monotone Operators. In such a problem, one is given a solid $X \subset \mathbf{R}^n$ and a vector field $F(x) : \mathrm{int}X \to \mathbf{R}^n$, which is *monotone*:

$$\langle F(x) - F(x'), x - x' \rangle \ge 0 \ \forall x, x' \in \mathrm{int}X,$$

and the goal is to find a (weak) solution to the variational inequality given by X, F, i.e., a point $x^* \in X$ such that

$$\langle F(x), x - x^* \rangle \ge 0 \ \forall x' \in \mathrm{int}X.$$

A *black box representation* of a variational inequality with monotone operator is given by

- a Separation oracle for the domain X of the problem, and
- a First Order oracle representing F. Given on input $x \in \mathrm{int}X$, this oracle returns $F(x)$.

We measure the inaccuracy of a candidate solution $x \in X$ to a variational inequality given by X, F by the quantity

$$\varepsilon_{\mathrm{vi}}(x) = \sup_{x' \in \mathrm{int}X} \langle F(x'), x - x' \rangle$$

which is sometimes called "dual gap function" [9].

Intermediate summary. Note that Variational Inequality with Monotone Operator is in fact a "common denominator" of all four problems with convex structure as defined above: the vector field F which we have associated with every one of the four problems are monotone on the interior of problem's domain X, and the solutions to the problem are nothing but the weak solutions to the variational inequality given by (X, F). This observation does not mean that instead of investigating four generic problems, we could investigate just the most general of them, the variational inequality with monotone operator. The reason for separate treatment of separate problems is that in reality we are interested in *approximate* solutions and thus – in accuracy measures, and in this respect the problems are not equivalent. E.g., the "optimization inaccuracy" $\varepsilon_{\mathrm{opt}}(x)$ associated with a candidate solution x of a convex minimization problem (2) is *not* the same as the "variational inaccuracy" $\varepsilon_{\mathrm{vi}}(x)$ of x considered as an approximate solution to the variational inequality associated with (2); in fact, $\varepsilon_{\mathrm{vi}}(x)/\varepsilon_{\mathrm{opt}}(x) \leq 1$, and the ratio can be quite large.

1.2.1.2 Accuracy Certificates

Black box oriented algorithms and Execution protocols. By definition, a *black box oriented algorithm* for solving a problem with convex structure is a procedure which, given access to the Separation and the First Order oracles associated with this problem, generates subsequent *search points* $x_1, x_2, ... \in \mathbf{R}^n$ (here $\mathbf{R}^n$ is an embedding space of the domain X of the problem being solved), and at a search point x_t calls the Separation oracle, x_t being the input, and then, if possible (that is, if it turns out that $x_t \in \mathrm{int}X$), calls the First Order oracle, x_t being the input. While we do not restrict the abilities of an algorithm to "learn" the problem under consideration by those offered by the Separation and the First Order oracles, we do assume that at every search point x_t the Separation oracle was invoked, and at every search point $x_t \in \mathrm{int}X$, in addition, the First Order oracle is invoked. It follows that after a number τ of steps the algorithm has at its disposal the *execution protocol* $P_\tau = (\{x_t, e_t\}_{t=1}^{\tau}, I_\tau, J_\tau)$, where

- x_t, $1 \leq t \leq \tau$, are the search points generated in course of τ steps,

- $I_\tau \subset \{1, ..., \tau\}$ and $J_\tau \subset \{1, ..., \tau\}$ are, respectively, the sets of indexes t of *productive* ($x_i \in \mathrm{int}X$) and *nonproductive* ($x_t \notin \mathrm{int}X$) search points with indexes $\leq \tau$, and
- $e_t = F(x_t)$ for $t \in I_\tau$ (note that at a productive step t, the First Order oracle is invoked, so that $F(x_t)$ indeed becomes known), and $e_t \neq 0$ separates x_t and X for $t \in J_\tau$ (note that at a nonproductive step t, the Separation oracle does report a separator e_t of x_t and X).

Accuracy certificate. An *accuracy certificate* ξ^τ is defined in terms of a *productive* (with $I_\tau \neq \emptyset$) execution protocol P_τ and is, by definition, a collection $\{\xi_t^\tau\}_{t=1}^\tau$ of *nonnegative* reals such that

$$\sum_{t \in I_\tau} \xi_t^\tau = 1.$$

The *approximate solution* associated with an execution protocol P_τ and accuracy certificate ξ^τ is defined as

$$\widehat{x}(P_\tau, \xi^\tau) = \sum_{t \in I_\tau} \xi_t^\tau x_t;$$

note that this solution belongs to $\mathrm{int}X$ (as a convex combination of the points $x_t \in \mathrm{int}X$, $t \in I_\tau$.

Given a solid $\mathbf{B}$ known to contain the domain X of the problem of interest, we define the *resolution* of an accuracy certificate ξ^τ as the quantity

$$\varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}) = \max_{x \in \mathbf{B}} \sum_{t=1}^{\tau} \xi_t^\tau \langle e_t, x_t - x \rangle.$$

Basic properties of accuracy certificates. At a first glance, the notion of an accuracy certificate seems highly bizarre and completely unrelated to certification. Nevertheless, this notion does work. The first – and the major – argument in its favor is given by the following result.

Theorem 2.2.1[27] *Let a pair (X, F) originate from a problem with convex structure, that is, from*

(a) a convex optimization problem,

(b) a convex-concave saddle point problem,

(c) a convex Nash equilibrium problem, or

(d) a variational inequality with a monotone operator.

Let P_τ be an execution protocol for (X, F), let ξ^τ be an accuracy certificate associated with this protocol, and let $\widehat{x} = \widehat{x}(P_\tau, \xi^\tau)$.

Then:
(i) $\widehat{x} \in X$, so that $\widehat{x}$ is a feasible solution to the problem underlying X, F, and
(ii) for every solid $\mathbf{B} \supset X$ the resolution $\varepsilon_{\rm cert}(\xi^\tau | P_\tau, \mathbf{B})$ of ξ^τ w.r.t. P_τ and $\mathbf{B}$ is an upper bound on the corresponding accuracy measure of $\widehat{x}$, so that

- $\varepsilon_{\rm opt}(\widehat{x}) \le \varepsilon_{\rm cert}$ *in the case of (a),*
- $\varepsilon_{\rm sad}(\widehat{x}) \le \varepsilon_{\rm cert}$ *in the case of (b),*
- $\varepsilon_{\rm Nash}(\widehat{x}) \le \varepsilon_{\rm cert}$ *in the case of (c), and*
- $\varepsilon_{\rm vi}(\widehat{x}) \le \varepsilon_{\rm cert}$ *in the case of (d).*

Building accuracy certificates. Theorem 2.2.1 says, roughly speaking, that *when solving a problem with convex structure,* an accuracy certificate with small resolution is a "highly valuable commodity" it allows for both

• generating a strictly feasible approximate solution to the problem with inaccuracy, defined in terms of the problem, not exceeding the resolution of the certificate on (any) convex solid $\mathbf{B}$ known to contain the domain X of the problem.

• certifying the quality of the above solution. Indeed, given an execution protocol, it is easy to check whether a candidate accuracy certificate is a valid one. In the latter case, assuming that the solid $\mathbf{B}$ is simple (like box, or ball, or simplex), meaning that it is easy to minimize a linear function over $\mathbf{B}$, it is easy to compute the resolution of the certificate and thus it is easy to access the "level of non-optimality" of the (automatically strictly feasible!) solution yielded by the certificate.

Taking alone, these observations still do not explain why accuracy certificates are of actual interest; to the latter end, they should be computable at a computationally reasonable price. The "ideal" situation here would be as follows: Given a generic problem (i.e., family of instances) $\mathcal{P}$ with convex structure and a "prototype" black box oriented algorithm $\mathcal{B}$ for solving problems from $\mathcal{P}$ with a complexity estimate $T_{\mathcal{B}}(\epsilon)$ (that is, it takes at most $T_{\mathcal{B}}(\epsilon)$ steps of the algorithm to solve any instance of $\mathcal{P}$ within accuracy ϵ), we can augment $\mathcal{B}$ with on-line rules for building accuracy certificates in such a way that

A: computing the certificates is cheap – it does not increase significantly the average, over the steps, computational effort per step;

B: the certificates are consistent with the complexity bound, meaning that for every $\epsilon > 0$

and every instance of $\mathcal{P}$, the number of steps of the augmented method in which a certificate with resolution $\leq \epsilon$ is built, does not exceed $O(1)T_{\mathcal{B}}(\epsilon)$.

Taken together, these two properties say that when augmenting $\mathcal{B}$ with an appropriate technique for building the accuracy certificates and terminating the resulting algorithm when for the first time an accuracy certificate with resolution not exceeding a required tolerance ϵ is built, we neither increase significantly the computational effort per step, nor spoil the complexity bound.

Now, the above "ideal situation" indeed takes place for *some* of known black box oriented methods for solving problems with convex structure. Specifically, in the hindsight one understands that the accuracy certificates satisfying the above requirements indeed are built in the simplest Subgradient Descent and Mirror Descent methods (originating, respectively, from [38, 37] and [24, 26], see also [3]) and in the full memory Bundle-Level method [21]. On the other hand, the most interesting, at least from the theoretical viewpoint, *polynomial time cutting plane algorithms for problems with convex structure,* most notably, the Ellipsoid method (see, e.g., [46, 12, 2] and references therein) and the Inscribed Ellipsoid method [41], by themselves do *not* produce accuracy certificates. For example, the standard result on the Ellipsoid method is as follows (see, e.g., [2]):

> **Fact I:** *Let the n-dimensional domain X of a Convex Minimization problem (2) be contained in the centered at the origin Euclidean ball $\mathbf{B}$ of a given radius R, and let the objective f of the problem be convex and Lipschitz continuous, with a known constant L taken w.r.t. $\|\cdot\|_2$, on X. Then for every $\epsilon > 0$, the number of steps of the Ellipsoid method resulting in a strictly feasible approximate solution x_ϵ of inaccuracy $\varepsilon_{\rm opt}(x_\epsilon) \leq \epsilon$, does not exceed $T(\epsilon) = O(1)n^2 \ln\left(\frac{LR^2}{r\epsilon} + 2\right)$, where $O(1)$ is an absolute constant and r is the largest among the radii of Euclidean balls contained in X. For every τ, the computational effort at the first τ steps, modulo τ calls to the Separation oracle and at most τ calls to the First Order oracle, does not exceed $O(1)\tau n^2$ operations.*

In this result, the approximate solution x^τ generated in course of τ steps is defined as the best (with the smallest value of f) among the strictly feasible search points x_t generated in course of the first τ steps. While the method is equipped with on-line termination rules (which do not require a priori knowledge of L, although do require a priori knowledge of r), these rules do not use accuracy certificates. Note that the option of choosing, as an approximate solution, the best, in terms of the objective, of the strictly feasible search points generated so far, exists only when solving convex minimization problems (and requires from the First Order oracle to report the values of the objective, and not only the subgradients) and does not exist when solving other problems with convex structure, e.g., convex-concave saddle point problems. In the latter case, the traditional polynomial time black box oriented algorithms *at best* end up with a large-scale convex minimization problem with piecewise

linear objective such that a good approximate solution to this problem induces an ϵ-solution to the problem of interest. In the hindsight, this phenomenon can be explained by the fact that the methods in question do not produce on line accuracy certificates, and the above "large-scale convex minimization problem with piecewise linear objective" is, again in the hindsight, nothing but the problem of building an accuracy certificate with the best possible resolution for the execution protocol built at the first phase of the method. Note that solving this auxiliary problem from scratch can be much more time-consuming than the first phase of the algorithm.

The issue of equipping polynomial time cutting plane algorithms with accuracy certificates satisfying the requirements A and B was fully resolved in [27]. For example, here is the result on Ellipsoid method with accuracy certificates from [27]:

> **Fact II:** *Let X be an n-dimensional solid given by a Separation oracle and contained in the centered at the origin Euclidean ball $\mathbf{B}$ of known radius R, and let $F : \mathrm{int}X \to \mathbf{R}^n$ be a vector field, represented by the First Order oracle computing $F(x)$ at a query point $x \in \mathrm{int}X$. Assume that F is semibounded:*
>
> $$V := \sup_{x \in \mathrm{int}X, y \in X} \langle F(x), y - x \rangle < \infty.$$
>
> *Then the Ellipsoid method, as applied to (X, F), can be augmented with on-line techniques for building accuracy certificates in such a way that*
>
> - *for every tolerance $\epsilon > 0$, the number of steps before an accuracy certificate with resolution $\le \epsilon$ on $\mathbf{B}$ is built, does not exceed $T_a(\epsilon) = O(1)n^2 \ln\left(\frac{RV}{r\epsilon} + 2\right)$, where, same as above, r is the largest of the radii of Euclidean balls contained in X, and*
>
> - *for every τ, the computational effort at the first τ steps, modulo τ calls to the Separation oracle and at most τ calls to the First Order oracle, does not exceed $O(1)\tau n^2$ operations.*

Note that in the situation of Convex Minimization problem $\min_{x \in X} f(x)$, the subgradient field F of the objective f is semibounded on X with $V \le \max_X f - \min_X f$, that is, under the premise of Fact I one has $V \le 2LR$. It follows that the complexity bound stemming from Fact II can be only better than the bound stemming from Fact I. At the same time, Fact II has significant advantages as compared to Fact I:

1. First and foremost, it allows to equip the Ellipsoid method with accuracy certificates with the on-line termination rule "terminate when the resolution of the current accuracy certificate on $\mathbf{B}$ becomes $\le \epsilon$, and output the approximate solution yielded by this certificate," $\epsilon > 0$ being the desired tolerance. This rule works equally well for *all* problems with convex structure

2. Even in the Convex Minimization case, the termination rule above is "less demanding" than the termination rules associated with the standard Ellipsoid method — now we neither require Lipschitz continuity of the objective, nor the necessity to know the values of the objective at strictly feasible search points, nor (and this is most important) the quantity r.

1.2.2 Goals and Main Results of the Thesis

After background on problems with convex problems and accuracy certificates is presented, we can formulate in more details the goals and outline the main results of the Thesis.

The summary of our goals can be expressed by a single statement: we intend to provide novel and important theoretical evidence of the fact that machinery of accuracy certificates does work. More specifically, our goals are twofold:

- We intend to demonstrate that the state of the art black-box-oriented methods for solving *large scale* problems with convex structure, same as the polynomial time cutting plane algorithms, can be augmented with on line techniques for building accuracy certificates which satisfy the above requirements A (computational simplicity) and B (compatibility with the standard complexity bounds);

- We intend to present a spectrum of novel and important by their own right applications of accuracy certificates.

We are about to present the outlined goals in more details and to outline the relevant results of the Thesis.

1.2.2.1 Novel Techniques for Building Accuracy Certificates for Large-Scale Problems with Convex Structure

Goal and Motivation As it was already explained, in order to make the machinery of accuracy certificates indeed useful, one needs to know how to augment "good" – admitting attractive complexity estimates – black box oriented methods with computationally cheap techniques for building accuracy certificates compatible with these complexity estimates. This crucial problem was resolved in [27] for theoretically most important *polynomial time cutting plane algorithms*, like the Ellipsoid methods; besides this, we have already mentioned that this problem, essentially does not arise when speaking about some other "good" algorithms, primarily, Subgradient- and Mirror Descent algorithms and the full memory Bundle Level methods – the accuracy certificates are, in the hindsight, built into the latter algorithms from the very beginning. This, however, falls short of the present state of black-box-oriented machinery for problems with convex structure, specifically, does

not cover the most attractive bundle type algorithms for extremely large-scale problems of this type. The situation here is as follows: whatever be the theoretical importance of polynomial time cutting plane algorithms, these algorithms can be considered as practical computational tools only in the low-dimensional case (at most few tens of decision variables). The Subgradient- and Mirror Descent algorithms are better suited for large-scale problems and, moreover, as applied to nonsmooth problems with favorable geometry, exhibit the best possible *worst-case* behavior in the large scale black box setting. From the practical viewpoint, however, a severe shortcoming of these algorithms is that their *typical* convergence rate is more or less the same as the worst-case one — the accuracy goes to zero with the number τ of steps as slowly as $1/\sqrt{\tau}$, which is really slow from the practical viewpoint. The usual way to improve practical behavior of Subgradient- and Mirror Descent type algorithms is to pass to their *bundle* versions, where the first order information obtained at previous steps is efficiently utilized at subsequent steps as well[1]. However, the only bundle-type algorithm which, in the hindsight, produces accuracy certificates is the full memory Bundle-Level algorithm [21], and this algorithm, because of the necessity to keep *the full* first order information accumulated so far (or, in the most advanced implementations, something like $2n$ "pieces" of this information, n being the design dimension of the problem), is well suited for medium-size problems (n of order of at most few hundreds). The bottom line is that *at present, as far as large-scale (n in the range of thousands and more) nonsmooth problems with convex structure are concerned, no general purpose black-box-oriented algorithms with accuracy certificates are known.* This is a severe shortcoming of the existing results on accuracy certificates, since from the practical perspective, large-scale (tens and hundreds of thousands of decision variables) general-type problems with convex structure form the most important area of applications of black-box-oriented optimization techniques. One of the major goals of the Thesis was to overcome this severe shortcoming. To this end, we have investigated the possibilities to equip with accuracy certificates the most attractive state-of-the-art bundle-type algorithm for solving large-scale nonsmooth problems with convex structure – the Non-Euclidean Restricted Memory Level (NERML) algorithm proposed in [4][2]. The corresponding developments form the subject of Chapter 3 – this first fully original chapter of the Thesis.

Results. Consider a problem with convex structure, and let X be its domain, and F be the associated monotone operator. In order to solve the problem by NERML (without or with certificates), one should equip the embedding space $E = \mathbf{R}^n$ of X with a norm $\|\cdot\|$ (not necessarily the Euclidean one) and, in addition, specify

[1] Note that the Subgradient- and Mirror Descent algorithms "as they are" are *memoryless* – all information accumulated so far is summarized in a single entity – the current search point. Bundle versions of gradient type algorithms take their origin in [17, 23, 20] and over the years were the subject of numerous studies, see, e.g., [39, 21, 18, 19] and references therein.

[2] The Euclidean version of this algorithm was proposed in [18].

- a solid $\mathbf{B} \subset E$ which contains X, and
- a *distance-generating function* $\omega(x) : X \to \mathbf{R}$, which should be continuously differentiable and strongly convex, modulus $\alpha > 0$, w.r.t.

$$\langle \omega'(x) - \omega'(x'), x - x' \rangle \geq \alpha \|x - x'\|^2.$$

In order for the NERML algorithm associated with the above setup to be practical, $\mathbf{B}$ and $\omega(\cdot)$ should be simple and fit each other, meaning that it is easy to solve auxiliary problems of the form

$$\min_{x \in \mathbf{B}} \left[\omega(x) + a^T x \right]. \tag{5}$$

We consider two cases:

Case I: X is simple, specifically, $X = \mathbf{B}$. We assume also that the monotone operator F (which we usually treat as defined on int$X$ only is in fact defined on the entire X [3]. Our main result here (see Proposition 3.3.1) is as follows:

Proposition *In addition to the assumptions we have just made, let there exist $c \in X = \mathbf{B}$, $\mathbf{V} < \infty$, and a $\Theta > 0$ such that*

$$\forall (x \in \mathbf{B}, y : \|y\| \leq 1) : \langle F(x), c + \Theta y - x \rangle \leq \mathbf{V}, \tag{6}$$

e.g., F is bounded on $X = \mathbf{B}$, in which case one can choose $c \in X$ arbitrarily and set

$$\Theta = R(\mathbf{B}),\ \mathbf{V} = 2\|F\|_* R(\mathbf{B}),\ \|F\|_* = \max_{x \in X} \|F(x)\|_*,\ R(\mathbf{B}) = \max_{x, x' \in \mathbf{B}} \|x - x'\|,$$

$\|\cdot\|_*$ being the norm conjugate to $\|\cdot\|$.

As applied to (X, F), the prototype NERML algorithm from [4] can be equipped with on-line techniques for building accuracy certificates in such a way that

- *for every $\epsilon > 0$, a certificate with resolution $\leq \epsilon$ on $X = \mathbf{B}$ is built after at most*

$$T(\epsilon) = O(1) \frac{\Omega(\mathbf{V} + \epsilon)^2}{\alpha \Theta^2 \epsilon^2} \tag{7}$$

steps, where α is the modulus of strong convexity of $\omega(\cdot)$ w.r.t. $\|\cdot\|$ and

$$\Omega = \max_{u, v \in \mathbf{B}} \left[\omega(u) - \omega(v) - \langle \omega'(v), u - v \rangle \right].$$

[3] For example, in the case of Convex Minimization we assume that the First Order oracle provides subgradients of the objective both at the interior and at the boundary points of the domain X.

- *the computational effort per step is dominated by the necessity to compute F at a point and to solve two auxiliary problems, one of the form*

$$\max_{x\in\mathbf{B}} \min_{1\le i\le m+1} h_i(x),$$

and the other one of the form

$$\min_{x\in\mathbf{B}} \left\{\omega(x) + a^T x : h_i(x) \le 0,\ 1 \le i \le m+1\right\}$$

where $h_i(\cdot)$ are given affine functions and a nonnegative integer m is a parameter of the construction ("memory depth") of the algorithm.

It should be stressed that the structure of a step and the computational effort per step of the NERML algorithm with certificates are completely similar to those of the prototype algorithm without certificates as presented in [4], and the complexity bound (7) is, within an absolute constant factor, the same as for the prototype.

Case II: $X \subset \mathbf{B}$ is given by Separation oracle. Here we assume neither that X is simple, nor that F is defined on the entire X; instead, we assume that F is defined on $\mathrm{int}X$, and is semibounded:

$$\forall(x \in \mathrm{int}X, y \in X) : \langle F(x), y - x\rangle \le \mathbf{V}$$

with known $\mathbf{V} < \infty$. In addition, we assume that X contains a $\|\cdot\|$-ball of radius $r > 0$.

Our main result here (see Proposition 3.4.1) is as follows:

Proposition *Under assumptions we just have made, one can point out a NERML-type algorithm equipped with on-line techniques for building accuracy certificates in such a way that*

- *for every $\epsilon > 0$, a certificate with resolution $\le \epsilon$ on $\mathbf{B} \supset X$ is built after at most*

$$T(\epsilon) = O(1)\frac{\Omega R^2(\mathbf{B})(\mathbf{V}+\epsilon)^2}{\alpha r^4 \epsilon^2} \tag{8}$$

steps, with the same as above α, Ω and $R(\mathbf{B})$;

- *the computational effort per step is dominated by one call to the Separation oracle representing X, at most one call to the First Order oracle computing F, and solving the same auxiliary problems as in Case I.*

Note that even aside of the issue of accuracy certificates, Case II goes beyond the scope of the prototype NERML algorithm as presented in [4].

1.2.2.2 Novel Academic Applications of Accuracy Certificates

Chapter 4 of the Thesis is devoted to several novel academic applications of accuracy certificates. We are about to outline these applications.

Certifying emptiness of intersection of convex sets. Let $X_1, ..., X_m$ be convex solids in $\mathbf{R}^n$ given by Separation oracles and known to belong to the centered at the origin Euclidean ball V_R of a given radius R. It is easy to certify that the intersection $\cap_i X_i$ of these solids is nonempty: a certificate is just a point $\bar{x}$ in $\cap_i X_i$, and the validity of a candidate certificate is easy to check; indeed, all we need in order to verify that $\bar{x} \in \cap_i X_i$ is to call the Separation oracles for X_i, $1 \leq i \leq m$, $\bar{x}$ being the input. A nontrivial (and, to the best of our knowledge, open) question is how to certify that the intersection of $X_1, ..., X_m$ is empty. We resolve this challenging issue as follows. Let

$$X = X_1 \times ... \times X_m \subset \mathbf{R}^{mn}.$$

Observe that Separation oracles for X_i straightforwardly induce a Separation oracle $\mathcal{S}$ for X.

1. We demonstrate (Proposition 4.1.1) that *if we can point out a finite collection of points $w_s \in \mathbf{R}^{mn}$ such that $w_s \notin \mathrm{int}X$ along with vectors $\eta_s = [\eta_s^1; ...; \eta_s^m] \in \mathbf{R}^{mn}$ and nonnegative weights ζ_s, $1 \leq s \leq S$, such that η_s is the separator of w_s and X as reported by $\mathcal{S}$ invoked at w_s, so that*

$$\langle \eta_s, w_s \rangle \geq \sup_{w \in X} \langle \eta_s, w \rangle,$$

and the linear inequality

$$\begin{gathered} \langle P\left[\sum_{s=1}^{S} \zeta_s \eta_s\right], y \rangle \leq \sum_{s=1}^{S} \zeta_s \langle \eta_s, w_s \rangle \\ \left[P[x^1; ...; x^m] = \sum_{i=1}^{m} x^i : \mathbf{R}^{mn} \to \mathbf{R}^n\right] \end{gathered} \tag{9}$$

in variable $y \in \mathbf{R}^n$ has no solutions in the ball V_R, then $\cap_i X_i = \emptyset$.

The above result states that a collection $\{w_s, \eta_s, \zeta_s\}$ with the outlined properties can be considered as a certificate of emptiness of $\cap_i X_i$; note that the validity of a candidate certificate of this type is easy to check: given w_s, we check whether η_s indeed is a separator of w_s and X by calling $\mathcal{S}$, and given $\{w_s, \eta_s, \zeta_s\}$, it is easy to verify that (9) has no solutions in V_R, since the latter merely means that

$$-R\|P\left[\sum\nolimits_{s=1}^{S} \zeta_s \eta_s\right]\|_2 > \sum_{s=1}^{S} \zeta_s \langle \eta_s, w_s \rangle.$$

2. We demonstrate further (Proposition 4.1.2) that the *sufficient* condition for emptiness of $\cap_i X_i$ as expressed in the previous statement is in fact *necessary and sufficient*: $\cap_i X_i$ is empty *if and only if* there exists an emptiness certificate of the just outlined type.

3. Finally, we demonstrate (Proposition 4.1.2) that in order to certify emptiness of $\cap_i X_i$, it suffices to apply to the Convex Minimization problem

$$\mathrm{Opt} = \min_{x \in X} f(x) := \frac{1}{2} \sum_{i=1}^{m} \|x^i - x^{i+1}\|_2^2, \qquad (P)$$

where $x^{m+1} \equiv x^1$, an algorithm with accuracy certificates. Specifically, we show that

(a) an accuracy certificate for the latter problem satisfying certain easy to verify condition can be straightforwardly converted into an emptiness certificate as defined above;

(b) if $\cap_i X_i = \emptyset$ (or, equivalently, $\mathrm{Opt} > 0$), then every accuracy certificate with resolution $< 2\mathrm{Opt}$ on $\mathbf{B} = V_R \times ... \times V_R$ definitely satisfies the "easy to verify condition" in (a) and thus implies an emptiness certificate.

It follows that if $\cap_i X_i$ indeed is empty and the algorithm with certificates used to solve (P) converges (meaning that the resolution of the associated certificates goes to 0 as the number of steps grows), an emptiness certificate will be eventually found. For example, when $\mathrm{Opt} > 0$ and (P) is solved by the Ellipsoid method with certificates, the emptiness will be certified not later than in $O(1)(mn)^2 \ln\left(\frac{nR^3}{r\mathrm{Opt}}\right)$ steps, where $r = \min_i r(X_i)$ and $r(X_i)$ is the largest of the radii of Euclidean balls contained in X_i.

Convex Minimization under Linear Optimization oracle. In the standard black box setting of a Convex Minimization problem

$$\mathrm{Opt} = \min_{x \in X} f(x), \qquad (2)$$

the feasible domain X of the problem is a solid represented by a Separation oracle, and the objective f is represented by the First Order oracle. Now, representation by a Separation oracle is not the only natural way of describing a solid $X \subset \mathbf{R}^n$; another natural way to represent X is via a *Linear Optimization* (LO) oracle. The latter, given on input a vector $e \in \mathbf{R}^n$, returns a minimizer x_e of the linear function $e^T x$ over $x \in X$. There are situations (see examples in section 4.2.3) where representation by an LO oracle is the only one available (or is much more preferable from the computational viewpoint than the representation by a Separation oracle). Now, it is well known (see, e.g., [12]) that given a solid X, both Separation and LO oracles are polynomially reducible to each other: roughly speaking, with LO oracle available, one can mimic a single call to Separation oracle via a polynomial time number of calls to the LO oracle, and vice versa. This equivalence, however, is primarily of

theoretical value: mimicking Separation oracle via an LO one, while being a polynomially solvable task, usually is too computationally expensive to be practical. We demonstrate (section 4.2) that when solving a Convex Minimization problem (2), LO representation is no less "practical" than the Separation one (and can be even much more practical than the latter), *provided that f admits a Fenchel-type representation*, that is, representation of the form

$$f(x) = \max_{y \in Y} \left\{ x^T(Ay + a) - h(y) \right\},$$

where Y is a solid in some $\mathbf{R}^m$ given by a Separation oracle, and $h : Y \to R$ is a convex continuous function given by a First Order oracle. Specifically, we propose the following strategy for solving the problem of interest (2):

- Given Fenchel-type representation of f, the problem of interest is nothing but the primal optimization problem associated with the convex-concave saddle point problem

  $$\mathrm{SadVal} = \mathrm{Opt} = \min_{x \in X} \max_{y \in Y} \left[x^T(Ay + a) - h(y) \right]$$

  The dual to (2) problem induced by this saddle point reformulation of the problem of interest is

  $$-\mathrm{Opt} = \min_{y \in Y} g(y) := h(y) - \min_{x \in X} x^T(Ay + a) \qquad (D)$$

  (it is convenient for us to rewrite the dual problem, which by itself is a maximization program, in the equivalent minimization form).

- Observe that the LO oracle for X clearly induces the First Order oracle for the convex function $f_*(y) = -\min_{x \in X} x^T(Ay + a)$: given y, we form the linear form $-(Ay + a)^T x$ of x and call the LO oracle for X to get a maximizer x_y of this form over X, thus getting $f_*(y) = -x_y^T(Ay + a)$ and $f'_*(y) = -A^T x_y$. Since the First Order oracle for h also is given to us (as a part of the Fenchel-type representation of f), we see that (D) is equipped with both a Separation oracle for Y (it is the remaining part of the Fenchel-type representation of f) and a First Order oracle (which is obtained by combining the LO oracle for X and the First Order oracle for h). The bottom line is that we can solve the dual problem (D) by a black-box-oriented method.

In order for the outlined strategy to be meaningful, we should answer two questions:

(?.a): *How wide is the family of convex functions admitting explicit Fenchel-type representations, as compared to the family of functions admitting explicit representations by First Order oracles?*

(?.b): *How to convert a good approximate solution to the dual problem (D) into an equally good approximate solution to the problem of actual interest(2) ?*

The first of these questions has to do with the scope of potential applications of the outlined strategy, which, of course, is an important consideration. The second question is not merely

important, it is crucial — without possibility to convert good solutions to (D) into equally good solutions to the problem of interest, the proposed strategy is no more than wishful thinking.

Fortunately, it turns out that both questions $(?.a, b)$ admit quite satisfactory answers. Specifically,

(!.a): While the *existence* of Fenchel-type representation of a convex function is not an issue at all — *every* lower semicontinuous function of this type admits even the "pure" Fenchel representation $f(x) = \sup_{y \in Y} \left\{x^T y - f_*(y)\right\}$, *availability* of such a representation could be an issue (e.g., pure Fenchel representations available in closed analytic form, or easy-to-compute algorithmic form, are a rare commodity). Surprisingly, availability of *Fenchel-type* representations, in contrast to pure ones, is *not* an issue. Specifically, we present in section 4.2.2 a simple, albeit important by its own right, *calculus* of Fenchel-type representations which demonstrates that all basic convexity-preserving operations with convex functions (e.g., taking linear combinations with nonnegative coefficients, maxima of finite families, and affine substitution of variables) preserves availability of Fenchel-type representations: given these representations for the operands, one can easily convert them into a Fenchel-type representation of the result. Since the Fenchel-type representations of "basic" convex functions (like the exponent and other elementary univariate functions) is not an issue (for these functions, the pure Fenchel representation is easy to compute), it follows that to assume availability of explicit and easy-to-compute Fenchel-type representation of a convex function is no more restrictive than assuming availability of easy to compute First Order oracle.

(!.b): It turns out – and this is the main result of section 4.2 – that the crucial issue (?.b) can be fully resolved via the machinery of accuracy certificates. Specifically, we demonstrate (Theorem 4.2.1) that when the dual problem (D) is solved by a black box oriented method with accuracy certificates, the latter allow to convert "certifiably good" approximate solutions to (D) to equally good feasible approximate solutions to (2), namely, *an accuracy certificate ξ^τ for (D) induces a feasible solution $x^\tau = \sum_{t \in I_\tau} \xi_t^\tau x_{y_t}$ such that* $\varepsilon_{\text{opt}}(x^\tau) \leq \varepsilon_{\text{cert}}(\xi^\tau | P_\tau, \mathbf{B})$. Here $P_\tau = (\{y_t, e_t\}_{t=1}^\tau, I_\tau, J_\tau)$ is the execution protocol for (D) underlying the certificate ξ^τ, $x_y \in X$ is the vector reported by the LO oracle when computing the first order information for (D) at a point $y \in Y$, and $\mathbf{B}$ is a solid containing the domain Y of (D).

The outlined results of section 4.2 are further extended to cover

- Convex Minimization problems with functional constraints, that is, problems of the form
$$\min_{x \in X} \{f_0(x) : f_i(x) \leq 0, 1 \leq i \leq m\}$$
where the solid X is given by an LO oracle, and convex functions $f_0(x), ..., f_m(x)$ –

by Fenchel-type representations (section 4.3.1);

- the situation where the domain X of (2) is given as an intersection of solids represented by LO oracles.

  When intersecting finitely many solids given by Separation oracles, these oracles straightforwardly induce a Separation oracle for the intersection of the solids. In contrast to this, when intersection solids given by LO oracles, there is no simple way to get from these oracles an LO oracle for the intersection. This is why this situation needs a dedicated treatment, presented in section 4.3.2;

- the situation of problem (2) with f given by Fenchel-type representation and *computationally intractable* solid X equipped with *approximate* LO oracle (section 4.3.3).

  To illustrate the contents of section 4.3.3, consider an instructive particular case of the situation investigated in this section, specifically, as follows. Let $X \subset \mathbf{R}^n$ be a solid, and Ξ be a collection of vectors from $\mathbf{R}^n$. Assume that we know how to maximize in polynomial time linear functions $\langle \xi, x\rangle$ given by $\xi \in \Xi$ over $x \in X$ within some *approximation ratio* $\alpha \in (0, 1]$. In other words, we have at our disposal a polynomial time algorithm $\mathcal{B}$ which, given on input $\xi \in \Xi$, returns a vector $\widehat{x}_\xi \in X$ such that $\langle \xi, x_\xi\rangle \geq \alpha \mathrm{Opt}(\xi)$, $\mathrm{Opt}(\xi) = \max\limits_{x\in X} \langle \xi, x\rangle$. Note that this assumption implies (at least for $\alpha < 1$) that $\mathrm{Opt}(\xi) \geq 0$ for every $\xi \in \Xi$, which we assume from now on. The question is, to which extent these approximation guarantees can be extended from maximizing over X *linear* functions $\langle \xi, \cdot\rangle$ with $\xi \in \Xi$ to maximizing over X *concave* functions ψ which in some sense are "generated" by the linear functions in question. In the case under consideration the main result of section 4.3.3 – Theorem 4.3.1 — states that if the convex function $-\psi$ admits a Fenchel-type representation

$$-\psi(x) = \max_{y\in Y} [-\langle x, Ay + a\rangle - h(y)] \quad [\Leftrightarrow \psi(x) = \min_{y\in Y} [\langle x, Ay + a] + h(y)]]$$

with convex *nonnegative* $h(y)$ and such that $Ay + a \in \Xi$ for all $y \in Y$, then, applying an algorithm with accuracy certificates to (Y, F) with

$$F(y) = -A^T \widehat{x}_{Ay+a} - h'(y),$$

an accuracy certificate ξ^τ for an execution protocol $P_\tau = (\{y_t, e_t\}_{t=1}^\tau, I_\tau, J_\tau)$ induces a feasible approximate solution $x^\tau = \sum_{t\in I_\tau} \xi_t^\tau \widehat{x}_{y_t}$ to the problem of interest

$$\mathrm{Opt} = \max_{x\in X} \psi(x)$$

such that

$$\psi(x^\tau) \geq \alpha \mathrm{Opt} - \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}),$$

$\mathbf{B}$ being a solid containing Y. Specifying the algorithm in question as, e.g., the Ellipsoid algorithm with certificates, we conclude that the problem of maximizing ψ over X within approximation factor $\alpha' < \alpha$ is polynomially solvable.

To give an instructive example, consider the set $X \subset \mathbf{S}^n$ comprised of all convex combinations of dyadic $n \times n$ matrices uu^T with $\|u\|_\infty \leq 1$; here $\mathbf{S}^n$ is the linear space of $n \times n$ symmetric matrices equipped with the Frobenuis inner product. The set X is known to be heavily computationally intractable: for a general-type positive semidefinite matrix $\xi \in \mathbf{S}^n$, it is NP-hard to maximize the linear form $\mathrm{Tr}(\xi x)$ over $x \in X$ within relative accuracy like 4%, even when randomized algorithms are allowed. On the other hand, Nesterov's $\pi/2$ Theorem [32] states that if Ξ is the positive semidefinite cone in $\mathbf{S}^n$, then Semidefinite relaxation allows to maximize over X efficiently, within approximation factor $2/\pi$, every linear form $\mathrm{Tr}(\xi x)$ given by $\xi \in \Xi$. Applying the above result to the concave function $\psi(x)$ defined as the sum of k smallest eigenvalues of a matrix $x \in \mathbf{S}^n$ and utilizing the Fenchel representation of this function:

$$\psi(x) = \max_{y \in Y} \mathrm{Tr}(yx), \;\; Y = \{y \in \mathbf{S}^n : 0 \preceq y \preceq I, \mathrm{Tr}(y) = k\},$$

we conclude that for every $\alpha' < 2/\pi$, one can efficiently find a matrix $\bar{x} \in X$ such that $\psi(\bar{x}) \geq \alpha' \max_{x \in X} \psi(x)$.

1.2.2.3 Accuracy Certificates and Decomposition of Large-Scale Linear Programs

The concluding chapter 5 of the Thesis, while still theoretical in its nature, is motivated by "fully practical" considerations stemming from decomposition of large-scale LPs.

Motivation. Consider a large scale *solvable* Linear Programming program

$$\mathrm{Opt} = \min_{x=[x_1;x_2]} \left\{ c_1^T x_1 + c_2^T x_2 : \begin{array}{r} A^{11}x_1 + A^{12}x_2 \leq b_1 \\ A^{21}x_1 + A^{22}x_2 \leq b_2 \\ \|x\|_\infty \leq R \end{array} \right\} \qquad \text{(LP)}$$

with $n = n_1 + n_2$ variables $x = [x_1; x_2]$, and $m = m_1 + m_2$ linear inequality constraints, where the sizes of A^{11}, A^{12} and A^{21} are, $m_1 \times n_1$, $m_1 \times n_2$ and $m_2 \times n_1$ respectively. Note that the bounds on variables $\|x\|_\infty \leq R$, which we add for technical reasons, are of no real importance from the practical viewpoint, since R can be arbitrarily large. The Lagrange function of the problem (with the bounds of variables not included into the list

of constraints) is

$$\begin{array}{rl} & L(x_1, x_2; y_1, y_2) \\ = & [c_1 + [A^{11}]^T y_1 + [A^{21}]^T y_2]^T x_1 + [c_2 + [A^{21}]^T y_1 + [A^{22}]^T y_2]^T x_2 : X \times Y^\infty \to \mathbf{R} \\ & \left[\begin{array}{c} X = X_1 \times X_2,\, Y^\infty = Y_1^\infty \times Y_2^\infty, \\ X_i = \{x_i : \|x_i\|_\infty \le R\},\ Y_i^\infty = \{y_i \ge 0\},\ i = 1, 2. \end{array} \right] \end{array} \tag{10}$$

Note that the saddle points of this bilinear (and thus convex-concave) function on $X \times Y$ are exactly the optimal solutions to (LP) augmented by optimal solutions to the dual problem.

Assume that

(a) It is relatively easy to solve Linear Programming problems of the form

$$\min_{x_1} \left\{ c^T x_1 : A^{11} x_1 \le b, \|x_1\|_\infty \le R \right\}.$$

More precisely, we assume that for every fixed x_2, y_2, it is easy to solve the *induced saddle point problem*

$$\min_{x_1 \in X_1} \max_{y_1 \in Y_1} L(x_1, x_2; y_1, y_2) \tag{11}$$

where Y_1 is a simple bounded part of Y_1^∞, specifically, a box of the form $\{y_1 \ge 0, \|y_1\|_\infty \le L\}$.

> The simplest (although by far not the only) case where (a) does take place is when A^{11} is a block-diagonal matrix comprised of a large number N of relatively low dimensional blocks (For example, these blocks can describe local technological relations at N branches of a large corporation). In this situation, the induced saddle point problem, due to the box structure of X_1 and Y_1, decomposes into N independent of each other low dimensional (and thus easy to solve) bilinear saddle point problems on products of two boxes.

(b) The number of $m_2 = \dim b_2$ of *linking constraints* is $\ll$ the total number of constraints m in (LP), and the number $n_2 = \dim y_2$ of *linking variables* is $\ll$ the total number of variables n in (LP).

> The above "corporation – branches" example usually meets the assumption (b). Indeed, in the situation considered in this example, the linking constraints usually correspond to bounds on main resources (capital, well-trained human resources, etc.) which are distributed between branches by the central management, while linking variables represent "strategic decisions" (directions of development, advertisement policies, etc.) made at the level of this management. Usually, the number of these resources/decisions is much smaller that the total, over all branches, number of branch-specific constraints/decisions.

The question (by far not new) is, how can we exploit the specific structure of (LP) in order to accelerate its solution.

There are two well studied "extreme cases" of the outlined problem – one without linking constraints, and another one without linking variables.

Case I: No linking constraints ($m_2 = 0$). In this case, the most popular way to utilize the specific structure of the problem is to use *Benders decomposition* (see, e.g., [5]). To the best of our knowledge, this decomposition, while often being extremely efficient in practice, in general is *not* supported by theoretical complexity bounds.

Case II: No linking variables ($n_2 = 0$). One way to exploit this structure is to use *Dantzig-Wolfe decomposition* (see, e.g., [5]), which is a specific implementation of the Revised Primal Simplex method. Another well-known decomposition scheme, much more flexible in the sense that it does *not* rely upon particular solution algorithm, is *Lagrangian decomposition*, where one dualizes the linking constraints, thus arriving at the *partial dual* of (LP), specifically, the problem

$$\mathrm{Opt} = \max_{y_2 \geq 0} \psi(y_2) := \min_{x_1} \left\{ [c_1 + [A^{21}]^T y_2]^T x_1 : \|x_1\|_\infty \leq R, A^{11} x_1 \leq b_1 \right\}. \tag{12}$$

Assuming that we can point out a "large enough" L, specifically, such that

$$\max_{y_2 \geq 0} \psi(y_2) = \max_{y_2 \geq 0, \|y_2\|_1 \leq L} \psi(y_2)$$

problem (12) becomes the Convex Minimization problem

$$-\mathrm{Opt} = \min_{y_2 \in Y^L} [-\psi(y_2)], \quad Y^L = \{y_2 \geq 0, \sum_i (y_2)_i \leq L\} \tag{13}$$

on a simple solid. Now, assumption (a) says that the first order information on ψ is not too costly. Specifically, given y_2, the optimization problem specifying $\phi(y_2)$ according to (12) is a problem of the form considered in (a) and thus is relatively easy to solve; after the optimal solution $x_1(y_2)$ to the latter problem is found, the first order information on ψ is readily given by the relations

$$\psi(y_2) = [c_1 + [A^{21}]^T y_2]^T x_1(y_2), \ \psi'(y_2) = A^{21} x_1(y_2).$$

Thus, the Convex Minimization problem (13) is well suited for solving by black box oriented methods. Of course, methods of this type are much slower than those oriented at well structured problems of comparable sizes, e.g., LPs. The point, however, is that according to (b), the design dimension of (13), that is, m_2, is much smaller that the sizes of the original LP, so that it well may happen (and indeed happens in numerous applications) that a "slow by itself" black-box-oriented method, like NERML or the Ellipsoid algorithm, as applied

to a *relatively low dimensional* problem (13) by far outperforms fast state-of-the-art LP solvers as applied to the *large scale* original problem (LP).

In order for the outlined Lagrange decomposition scheme to work, one should resolve a nontrivial question of how to pass from a good approximate solution to (13) to equally good approximate solution to the problem of actual interest (LP). There are various ways to resolve this challenging issue (e.g., to pass from the usual Lagrange function underlying (13) to an augmented Lagrangian). We demonstrate in section 5.2.2 that the machinery of accuracy certificates allows to resolve this issue in a pretty general and quite attractive, at least from the theoretical viewpoint, manner. The corresponding result is as follows:

Proposition 5.2.1 *Assume that $n_2 = 0$ and that* (13) *is solved by an algorithm with accuracy certificates. By construction of the First Order oracle for the objective of this problem, at every productive step t of this algorithm, the search point being $y_2^t \in Y$, we have at our disposal a point*

$$x_1^t = x_1(y_2^t) \in \operatorname*{Argmax}_{x_1} \left\{-[c_1 + [A^{21}]^T y_2^t]^T x_1 : A^{11}x_1 \le b_1, \|x_1\|_\infty \le R\right\}.$$

Now let τ be a step such that the accuracy certificate ξ^τ associated with the corresponding execution protocol $P_\tau = \left\{\{y_2^t, e_t\}_{t=1}^\tau, I_\tau, J_\tau\right)$ is well defined. Setting

$$\widehat{x}_1^\tau = \sum_{t \in I_\tau} \xi_t^\tau x_1^t,$$

we get an approximate solution to the problem of interest (LP) *such that*

$$\begin{aligned} &A^{11}\widehat{x}_1^\tau \le b_1 \ \&\ \|x_1^\tau\|_\infty \le R, \\ &A^{21}\widehat{x}_1^\tau \le b_2 + L^{-1}[\mathrm{Opt} + \varepsilon_{\mathrm{cert}}(\xi^\tau|P_\tau, Y^L) + R\|c_1\|_1]\mathbf{1} \\ &c_1^T\widehat{x}_1^\tau \le \mathrm{Opt} + \varepsilon_{\mathrm{cert}}(\xi^\tau|P_\tau, Y), \end{aligned} \tag{14}$$

where $\mathbf{1}$ is the all-ones vector of dimension m_2. In other words, the approximate solution $\widehat{x}_1^\tau$ exactly satisfies all but the linking constraints of the problem of interest, violates every one of the linking constraints by at most $L^{-1}[\mathrm{Opt} + \varepsilon_{\mathrm{cert}}(\xi^\tau|P_\tau, \mathbf{B}_L) + R\|c_1\|_1]$ – the quantity which is small when L is large enough and $\varepsilon_{\mathrm{cert}}(\xi^\tau|P_\tau, Y^L)$ is small enough, and is non-optimal in terms of the objective by at most $\varepsilon_{\mathrm{cert}}(\xi^\tau|P_\tau, Y^L)$.

In addition, let $\widetilde{y} = [\widetilde{y}_1; \widetilde{y}_2; \widetilde{y}_+; \widetilde{y}_-] \ge 0$ be the vector of optimal Lagrange multipliers for (LP)*, so that*

$$c_1^T x_1 + \widetilde{y}_1^T[A^{11}x_1 - b_1] + \widetilde{y}_2^T[A^{21}x_1 - b_2] + \widetilde{y}_+^T[x_1 - R\mathbf{1}] + \widetilde{y}_-^T[-x_1 - R\mathbf{1}] \equiv \mathrm{Opt}\,\forall x_1.$$

When $\ell := L - \sum_{i=1}^{m_2}[\widetilde{y}_2]_i > 0$, then, in addition to the second relation in (14)*, we have*

$$A^{21}\widehat{x}_1^\tau \le b_2 + \ell^{-1}\varepsilon_{\mathrm{cert}}(\xi^\tau|P_\tau, Y^L)\mathbf{1}$$

In particular, with properly chosen L one has $\ell \geq 1$, so that both infeasibility of $\widehat{x}_1^\tau$ w.r.t. the linking constraints and nonoptimality of this solution in terms of the objective are bounded from above by $\varepsilon_{\text{cert}}(\xi^\tau | P_\tau, Y^L)$.

General case. Our main emphasis in chapter 5 is on the *general case* where both linking variables and linking constraints are present. To the best of our knowledge, the only well known decomposition scheme proposed for this case is "cross decomposition" originating from T.J. Van Roy [42, 43], see also [45, 13, 14, 15, 16, 36] and references therein. In this scheme, essentially, one alternates iteratively between the Benders and the Lagrange decompositions. To the best of our understanding, no complexity results for this scheme are known (that is natural: the primary motivation and application of cross decomposition is the Mixed Integer version of (LP) where the linking variables x_2 are subject to additional constraints of integrality, the situation where no good complexity bounds could be expected). When completing our research, we became aware of a single paper [11] where, in hindsight, we recognized the accuracy-certificate-based approach similar to the one we developed in the Thesis. Note, however, that [11] is restricted to the case of problem (LP) and the full memory Bundle-Level algorithm as a working horse, while the approach we are about to present is much more general.

Our strategy is as follows.

1. We assume that our a priori understanding of (LP) allows to point out a finite upper bound S on the dual variables such that solving (LP) reduces (exactly or within accuracy sufficient for our goals) to solving the bilinear saddle point problem

$$\text{SadVal} = \min_{(x_1,x_2)\in X_1\times X_2} \max_{(y_1,y_2)\in Y_1\times Y_2} L(x_1,x_2;y_1,y_2), \tag{15}$$

   where $L(\cdot)$, X_1, X_2 are as in (10), and $Y_1 = \{y_1 : 0 \leq y_1 \leq L\mathbf{1}\}$, $Y_2 = \{y_2 : 0 \leq y_2 \leq L\mathbf{1}\}$ are *bounded* "approximations" of Y_1^∞, Y_2^∞.
   By assumption (a), given $x_2 \in X_2, y_2 \in Y_2$, it is easy to solve the saddle point problem

$$\widehat{L}(x_2,y_2) = \min_{x_1\in X_1} \max_{y_1\in Y_1} L(x_1,x_2;y_1,y_2); \tag{16}$$

   it is immediately seen that $\widehat{L}$ is convex-concave and continuous, and that first order information on $\widehat{L}$ is readily given by (any) solution $x_1 = x_1(x_2,y_2)$, $y_1 = y_1(x_2,y_2)$ of the right hand side saddle point problem. We now can form the *induced* saddle point problem

$$\text{SadVal} = \min_{x_2\in X_2} \max_{y_2\in Y_2} \widehat{L}(x_2,y_2); \tag{17}$$

   as we have already mentioned, we have at our disposal First Order oracle associated with this problem, and thus can solve it by a black-box-oriented saddle point algorithm.

By assumption (b), the dimension of this problem is much less than the one of the problem of interest (15), and under favorable circumstances it may happen that the total computational effort of solving (17) (including the expenses to generate the first order information on $\widehat{L}$) in this fashion is much smaller as compared to the effort of solving (LP) by a general purpose LP solver. Whenever this is the case, *the only issue we should resolve in order to use the outlined approach, is how to convert a good approximate solution to the induced saddle point problem* (17) *into an equally good approximate solution to the saddle point of interest* (15). Our major effort in chapter 5 is to resolve this issue by using the machinery of accuracy certificates.

Note that the outlined strategy is a natural extension of the one used in Lagrange decomposition, with the only (although essential) difference that in the latter, the induced saddle point problem is just a Convex Minimization program.

In fact, in the main body of chapter 5 we consider a situation more general than the one outlined above. Specifically, we assume that the problem of interest is a convex-concave saddle point problem of the form (15), with arbitrary solids X_1, X_2, Y_1, Y_2 and a not necessarily bilinear cost function $L(\cdot)$ which should be convex-concave and continuous and, in addition, should satisfy mild regularity assumption which for sure is satisfies when L is continuously differentiable on $X_1 \times X_2 \times Y_1 \times Y_2$. Same as above, we assume that whenever $x_2 \in X_2$, $y_2 \in Y_2$, it is easy to solve the saddle point problem in the right hand side of (16). Under the regularity assumption we have mentioned, a solution $x_1(x_2, y_2), y_1(x_2, y_2)$ straightforwardly induces first order information for the function $\widehat{L}(x_2, y_2)$ which automatically is convex-concave and continuous on $X_2 \times Y_2$. As a result, the *induced saddle point problem* (17) is well suited for solving by black box oriented methods. Our main result in chapter 5 is how to resolve the issue of converting a good approximate solution to the induced problem into an equally good solution to the problem of interest. Here is the result (cf. Theorem 5.3.1):

Theorem *Let the induced saddle point* (17) *be solved by an algorithm with accuracy certificates. Assume that after a number τ of steps, we have at our disposal execution protocol $P_\tau = \big(\{(x_2^t, y_2^t), e_t\}_{t=1}^\tau, I_\tau, J_\tau\big)$ along with an accuracy certificate ξ^τ. For $t \in I_\tau$, as a byproduct of getting first order information on $\widehat{L}$ at the productive search points (x_2^t, y_2^t), we have at our disposal vectors $x_1^t \in X_1, y_1^t \in Y_1$ such that (x_2^t, y_1^t) is a solution to the right hand side saddle point problem in* (16) *corresponding to $(x_2, y_2) = (x_2^t, y_2^t)$. Now let*

$$(x_1^\tau, x_2^\tau, y_1^\tau, y_2^\tau) = \sum_{t \in I_\tau} \xi_t^\tau (x_1^t, x_2^t, y_1^t, y_2^t).$$

Then $(x_1^\tau, x_2^\tau, y_1^\tau, y_2^\tau)$ is a feasible approximate solution to the problem of interest (15)*, and*

$$\varepsilon_{\text{sad}}(x_1^\tau, x_2^\tau, y_1^\tau, y_2^\tau) \le \varepsilon_{\text{cert}}(\xi^\tau | P_\tau, \mathbf{B})$$

for every solid $\mathbf{B}$ containing $X_2 \times Y_2$. In other words, a good approximate solution to the induced saddle point problem (one induced by an accuracy certificate with resolution $\le \epsilon$)

can be easily converted into equally good (with $\varepsilon_{\text{sad}}(\cdot) \leq \epsilon$) approximate solution to the problem of interest.

The above outline makes it clear that our research is theoretical in its nature. We, however, do believe that some of our findings, especially those related to decomposition, possess certain practical potential. The specific time limitations imposed on our research did not allow for extensive numerical experimentation aimed at investigations this potential; we consider such an investigation as a subject of future research.

CHAPTER II

BACKGROUND: ACCURACY CERTIFICATES FOR PROBLEMS WITH CONVEX STRUCTURE

The notion of a *certificate* — a well-structured entity which provides an easy-to-verify proof of a certain statement — is widely used in Computer Science and Optimization. For example,

- A simple certificate of the solvability of a system of linear inequalities $Ax \geq b$ is a solution $\bar{x}$ to this system, while a simple certificate of *in*solvability of the system is a solution $\bar{y}$ to the alternative system $A^T y = 0$, $y \geq 0$, $y^T b > 0$ (General Theorem on Alternative).

- Given a Linear Programming problem $\mathrm{Opt} = \min_x\{c^T x : Ax \geq b\}$ and a tolerance $\epsilon \geq 0$, a simple certificate of the fact that a candidate solution $\bar{x}$ is feasible and ϵ-optimal (i.e., $c^T\bar{x} \leq \mathrm{Opt} + \epsilon$) is given by a feasible solution to the primal problem $\bar{x}$, and a feasible solution $\bar{y}$ to the dual problem $\max_y\{b^T y : A^T y = c, y \geq 0\}$ such that the corresponding duality gap $c^T\bar{x} - b^T\bar{y}$ is $\leq \epsilon$.

Note that both these constructions can be naturally extended from Linear Programming problems and systems of linear inequalities, to conic programming problems $\min_x\{c^T x : Ax - b \in K\}$, and conic inequalities $Ax - b \in K$, where $K \subset \mathbf{R}^n$ is a closed convex pointed cone with a nonempty interior, see, e.g., [2]. [1]

As a matter of fact, certificates for feasibility and ϵ-optimality of candidate solutions to Convex Programming problems were until recently known solely for "well-structured" problems, like those in the examples above. Surprisingly, till very recently no notion of a certificate for *black box represented* convex problems — those where the objective and the constraints are represented by oracles capable of computing their values and derivatives – was known. This is quite unfortunate considering the fact that black box representation of a problem is the standard "computational environment" for general-type Convex Programming algorithms. A general concept of an "accuracy certificate" for a black box represented convex program

[1] It should be mentioned that in the case of an LP problem/a system of linear inequalities the existence of the outlined certificate is not only a sufficient, but also a *necessary* condition for the validity of the fact we intend to certify. In the case of Conic Programming problems/conic inequalities, the existence of a certificate is merely sufficient; to make it necessary, mild regularity assumptions should be made, see [2].

was introduced only recently in [27].

In this chapter we will introduce four problems with convex structure along with their associated accuracy measures, we will then introduce the idea of a ξ certificate first detailed in [27], we will show that this ξ certificate is sufficient to prove that we have achieved an accuracy of ϵ for any $\epsilon > 0$. Further we will briefly exam the necessity of these ξ certificates.

2.1 Problems with Convex Structure and Accuracy Measures

In this section, we present four generic "problems with convex structure" — convex minimization, convex-concave saddle point problem, convex Nash equilibrium problem, and variational inequality with monotone operator — which will be of primary interest in the sequel. Each problem will be associated with a natural accuracy measure and a specific operator (a vector field on the problem's domain). The "common ground" for these seemingly highly diverse problems is that up to certain point they admit a unified treatment that covers the most basic first-order solution algorithms (Cutting Plane, Subgradient Descent, and Bundle methods), as well as the issue of accuracy certificates, the latter being our primary focus for this study.

2.1.1 Convex Minimization

A convex minimization problem is:

$$\mathrm{Opt} = \min_{x \in X} f(x), \tag{18}$$

where X is a *solid* (a compact convex set with a nonempty interior) in $\mathbf{R}^n$, and f is a convex continuous function on X, s.t. $\mathrm{int}X \subset \mathrm{Dom}\, f$.

We focus on *black box represented* convex minimization problems, meaning that:

1. X is given by a *Separation oracle* — a routine which, given on input a vector $x \in \mathbf{R}^n$, reports one of three possible situations: $x \in \mathrm{int}X$, $x \in \partial X$, or $x \notin X$. If $x \notin \mathrm{int}X$, the routine returns a *separator* — a nonzero linear form $\langle e, \cdot \rangle$ on $\mathbf{R}^n$ such that

$$\langle e, x \rangle \geq \max_{y \in X} \langle e, y \rangle.$$

Example 2.1.1. A standard example of a Separation oracle corresponds to the case when X is given by a finite set of strictly feasible convex constraints: $X = \mathrm{cl}\,\{x \in \mathbf{R}^n; g_i(x) < 0,\, 1 \leq i \leq m\}$, where the convex functions $g_i : \mathbf{R}^n \to \mathbf{R}$ are computable along with their subgradients. The associated Separation oracle is a routine which,

given on input a query point x, computes the values $g_i(x)$ and subgradients $g_i'(x)$ of all the constraints at x and checks whether $g_i(x) < 0$ for all i. If it is the case, the oracle reports that $x \in \mathrm{int}X$; otherwise, it reports that $x \notin \mathrm{int}X$, finds $i = i_x$ such that $g_{i_x}(x) \geq 0$, and returns the separator $e = g_{i_x}'(x)$.

2. f is represented by a *First Order Oracle* — a routine which, given on input a vector $x \in \mathrm{int}X$, returns the value $f(x)$, as well as a subgradient $f'(x) \in \partial_X f(x)$ of f at x, so that

$$\forall y \in X : f(y) \geq f(x) + \langle f'(x), y - x \rangle.$$

Sometimes we will assume that f possesses subgradients everywhere on X and that the first Order oracle is capable of computing $f(x)$ and $f'(x) \in \partial_X f(x)$ at every point $x \in X$; in this case, we say that the First Order oracle serves the entire X, while in the former case we say that it serves int$X$ only.

Accuracy Measure: We quantify the (in)accuracy of a candidate solution $x \in X$ for a convex minimization problem (18) via the *accuracy measure*

$$\varepsilon_{\mathrm{opt}}(x) := f(x) - \mathrm{Opt} \tag{19}$$

Comment. While the accuracy measure we have introduced (usually called the *residual in terms of the objective*) is quite natural, it is not the only natural accuracy measure in convex minimization. Perhaps a more intuitive accuracy measure would be the "accuracy in argument" — the distance $\mathrm{dist}(x, X_*)$ from a candidate solution $x \in X$ to the set X_* of optimal solutions to (18):

$$\mathrm{dist}(x, X_*) = \min_{x_* \in X_*} \|x - x_*\|.$$

The rationale behind our choice of accuracy measure is completely pragmatic: *namely, in general-type convex minimization $\varepsilon_{\mathrm{opt}}$ is the only known accuracy measure which gives rise to affirmative complexity results.* For example, assume that in addition to convexity, f is bounded on X; then it is known ([46]; also Proposition 2.2.1 below) that for every $\epsilon > 0$ one can find an approximate solution $x_\epsilon \in X$ to (18) such that $\varepsilon_{\mathrm{opt}}(x) \leq \epsilon$ in no more than

$$N = O(1)n^2 \ln\left(r(X)\frac{\max_X f - \min_X f}{\epsilon}\right)$$

calls to the Separation and First Order oracles, with $O(1)n^2$ additional arithmetic operations per oracle call. (Here $r(X) < \infty$ depends solely on X, and all $O(1)$'s are absolute constants.) In contrast to this, the complexity of approximating the optimal solution in terms of the argument can be disastrous.

Specifically, consider the case of minimizing a C^∞ convex function f of two variables over the unit disk $X \subset \mathbf{R}^2$, where we know in advance that, for a given k, the k^{th}-order

derivatives of f are bounded by 1 within a disk which is twice as large. Let us also assume that instead of a First Order oracle, we have at our disposal an oracle which reports *all* partial derivatives of f at a query point. Let us now impose on an algorithm, based on this powerful oracle, the requirement that for any given objective f with the outlined properties, the algorithm terminates after a finite number of steps (this number can depend on f) and outputs a point $\bar{x} = \bar{x}_f$ such that $\mathrm{dist}(\bar{x}_f, \mathrm{Argmin}_X f) \leq \frac{1}{4}$. The fact is that *no such algorithm exists!* [27]. It should be stressed that while there exist numerous — and simple — algorithms which generate sequences of iterates converging to $\mathrm{Argmin}_X f$; *the convergence of these algorithms is completely "unobservable" — there is no on-line test which allows one to conclude that the distance from an iterate to the optimal set is already* $\leq \frac{1}{4}$.

Of course, when imposing on f additional assumptions (e.g., strong convexity), then measuring the inaccuracy in the argument becomes a tractable task. Note, however, that as far as we know all assumptions of this type allow one to bound the accuracy in argument via $\varepsilon_{\mathrm{opt}}$, such that we may focus solely on our chosen accuracy measure.

We will follow this "pragmatic" approach (namely "a good (in)accuracy measure is the one which we can make small in an efficient fashion") when introducing accuracy measures for other problems with convex structure.

Monotone Operator: The crucial entity associated with the convex minimization problem (18) is the operator $F(x) = f'(x)$, the domain $\mathrm{Dom}\, F$ of this operator being $\mathrm{int}X$ or X, depending on whether the First Order oracle serves $\mathrm{int}X$ or X. Note that $\mathrm{Dom}\, F$ is always convex, and F is *monotone* on its domain:

$$\langle F(x) - F(y), x - y\rangle \geq 0 \ \forall x, y \in \mathrm{Dom}\, F. \tag{20}$$

For the sake of completeness, here is the verification of monotonicity: by construction combined with the subgradient inequality, for $x, y \in \mathrm{Dom}\, F$ we have $\langle F(x), y - x\rangle \leq f(y) - f(x)$ and $\langle F(y), x - y\rangle \leq f(x) - f(y)$; summing up these relations, we arrive at (20).

2.1.2 Convex-Concave Saddle Point Problem

A convex-concave saddle point problem is

$$\text{Find } x^* = (x_1^*, x_2^*) \in X = X_1 \times X_2 : f(x_1, x_2^*) \geq f(x_1^*, x_2^*) \geq f(x_1^*, x_2) \ \forall (x_1, x_2) \in X, \tag{21}$$

where $X_i \subset \mathbf{R}^{n_i}$, $i = 1, 2$, are convex solids and f is a continuous function on $X = X_1 \times X_2$ which is convex in $x_1 \in X_1$, and concave in $x_2 \in X_2$.

We focus on *black box represented* saddle points problems, meaning that X_1, and X_2 are given by Separation oracles, and f is given by a First Order oracle capable of computing the value $f(x)$ along with a subgradient $f'_{x_1}(x)$ of f with respect to x_1 and a super-gradient $f'_{x_2}(x)$ of f with respect to x_2 at every point $x \in \mathrm{int}X$ ("the oracle serves int$X$") or at every point $x \in X$ ("the oracle serves the entire X"); of course, in the latter case we assume that the required sub- and supergradients exist.

Accuracy Measure: A convex-concave saddle point problem (21) gives rise to a primal-dual pair of convex optimization problems

$$\begin{array}{rcl} \mathrm{Opt}(P) & = & \min\limits_{x_1 \in X_1} \overline{f}(x_1), \ \text{where } \overline{f}(x_1) = \max\limits_{x_2 \in X_2} f(x_1, x_2), \\ \mathrm{Opt}(D) & = & \max\limits_{x_2 \in X_2} \underline{f}(x_2), \ \text{where } \underline{f}(x_2) = \min\limits_{x_1 \in X_1} f(x_1, x_2) \end{array} \tag{22}$$

with equal optimal values. The solutions to (21) are exactly the pairs (x_1^*, x_2^*) comprised of optimal solutions to the respective optimization problems. This leads us to equip (21) with a natural accuracy measure

$$\varepsilon_{\mathrm{sad}}(x) := \overline{f}(x_1) - \underline{f}(x_2) = [\overline{f}(x_1) - \mathrm{Opt}(P)] + [\mathrm{Opt}(D) - \underline{f}(x_2)]. \tag{23}$$

We note that $\varepsilon_{\mathrm{sad}}(x_1, x_2)$ is the sum of the nonoptimalities of x_1 as a solution to the primal problem and of x_2 as a solution to the dual problem. As such, the duality gap is always nonnegative and equals zero iff x_1 is an optimal solution to (P) and x_2 is an optimal solution to (D), i.e., iff $x = [x_1; x_2]$ is a solution to (21).

Monotone Operator: We associate with (21) the operator $F(x_1, x_2) = [f'_{x_1}(x); -f'_{x_2}(x)]$ with the domain which is either int$X$ or X, depending on whether the First Order oracle serves int$X$ or X. It is well known that this operator is monotone on its domain.

For the sake of completeness, here is the verification of the monotonicity: by construction combined with the definitions of sub/supergradients, for $x = [x_1; x_2], y = [y_1; y_2] \in \mathrm{Dom}\, F$ we have

$$\begin{array}{l} \langle F(x), y - x\rangle = \langle f'_1(x), y_1 - x_1\rangle - \langle f'_2(x), y_2 - x_2\rangle \\ \le [f(y_1, x_2) - f(x)] + [f(x) - f(x_1, y_2)] = f(y_1, x_2) - f(x_1, y_2), \end{array}$$

and similarly $\langle F(y), x - y\rangle \le f(x_1, y_2) - f(y_1, x_2)$. Summing up the resulting inequalities, we arrive at (20).

2.1.3 Convex Nash Equilibrium Problem

In a Nash equilibrium problem, one is given m sets $X_i \subset \mathbf{R}^{n_i}$ along with m functions $f_i(x)$ defined on $X = X_1 \times ... \times X_m$, and seeks a *Nash equilibrium*:

$$\begin{array}{l} \text{a point } x^* \in X \text{ such that for every } i \leq m \ x_i^* \text{ is a minimizer of the function} \\ f_i(x_1^*, ..., x_{i-1}^*, \xi_i, x_{i+1}^*, ..., x_m^*) \text{ over } \xi_i \in X_i. \end{array} \tag{24}$$

The "story" is that $x_i \in X_i$ represents the choice of the i-th player, and $f_i(x)$ represents the loss incurred by this player when the choices of all the players constitutes a point $x = [x_1; ...; x_m]$. Viewed in this fashion, an equilibrium is a collection of players choices such that no player can reduce his loss by his unilateral actions.

Following [27], we call a Nash equilibrium problem *convex*, if the X_i are solids, and the f_i are continuous functions on X such that f_i is convex in x_i and jointly concave in the remaining components of x, and the function $f(x) = \sum_i f_i(x)$ is convex. From now on, when speaking about Nash equilibrium problems we will always assume that the problem is convex.

We focus on *black box represented convex* Nash equilibrium problems, where X_i are represented by Separation oracles, and f_i, $i = 1, ..., m$ are represented by First Order oracles capable of computing $f_i(x)$ and subgradients $f_i'(x)$ of f_i w.r.t. x_i, $i = 1, ..., m$. As is now standard this holds for every $x \in \text{int}X$ ("first Order oracle serves int$X$"), or for every $x \in X$ ("the oracle serves the entire X").

Accuracy Measure: The *accuracy measure* for a convex Nash equilibrium problem is defined as

$$\varepsilon_{\text{N}}(x) := \sum_{i=1}^{m} \left[f_i(x) - \min_{\xi_i \in X_i} f_i(x_1, ..., x_{i-1}, \xi_i, x_{i+1}, ..., x_m) \right]$$

which can be interpreted as the sum, over the m players, of the incentive for the $i-th$ player to deviate from his choice x_i given that all the remaining players stick to their choices.

Note a convex minimization problem (18), with an objective f which is continuous on X, can be considered as a convex Nash equilibrium problem with $m = 1$ player (set $X_1 = X$, $f_1(x_1) = f(x)$), and in this case we clearly have $\varepsilon_{\text{opt}}(x) \equiv \varepsilon_{\text{N}}(x)$. Similarly, a convex-concave saddle point problem can be thought of as a convex Nash equilibrium problem with $m = 2$ players, such that $f_1(x) = f(x)$, $f_2(x) = -f(x)$, and thus $f_1(x) + f_2(x) \equiv 0$. In this case the Nash accuracy measure recovers $\varepsilon_{\text{sad}}(\cdot)$, and here again the accuracy measure ε_{N} reduces to

the saddle point accuracy measure $\varepsilon_{\rm sad}$, since when $f_1(x) = f(x)$, $f_2(x) = -f(x)$, we have

$$\begin{aligned}
&\varepsilon_{\rm N}(x_1, x_2) = [f_1(x_1, x_2) - \min x_1' \in X_1 f_1(x_1', x_2)] + [f_2(x_1, x_2) - \min x_2' \in X_2 f_2(x_1, x_2')] \\
&= [f(x_1, x_2) - \min x_1' \in X_1 f(x_1', x_2)] + [-f(x_1, x_2) + \max x_2' \in X_2 f(x_1, x_2')] \\
&= [f(x) - \underline{f}(x_2)] + [\overline{f}(x_1) - f(x)] = \varepsilon_{\rm sad}(x).
\end{aligned}$$

Monotone Operator: A convex Nash equilibrium problem can be associated with the operator $F(x) = [f_1'(x); ...; f_m'(x)]$ with a domain which is either int$X$, or X, depending on whether the First Order oracle serves int$X$ or the entire X. In both cases, F is monotone on its domain.

For the sake of completeness, here is the verification of the monotonicity as taken from [27]. Let $x, y \in \mathrm{Dom}\, F$, and let us prove that $\langle F(x) - F(y), x - y\rangle \geq 0$, or, equivalently, that $\langle F(\bar{x} + \Delta) - F(\bar{x} - \Delta), \Delta\rangle \geq 0$, where $\bar{x} = \frac{1}{2}(x + y)$ and $\Delta = \frac{1}{2}(x - y)$. For $z = [z_1; ...; z_m] \in X$ let us set $z^i = [z_1; ...; z_{i-1}; z_{i+1}; ...; z_m]$, so that $f_i(z) = f_i(z^i, z_i)$. We have

$$\begin{aligned}
&\langle F(\bar{x} + \Delta) - F(\bar{x} - \Delta), \Delta\rangle = \textstyle\sum_i \big[\underbrace{\langle f_i'(\bar{x} + \Delta), \Delta_i\rangle}_{\geq_{(a)} f_i(\bar{x}+\Delta) - f_i(\bar{x}^i + \Delta^i, \bar{x}_i)} + \underbrace{\langle f_i'(\bar{x} - \Delta), -\Delta_i\rangle}_{\geq_{(a)} f_i(\bar{x}-\Delta) - f_i(\bar{x}^i - \Delta^i, \bar{x}_i)} \big] \\
&\geq \textstyle\sum_i \big[f_i(\bar{x} + \Delta) - f_i(\bar{x}^i + \Delta^i, \bar{x}_i) + f_i(\bar{x} - \Delta) - f_i(\bar{x}^i - \Delta^i, \bar{x}_i)\big] \\
&= \Phi(\bar{x} + \Delta) + \Phi(\bar{x} - \Delta) - \textstyle\sum_i \big[\underbrace{f_i(\bar{x}^i + \Delta^i, \bar{x}_i) + f_i(\bar{x}^i - \Delta^i, \bar{x}_i)}_{\leq_{(b)} 2f_i(\bar{x})} \big] \\
&\geq \Phi(\bar{x} + \Delta) + \Phi(\bar{x} - \Delta) - 2\Phi(\bar{x}) \underset{(c)}{\geq} 0
\end{aligned}$$

where (a), (b) are due to the fact that $f_i(z)$ are convex in z_i and concave in z^i, and (c) is due to the convexity of $\Phi = \sum_i f_i$, along with definitions of $\bar{x}$ and Δ.

2.1.4 Variational Inequalities with Monotone Operators

Let $X \subset \mathbf{R}^n$ be a solid, and $F(x) : \mathrm{Dom}\, F \to \mathbf{R}^n$ be an operator with convex domain $\mathrm{Dom}\, F$ such that int$X \subset \mathrm{Dom}\, F \subset X$[2] and F is monotone on its domain. The pair X, F gives rise to the *variational inequality problem*

$$\text{Find } x_* \in X : \langle F(y), y - x_*\rangle \geq 0 \, \forall y \in \mathrm{Dom}\, F \tag{25}$$

[2]Note that while variational inequalities are quite general, our main interest is to set up the operator F as the monotone operator from one of our three prior functional problems. Hence it makes perfect sense to consider int$X \subset \mathrm{Dom}\, F \subset X$, since we are guaranteed that sub/super gradients are defined on at least all of int$X$, and *may* be defined over the entire X.

In literature, the just defined solutions to (25) are called *weak solutions*, as opposed to *strong solutions* — points $x_* \in \mathrm{Dom}\, F$ such that $\langle F(x_*), y - x_* \rangle \geq 0$ for all $y \in X$. Note that for a monotone F strong solutions are always weak solutions.

Indeed let us assume that we have a strong solution x^*, then $x^* \in \mathrm{Dom}\,(\Phi)$ is such that $\langle \Phi(x^*), x - x^* \rangle \geq 0 \ \forall x \in X$. This certainly implies that $\langle \Phi(x^*), x - x^* \rangle \geq 0 \ \forall x \in X \cap \mathrm{Dom}\,(\Phi)$, further monotonicity implies $\langle \Phi(x), x - x^* \rangle - \langle \Phi(x^*), x - x^* \rangle \geq 0$. Now since $\langle \Phi(x^*), x - x^* \rangle \geq 0$ we must have $\langle \Phi(x), x - x^* \rangle \geq 0$, and thus x is also a weak solution

Also have that under mild regularity assumptions (F is continuous on $X = \mathrm{Dom}\, F$) the inverse is also true (weak solutions are also strong solutions). The advantage of weak solutions is that they always exist.

For the sake of completeness, here is the proof of the aforementioned well known result.

Proposition 2.1.1. [Existence of weak solutions to VIs with monotone operators] *Under our assumptions on X and F, the variational inequality* (25) *always has a (weak) solution.*

Proof. Setting $\Pi_y = \{x \in X : \langle F(y), y - x \rangle \geq 0\}$, for every $y \in \mathrm{Dom}\, F$, we get a family of closed subsets of the compact set X; what we want to prove, is that this whole family of sets has a point in common, and to this end, by the standard compactness argument it suffices to prove that every finite collection $\{\Pi_{y_i} : y_i \in \mathrm{Dom}\, F\}_{i=1}^N$ of these sets has a nonempty intersection. Let, on the contrary to what should be proved, $\{\Pi_{y_i} : y_i \in \mathrm{Dom}\, F\}_{i=1}^N$ be a collection with $\bigcap\limits_{i=1}^N \Pi_{y_i} = \emptyset$, or, which is the same, such that

$$\min_{1 \leq i \leq N} \langle F(y_i), y_i - x \rangle < 0 \ \forall x \in X.$$

Then, invoking the von Neumann Lemma, there exists a convex combination

$$\sum_{i=1}^N \lambda_i \langle F(y_i), y_i - x \rangle$$

of the affine functions $\langle F(y_i), y_i - x \rangle$ of x which is negative everywhere on X:

$$\sum_{i=1}^N \lambda_j \langle F(y_i), y_i - x \rangle < 0 \ \forall x \in X. \tag{26}$$

Let us set $\bar{x} = \sum_{i=1}^N \lambda_i y_i$; this point belongs to $\mathrm{Dom}\, F$, since the latter set is convex, and thus belongs to X. By monotonicity of F, we have for every i

$$\langle F(y_i), y_i - \bar{x} \rangle \geq \langle F(\bar{x}), y_i - \bar{x} \rangle.$$

Multiplying both sides of these inequalities by λ_i and summing up, we get

$$\sum_{i=1}^N \lambda_i \langle F(y_i), y_i - \bar{x} \rangle \geq \langle F(\bar{x}), \sum_{i=1}^N \lambda_i y_i - \bar{x} \rangle = 0,$$

which contradicts (26). We have arrived at a desired contradiction. ∎

We are interested in *black box represented* variational inequalities with monotone operators, meaning that X is given by a Separation oracle, and F is represented by a First Order oracle capable of computing $F(x)$ for $x \in \mathrm{Dom}\, F$. In what follows, we restrict ourselves to the situations where either $\mathrm{Dom}\, F = \mathrm{int}X$ ("the First Order oracle serves int$X$"), or $\mathrm{Dom}\, F = X$ ("the First Order oracle serves the entire X").

Accuracy Measure: The *accuracy measure* associated with (25) is

$$\varepsilon_{\mathrm{vi}}(x) = \sup_{y \in \mathrm{Dom}\, F} \langle F(y), x - y \rangle;$$

when $x \in X$. This measure (also called the *dual gap function* in the terminology of [9]) originates from [1] and has been used in many papers, in particular, in [25, 21, 27]. It is immediately seen that $\varepsilon_{\mathrm{vi}}(x)$ satisfies the natural requirements for an accuracy measure, specifically, it is nonnegative and equals zero if and only if x is a weak solution to (25). The latter fact is evident; while the nonnegativity of $\varepsilon_{\mathrm{vi}}(\cdot)$ is readily given by the fact that this function is convex (as the supremum of a family of affine functions of x) and nonnegative on $\mathrm{Dom}\, F \supset \mathrm{int}X$ (indeed, for $x \in \mathrm{Dom}\, F$ one has $\varepsilon_{\mathrm{vi}}(x) \geq \langle F(x,), x - x \rangle = 0$).

Note: We have associated with our three prior functional problems, namely, convex minimization problems, convex-concave saddle point problems, and convex Nash equilibrium problems, a monotone operator F with a convex domain which is in-between the domain X of the problem of interest and int$X$. It is well known that the solutions to these problems are exactly the weak solutions of the variational inequality associated with X and F.

Proposition 2.1.2. *Under assumptions of the respective sections 2.1.1, 2.1.2, 2.1.3, weak solutions to the variational inequalities with monotone operators associated with convex minimization, convex-concave saddle point and convex Nash equilibrium problems are exactly the solutions to the respective problems.*

Proof. Let us start with the convex minimization problem (18).

⇒ First, we observe that an optimal solution x^* to the associated variational inequality is a weak solution to this inequality. Indeed, by construction of the monotone operator F associated with (18), for every $y \in \mathrm{dom}F$ we have $F(y) \in \partial_X f(y)$, whence by the gradient inequality and the optimality of x^*, $\langle F(y), y - x^* \rangle \geq f(y) - f(x^*) \geq 0$ for all $y \in \mathrm{Dom}\, F$, (i.e., x^* is a weak solution).

$\Leftarrow$ Vice versa, let x^* be a weak solution to the variational inequality associated with (18), and let $z \in \mathrm{int}X$. Then for every $\alpha \in (0,1]$ we have $y_\alpha = (1-\alpha)x^* + \alpha z \in \mathrm{int}X \subset \mathrm{Dom}\, F$, whence by (25) $\langle F(y_\alpha), y_\alpha - x^*\rangle \geq 0$, meaning that the function $\phi(\alpha) = f(y_\alpha)$ is nonincreasing in $\alpha \in (0,1]$. Since f is lower semicontinuous, we conclude that $f(x^*) = \phi(0) \leq \phi(1) = f(z)$ for any $z \in \mathrm{int}X$. The bottom line is that $f(x^*) \leq \inf_{z\in \mathrm{int}X} f(z) = \inf\limits_{z\in X} f(z)$, where the concluding relation follows from the fact that f is convex on X and $\mathrm{int}X \neq \emptyset$. The proof in the case of convex minimization is completed.

As we have seen, the convex-concave saddle point problem is a particular case of the convex Nash equilibrium problem, so that all that remains is to prove the Proposition for the case of a convex Nash equilibrium problem.

$\Rightarrow$ Let x^* be a Nash equilibrium, and let us prove that x^* is a weak solution to the corresponding variational inequality. Let $y \in \mathrm{dom}F$. As we remember from section 2.1.3, $\mathrm{Dom}\, F$ is either $\mathrm{int}X$, or the entire X; in both cases, setting $\Delta = \frac{1}{2}(y - x^*)$ and $\bar{x} = \frac{1}{2}(y + x^*)$, we have $\bar{x} \in \mathrm{Dom}\, F$. We have [3]

$$
\begin{aligned}
\tfrac{1}{2}\langle F(y), y - x^*\rangle = \langle F(\bar{x}+\Delta), \Delta\rangle &= \textstyle\sum_i \underbrace{\langle f_i'(\bar{x}+\Delta), \Delta_i\rangle}_{\overset{\geq}{(a)}\, f_i(\bar{x}+\Delta) - f_i(\bar{x}^i + \Delta^i, \bar{x}_i)} \\
&\geq \textstyle\sum_i \big[f_i(\bar{x}+\Delta) - f_i(\bar{x}^i+\Delta^i, \bar{x}_i) + \underbrace{f_i(\bar{x}-\Delta) - f_i(\bar{x}^i - \Delta^i, \bar{x}_i)}_{= f_i([x^*]^i, x_i^*) - f_i([x^*]^i, \bar{x}_i)\, \overset{\leq}{(b)}\, 0} \big] \\
&= \Phi(\bar{x}+\Delta) + \Phi(\bar{x}-\Delta) - \textstyle\sum_i \big[\underbrace{f_i(\bar{x}^i+\Delta^i, \bar{x}_i) + f_i(\bar{x}^i - \Delta^i, \bar{x}_i)}_{\overset{\leq}{(c)}\, 2f_i(\bar{x})} \big] \\
&\geq \Phi(\bar{x}+\Delta) + \Phi(\bar{x}-\Delta) - 2\Phi(\bar{x}) \overset{\geq}{(d)} 0
\end{aligned}
$$

where (a) is due to convexity of $f_i(y^i, y_i)$ in y_i, (b) due to the fact that $f_i([x^*]^i, x_i)$ attains it minimum in $x_i \in X_i$ at the point x_i^*, (c) is due to the concavity of $f_i(x^i, \bar{x}_i)$ in x^i and (d) is due to the convexity of Φ. We see that $\langle F(y), y - x^*\rangle \geq 0$ for all $y \in \mathrm{Dom}\, F$, so that x^* is a weak solution to the variational inequality in question.

$\Leftarrow$ Now let x^* be a weak solution to the variational inequality associated with (24), and let us prove that x^* is a Nash equilibrium. Assume, on the contrary, that for some i the function $f_i([x^*]^i, x_i)$ does *not* attain its minimum over $x_i \in X_i$ at the point x_i^*;

[3] Recall our notation, for $z = [z_1; ...; z_m] \in X$ let us set $z^i = [z_1; ...; z_{i-1}; z_{i+1}; ...; z_m]$, so that $f_i(z) = f_i(z^i, z_i)$

w.l.o.g., let it be the case for $i = m$. Let v be a minimizer of the convex continuous function $f_m([x^*]^m, \cdot)$ on X_m; then the function $f(s) = f_m([x^*]^m, x_m^* + s[v - x_m^*])$ is nonincreasing in $s \in [0,1]$ and there exists $\bar{s} \in (0,1)$ such that $f(\bar{s}) - f(1) > 0$, or, equivalently, $f_m([x^*]^m, x_m^* + \bar{s}[v - x_m^*]) > f_m([x^*]^m, v)$. Since $f_m(x)$ is continuous in $x \in X$, we can find, $\epsilon > 0$, $\bar{v} \in \mathrm{int} X_m$ close enough to v, and a small enough convex neighborhood U (in $X^m = X_1 \times ... \times X_{m-1}$) of the point $[x^*]^m$ such that

$$\forall (u \in U) : f_m(u, x_m^* + \bar{s}[\bar{v} - x_m^*]) - f_m(u, \bar{v}) \geq \epsilon. \tag{27}$$

Let us choose somehow $\bar{u} \in U \cap \mathrm{int} X^m$ and let

$$x[\rho, \delta] = (\underbrace{[x^*]^m + \rho[\bar{u} - [x^*]^m]}_{\bar{u}_\rho}, \underbrace{x_m^* + \delta[\bar{v} - x_m^*]}_{v_\delta}),$$

so that $x[\rho, \delta] \in \mathrm{int} X$ for $0 < \rho, \delta \leq 1$. For $1 \leq i < m$ and $0 \leq \rho < 1$, $0 < \delta \leq 1$ we have

$$\begin{array}{l} \langle f_i'(x[\rho,\delta]), x_i[\rho,\delta] - x_i^* \rangle = \frac{\rho}{1-\rho} \langle f_i'(x[\rho,\delta]), \bar{u}_i - x_i[\rho,\delta] \rangle \\ \overset{\leq}{(a)} \frac{\rho}{1-\rho} [f_i(x^i[\rho,\delta], \bar{u}_i) - f_i(x[\rho,\delta])] \leq \frac{\rho}{1-\rho} M, \end{array} \tag{28}$$

where $M/2$ is an upper bound on $|f_j(x)|$ over $j = 1, ..., m$ and $x \in X$; here (a) is given by the convexity of f_i in $x_i \in X_i$. We further have $\bar{u}_\rho \in U$, whence, by (27),

$$\begin{array}{l} -\epsilon \geq f_m(\bar{u}_\rho, \bar{v}) - f_m(\bar{u}_\rho, x_m^* + \bar{s}[\bar{v} - x_m^*]) = f_m(x[\rho, 1]) - f_m(x[\rho, \bar{s}]) \\ = \int_{\bar{s}}^1 \langle f_m'(x[\rho, s]), \bar{v} - x_m^* \rangle ds. \end{array}$$

We see that there exists $\delta = \delta_\rho \in [\bar{s}, 1]$ such that $\langle f_m'(x^m[\rho, \delta_\rho], \bar{v} - x_m^* \rangle \leq -\epsilon$, or, which is the same,

$$\langle f_m'(x[\rho, \delta_\rho]), x_m[\rho, \delta_\rho] - x_m^* \rangle = \delta_\rho \langle f_m'(x[\rho, \delta_\rho]), \bar{v} - x_m^* \rangle \leq -\delta_\rho \epsilon \leq -\bar{s}\epsilon.$$

Combining this relation with (28), we get

$$\langle F(x[\rho, \delta_\rho]), x[\rho, \delta_\rho] - x^* \rangle \leq (m-1) \frac{\rho}{1-\rho} M - \bar{s}\epsilon.$$

For small $\rho > 0$, the right hand side in this inequality is < 0, while the left hand side is nonnegative for all $\rho \in (0, 1]$ since x^* is a weak solution to the variational inequality. We now have the desired contradiction. ∎

Relations between accuracy measures. We have associated each of the three functional problems with convex structure — convex minimization, convex-concave saddle point and convex Nash equilibrium problems — with their own accuracy measure. We showed above that each of these problems can also be associated with variational inequalities with

monotone operators and as such with the accuracy measure $\varepsilon_{\rm vi}(\cdot)$ resulting from these variational inequalities. Consequently, each of our three problems is equipped with two accuracy measures, and a natural question is, what is the relation between these measures? This question admits a simple answer in the case of convex minimization and convex-concave saddle point problems:

Proposition 2.1.3. [27] *Consider the variational inequality associated with (a) a convex minimization problem* (18)*, or (b) with a convex-concave saddle point problem* (21)*. Then for every $x \in X$ one has*

$$\varepsilon_{\rm vi}(x) \le \begin{cases} \varepsilon_{\rm opt}(x), & \text{in the case of (a)} \\ \varepsilon_{\rm sad}(x), & \text{in the case of (b)} \end{cases} \tag{29}$$

Proof. Let (a) be the case. Then, by construction, for every $y \in {\rm dom} F$ we have $F(y) \in \partial_X f(x)$, whence $\langle F(y), x - y\rangle \le f(x) - f(y)$ for all $y \in {\rm Dom}\, F$, $x \in X$, whence

$$\varepsilon_{\rm vi}(x) = \sup_{y \in {\rm Dom}\, F} \langle F(y), x - y\rangle \le \sup_{y \in {\rm Dom}\, F} [f(x) - f(y)] \le f(x) - \min_{y \in X} f(y) = \varepsilon_{\rm opt}(x).$$

Now let (b) be the case. Then for every $y = [y_1; y_2] \in {\rm Dom}\, F$ we have $F(y) = [f_1'(y); -f_2'(y)]$, where $f_1'(y)$ is a partial subgradient of $f(y)$ in y_1, and $f_2'(y)$ is a partial supergradient of $f(y)$ in y_2. consequently, for every $x = [x_1; x_2] \in X$ we have

$$\begin{aligned} \varepsilon_{\rm vi}(x) &= \sup_{y \in {\rm Dom}\, F} \langle F(y), x - y\rangle = \sup_{y \in {\rm Dom}\, F} \big[\underbrace{\langle f_1'(y), x_1 - y_1\rangle}_{\ge f(x_1,y_2) - f(y_1,y_2)} + \underbrace{\langle f_2'(y), y_2 - x_2\rangle}_{\ge f(y_1,y_2) - f(y_1,x_2)} \big] \\ &\le \sup_{y \in {\rm Dom}\, F} [f(x_1, y_2) - f(y_1, x_2)] = \overline{f}(x_1) - \underline{f}(x_2) = \varepsilon_{\rm sad}(x), \end{aligned}$$

■

2.2 Accuracy Certificates for Problems with Convex Structure

In this section we shall follow material presented in [27]; specifically we will define and investigate the entity of primary importance for our study — namely the *accuracy certificate*.

2.2.1 Accuracy Certificates: the Goal

For the sake of definiteness, let us restrict ourselves for a moment to the convex minimization problem in the form of (18). An "intelligent" solution algorithm for this problem should be capable of generating a feasible solution $\bar{x}$ of a pre-specified quality: $\varepsilon_{\rm opt}(\bar{x}) \le \epsilon$, where ϵ is a given-in-advance positive tolerance (which can be arbitrarily small). A related question

is: How can we *certify* that the resulting approximate solution meets these properties? The feasibility of $\bar{x}$, can usually be certified by calling on a Separation oracle for the feasible domain (which, essentially, means just checking that the constraints defining the feasible domain X of the problem are satisfied at $\bar{x}$, cf. Example 2.1.1). However certifying the accuracy specification $\varepsilon_{\rm opt}(\bar{x}) \leq \epsilon$ is a much more challenging problem,[4] and the structure of the associated certificates depends heavily on the structure of the convex minimization problem in question. Here is a list of examples:

- When (18) is a Linear Programming (or, more generally, a Conic Programming) program, the usual way to certify accuracy is to point out a feasible solution to the dual program such that the duality gap associated with the resulting primal-dual feasible pair of solutions is $\leq \epsilon$.

- Another way to certify accuracy is offered by self-concordance-based feasible start path-following interior point methods [30]; here the accuracy certificate is given by the fact that $\bar{x}$ is "close" (in a certain precise and verifiable sense) to a point on the central path with appropriately large value of the penalty parameter.

- Finally it may happen that the problem (18) in question can be represented as the primal component (P) of the primal-dual pair of optimization programs associated with a convex-concave saddle point problem (21). Assuming that the domains X_1, X_2 and the cost function $f(x_1, x_2)$ of (21) are sufficiently simple so that it is easy to compute the accuracy measure $\varepsilon_{\rm sad}$.[5] In this case, similarly to the LP/conic Programming case, a certificate for the relation $\varepsilon_{\rm opt}(\bar{x}) \leq \epsilon$ can be given by a feasible solution $\bar{y}$ to the dual problem (D) such that $\varepsilon_{\rm sad}(\bar{x}, \bar{y}) \leq \epsilon$. This method of certifying accuracy is utilized, for example, in recent first order methods (smoothing [33] and Mirror Prox algorithm [29]) for "well-structured" large-scale convex minimization.

Unfortunately the above outlined approaches for certifying accuracy in convex minimization are applicable only to "well structured" problems, and do not work in the general black-box-oriented setting we are interested in. This does not mean, however, that no black-box-oriented convex minimization algorithms capable of guaranteeing a pre-specified accuracy are known; this feature is shared, for example, by the polynomial time Cutting Plane algorithms (most notably, by the Ellipsoid method) and by "intelligent" gradient-type methods for smooth and nonsmooth convex minimization (primarily, by Subgradient and Mirror Descent algorithms and their bundle versions, see, e.g., [21, 31, 3, 4] and references therein). While in hindsight these algorithms, (primarily, the Bundle-Level method [21]), do suggest the notion of accuracy certificates as defined below, this notion, to the best of

[4] This is quite natural, since we don't know the true minimum the goal then is to certify a *negative* statement "there does *not* exist a feasible solution x with $f(x) < f(\bar{x}) - \epsilon$".

[5] This is the case, for example, when $f(x_1, x_2)$ is bilinear and it is easy to minimize linear functions over X_1 and X_2, such that it is easy to compute $\overline{f}(x_1)$ and $\underline{f}(x_2)$.

our knowledge, was explicitly defined for the first time only recently (specifically, in [27]). We are about to define formally what an accuracy certificate is.

2.2.2 Accuracy Certificate: Definition

Let us consider the following situation: We are given a pair (X, F) comprised of a solid $X \subset \mathbf{R}^n$ given by a Separation oracle, and an operator $F : \mathrm{Dom}\, F \to \mathbf{R}^n$ with convex domain $\mathrm{Dom}\, F$ (which is either $\mathrm{int}X$, or X) represented via a First Order oracle that, given an input $x \in \mathrm{Dom}\, F$, returns $F(x)$. Note that we have associated with every one of the aforementioned four problems with convex structure such a pair (X, F), and in all these cases F is monotone on its domain.

Now assume that we have at our disposal an algorithm which generates a sequence of *search points* $x_t \in \mathbf{R}^n$, $t = 1, 2, ...$, where at every step t the algorithm queries the Separation oracle, inputting the point x_t, and if the Separation oracle says that $x_t \in \mathrm{Dom}\, F$ (such as when $x \in \mathrm{int}X$), the algorithm invokes the First Order oracle to get $F(x_t)$.

- We refer to the steps t where the First Order oracle is invoked as *productive steps*, and
- we refer to the remaining steps as *non-productive* ones. Note that at a non-productive step t we have $x_t \notin \mathrm{int}X$, so that the Separation oracle provides a vector $e_t \neq 0$ which separates x_t and X, i.e an e_t s.t. $\langle e_t, x_t - x\rangle \geq 0$ for all $x \in X$.

The information acquired by such an algorithm during the course of $\tau = 1, 2, ...$ steps contains at a minimum the *τ-step execution protocol*

$$P_\tau = (\{x_t, e_t\}_{t=1}^{\tau}, I_\tau, J_\tau)\,,$$

where

- x_t is t-th search point,
- I_τ is an index set of productive steps $t \leq \tau$,
- J_τ is an index set of non-productive steps $t \leq \tau$, and
- $e_t = F(x_t)$ when the step t is productive; or
- $e_t \neq 0$ separates x_t and X, when the step t is non-productive.

The central notions of an *Accuracy Certificate associated with an execution protocol P_τ* are:
Definition 2.2.1. [27] Let $P_\tau = (\{x_t, e_t\}_{t=1}^{\tau}, I_\tau, J_\tau)$ be an execution protocol for X, F, and let $\mathbf{B}$ be a solid containing X.

(i) An *accuracy certificate* associated with an execution protocol P_τ is an ordered collection $\xi = \{\xi_t \geq 0\}_{t=1}^{\tau}$ of nonnegative weights such that $\sum_{t\in I_\tau} \xi_t = 1$.
Note that a certificate can only be associated with a *productive* execution protocol — one with $I_\tau \neq \emptyset$.

(ii) The *resolution* of an accuracy certificate ξ is the quantity

$$\varepsilon_{\rm cert}(\xi|P_\tau, \mathbf{B}) = \max_{x\in\mathbf{B}} \sum_{t=1}^{\tau} \xi_t \langle e_t, x_t - x\rangle. \tag{30}$$

(ii) The *approximate solution* induced by an accuracy certificate ξ is the vector

$$\widehat{x} = \widehat{x}(P_\tau, \xi^\tau) := \sum_{t\in I_\tau} \xi_t x_t. \tag{31}$$

While at first glance these definitions may seem strange, we now provide the justification.

2.2.3 Accuracy Certificates: Justification

Theorem 2.2.1. [27] *Let a pair (X, F) originate from a problem with convex structure, that is, from*

(a) a convex optimization problem,

(b) a convex-concave saddle point problem,

(c) a convex Nash equilibrium problem, or

(d) a variational inequality with a monotone operator.

Let P_τ be an execution protocol for (X, F), let ξ^τ be an accuracy certificate associated with this protocol, and let $\widehat{x} = \widehat{x}(P_\tau, \xi^\tau)$.

Then:
(i) $\widehat{x} \in X$, so that $\widehat{x}$ is a feasible solution to the problem underlying X, F, and
(ii) for every solid $\mathbf{B} \supset X$ the resolution $\varepsilon_{\rm cert}(\xi^\tau|P_\tau, \mathbf{B})$ of ξ^τ w.r.t. P_τ and $\mathbf{B}$ is an upper bound on the corresponding accuracy measure of $\widehat{x}$, so that

- $\varepsilon_{\rm opt}(\widehat{x}) \leq \varepsilon_{\rm cert}$ *in the case of (a),*
- $\varepsilon_{\rm sad}(\widehat{x}) \leq \varepsilon_{\rm cert}$ *in the case of (b),*
- $\varepsilon_{\rm N}(\widehat{x}) \leq \varepsilon_{\rm cert}$ *in the case of (c), and*
- $\varepsilon_{\rm vi}(\widehat{x}) \leq \varepsilon_{\rm cert}$ *in the case of (d).*

Proof.
(i): The fact that $\widehat{x} \in \operatorname{Dom} F \subset X$ is evident, since by definition of an accuracy certificate $\widehat{x}$ is a convex combination of points x_t, $t \in I_\tau$, and all these points belong to the convex set $\operatorname{Dom} F \subset X$.
(ii): To prove the case (d) of $\varepsilon_{\rm vi}(\widehat{x}) \le \varepsilon_{\rm cert}$, note that

$$
\begin{aligned}
&\forall y \in \operatorname{Dom} F: \\
&\langle F(y), \bar{x} - y\rangle = \langle F(y), \textstyle\sum_{t\in I_\tau} \xi_t x_t - y\rangle \text{ [definition of } \widehat{x}] \\
&= \textstyle\sum_{t\in I_\tau} \xi_t \langle F(y), x_t - y\rangle \text{ [since } \textstyle\sum_{t\in I_\tau} \xi_t = 1] \\
&\le \textstyle\sum_{t\in I_\tau} \xi_t \langle F(x_t), x_t - y\rangle \text{ [since } \xi_t \ge 0 \text{ and } F \text{ is monotone]} \\
&\le \textstyle\sum_{t\in I_\tau} \xi_t \langle F(x_t), x_t - y\rangle + \sum_{t\in J_\tau} \xi_t \langle e_t, x_t - y\rangle \quad \left[\begin{array}{l} \text{since } e_t \text{ separates } x_t \text{ and } X \ni y \\ \text{when } t \in J_\tau \end{array}\right] \\
&= \textstyle\sum_{t=1}^{\tau} \xi_t \langle e_t, x_t - y\rangle \\
&\le \varepsilon_{\rm cert}(\xi|P_\tau, \mathbf{B}) \text{ [by definition of } \varepsilon_{\rm cert} \text{ and due to } y \in X \subset \mathbf{B}]
\end{aligned}
$$

Since this holds for $\forall y \in \operatorname{Dom} F$ then it certainly holds for $y^* = \operatorname{argmax}\langle F(y), x - y\rangle$. Hence $\varepsilon_{\rm vi} = \sup_{y \in \operatorname{Dom} F} \langle F(y), x - y\rangle \le \varepsilon_{\rm cert}$

To prove the case (a) of $\varepsilon_{\rm opt}(\widehat{x}) \le \varepsilon_{\rm cert}$, note that

$$
\begin{aligned}
&\forall y \in X: \\
&f(\widehat{x}) - f(y) = f(\textstyle\sum_{t\in I_\tau} \xi_t x_t) - f(y) \qquad \text{[definition of } \widehat{x}] \\
&\le \textstyle\sum_{t\in I_\tau} \xi_t f(x_t) - f(y) \quad \left[\begin{array}{l} \text{since } f \text{ is convex on } X,\ \xi_t \ge 0,\ \sum_{t\in I_\tau} \xi_t = 1 \\ \text{and } x_t \in X,\ \forall t \in I_\tau, \text{ and via Jensen's Inequality.} \end{array}\right] \\
&= \textstyle\sum_{t\in I_\tau} \xi_t [f(x_t) - f(y)] \text{ [since } \textstyle\sum_{t\in I_\tau} \xi_t = 1] \\
&\le \textstyle\sum_{t\in I_\tau} \xi_t \langle e_t, x_t - y\rangle \quad \left[\begin{array}{l} \text{since } e_t \in \partial f(x_t) \text{ when } t \in I_\tau \text{ and } \xi_t \ge 0, \text{ by applying} \\ \text{the gradient inequality for convex functions.} \end{array}\right] \\
&\le \textstyle\sum_{t=1}^{\tau} \xi_t \langle e_t, x_t - y\rangle \text{ [since } \xi_t \ge 0 \text{ and } e_t \text{ separates } x_t \text{ and } X \ni y \text{ when } t \in J_\tau] \\
&\le \varepsilon_{\rm cert}(\xi|P_\tau, \mathbf{B}) \text{ [by definition of } \varepsilon_{\rm cert} \text{ and due to } y \in X \subset \mathbf{B}],
\end{aligned}
$$

as required for the case of (a). Since this holds for any $y \in X$ it certainly holds for $y^* = \operatorname{argmin} f(x)$. Hence $\varepsilon_{\rm opt} \le \varepsilon_{\rm cert}$.

Recalling that a convex-concave saddle point problem is just the zero sum 2-player case of a convex Nash equilibrium problem, (see section 2.1.3), all that remains is to prove the case

(c) of $\varepsilon_{\rm N}(\widehat{x}) \le \varepsilon_{\rm cert}$. To this end note that[6]

$$
\begin{array}{l}
\forall y = [y_1; ...; y_m] \in X = X_1 \times ... \times X_m : \\
\sum_{i=1}^m \left[f_i(\widehat{x}) - f_i(\widehat{x}^i, y_i) \right] = \Phi(\widehat{x}) - \sum_{i=1}^m f_i(\widehat{x}^i, y_i) \ \text{[definition of } \Phi] \\
= \Phi(\sum_{t\in I_\tau} x_t) - \sum_{i=1}^m f_i(\sum_{t\in I_\tau} \xi_t x_t^i, y_i) \ \text{[definition of } \widehat{x}] \\
\le \sum_{t\in I_\tau} \xi_t \Phi(x_t) - \sum_{t\in I_\tau} \sum_{i=1}^m f(x_t^i, y_i) \quad \left[\begin{array}{l} \text{since } \Phi \text{ is convex, } f_i(z^i, z_i) \text{ is concave in} \\ z^i, \text{ while } \xi_t \ge 0 \text{ and } \sum_{t\in I_\tau} \xi_t = 1 \end{array} \right] \\
= \sum_{t\in I_\tau} \xi_t \sum_{i=1}^m \left[f_i(x_t) - f_i(x_t^i, y_i) \right] \ \text{[definition of } \Phi] \\
\le \sum_{t\in I_\tau} \xi_t \sum_{i=1}^m \langle f_i'(x_t), (x_t)_i - y_i \rangle \ \text{[since } f_i(z^i, z_i) \text{ is convex in } z_i] \\
= \sum_{t\in I_\tau} \xi_t \langle e_t, x_t - y \rangle \ \text{[by definition of } e_t \in \partial f(x_t), \text{ when } t \in I_\tau] \\
\le \sum_{t=1}^\tau \xi_t \langle e_t, x_t - y \rangle \ \text{[since } \xi_t \ge 0 \text{ and } e_t \text{ separates } x_t \text{ and } X \ni y \text{ when } t \in J_\tau] \\
\le \varepsilon_{\rm cert}(\xi | P_\tau, \mathbf{B}) \ \text{[by definition of } \varepsilon_{\rm cert} \text{ and due to } y \in X \subset \mathbf{B}],
\end{array}
$$

as required in the case of (c). ∎

Intermediate summary. Observe that when $\mathbf{B}$ is "simple", so that it is easy to maximize linear forms over $\mathbf{B}$ (for example, $\mathbf{B}$ is an Euclidean ball, or a box, or a simplex), it is easy to verify, given an execution protocol P_τ, that a given candidate accuracy certificate ξ^τ is indeed an accuracy certificate, and when this holds, it is also easy to compute its resolution. This fact combined with Theorem 2.2.1 implies that an accuracy certificate with resolution $\varepsilon_{\rm cert} \le \epsilon$ is both a simple way to build a feasible approximate solution to a problem with convex structure and a "simple proof" of the fact that this solution is ϵ-optimal.

This knowledge motivates a desire to equip black box oriented algorithms for problems with convex structure with computationally cheap techniques for building accuracy certificates. These certificates can be used in stopping rules for both identifying the termination step ("stop when the resolution of the current accuracy certificate is $\le$ a desired tolerance ϵ") and generating the resulting feasible and ϵ-optimal approximate solution. In respect to this desire, a significant step was made in [27], where computationally cheap certificate-generating techniques were proposed for the *cutting plane* algorithms, most notably, for the Ellipsoid method. Here is the corresponding result from [27]:

Proposition 2.2.1. *Let $X \subset \mathbf{R}^n$ be a solid given by Separation oracle and contained in the centered at the origin Euclidean ball $\mathbf{B} = \{x \in \mathbf{R}^n : \|x\|_2 \le R\}$ of a given radius R, and let*

$$F : {\rm int} X \to \mathbf{R}^n$$

[6] Recall our notation, for $z = [z_1; ...; z_m] \in X$ let us set $z^i = [z_1; ...; z_{i-1}; z_{i+1}; ...; z_m]$, so that $f_i(z) = f_i(z^i, z_i)$

be a semi-bounded vector field, meaning that the X-variation of F – the quantity

$$\mathrm{Var}_X(F) := \sup_{x\in \mathrm{int}X, y\in X} \langle F(x), y-x\rangle$$

– is finite.

The Ellipsoid algorithm can be equipped with rules for building accuracy certificates in such a way that

(i) *The resulting algorithm, as applied to (X,F), at every step τ*

1. *makes a single call to the Separation oracle at the current search point x_τ,*
2. *when Separation oracle reports that $x_\tau \in \mathrm{int}X$, the algorithm makes a single call to the F-oracle to get $F(x_\tau)$,*
3. *given the answer(s) of the oracle(s), produces*
   (a) *the current execution protocol*

$$\left[\begin{array}{c} P_\tau = (\{x_t, e_t\}_{t=1}^\tau, I_\tau, J_\tau)\,, \\ I_\tau = \{t\le\tau : x_t\in \mathrm{int}X\}, J_\tau = \{t\le\tau : x_t\notin \mathrm{int}X\}, \\ t\in I_\tau \Rightarrow e_t = F(x_t),\ t\in J_t \Rightarrow e_t\neq 0 \,\&\, \langle e_t, x_t - x\rangle \ge 0, \forall x\in X \end{array}\right]$$

   (b) *an accuracy certificate ξ^τ for P_τ, provided $I_\tau\neq\emptyset$*

   (c) *the next search point x_{t+1}*

(ii) *The resulting algorithm ensures that for every $\epsilon>0$, the number $\tau=\tau(\epsilon)$ of steps until an accuracy certificate ξ^τ with $\varepsilon_{\mathrm{cert}}(\xi^\tau|P_\tau,\mathbf{B})\le\epsilon$ is built, is upper-bounded by*

$$O(1)n^2\ln\left(\frac{nR[\mathrm{Var}_X(F)+\epsilon]}{r\epsilon}\right),$$

where r is the largest of radii of Euclidean balls contained in X;

(iii) *For the resulting algorithm, for every τ, the average, over the first τ steps, "computational overhead" per step (the number of arithmetic operations except for those carried out by the oracles) does not exceed $O(1)n^2$.*

This Proposition combines with Theorem 2.2.1 to imply that a problem with convex structure on a solid $X\subset\mathbf{R}^n$ (known to belong to a given Euclidean ball of radius R) can be solved by the Ellipsoid method within (any) desired accuracy ϵ in $O(1)n^2\ln\left(\frac{nR}{r}\frac{V+\epsilon}{\epsilon}\right)$ steps, where V is the X-variation of the monotone operator associated with the problem, and r is the maximum of radii of Euclidean balls contained in X.

In the next chapter we make another step in this direction, specifically, equip with accuracy certificates one of the most attractive state-of-the-art black-box oriented algorithms for *large-scale* non-smooth problems with convex structure - the (generalized) Non-Euclidian Restricted Memory Level Method (NERML, [4]).

CHAPTER III

ACCURACY CERTIFICATES IN LARGE-SCALE FIRST ORDER MINIMIZATION

3.1 Motivation

While cutting plane algorithms play a crucial role in the theory of Convex Programming — these are the algorithms underlying the most general results on polynomial time solvability of generic Convex Programming problems, see [2], their role as practical computational tools is rather restricted. The reason is that the operation count of these algorithms as applied to solving to within an accuracy of ϵ, n-dimensional problems with convex structure is *at least* $O(1)n^4\ln(1/\epsilon)$. While polynomial in n and $\ln(1/\epsilon)$, this operation count grows rapidly with n, which makes these cutting plane algorithms impractical for problems with several hundred or more variables. For well structured convex problems, like Linear, Conic Quadratic, and Semidefinite Programming, a viable alternative is offered by *polynomial time interior point methods* (see, e.g., [2]); however, a single iteration of an interior point method is computationally demanding in the large scale case.[1] As a result, interior point methods as applied to really large-scale problems (tens of thousands or more variables) often become completely impractical. Hence, when speaking about these huge-scale well structured convex problems (or about medium- or large-scale convex problems with no transparent structure), computationally cheap gradient-type methods are the algorithms of choice. These algorithms are black box oriented, and as such obey the theoretical limits of performance established in Information-based Complexity Theory (see [46]) stating, essentially, that the number of iterations required to solve an n-dimensional problem with convex structure to within an accuracy of ϵ for large n is at least $O(\epsilon^{-2})$. Since this lower bound grows rapidly when $\epsilon \to +0$, it follows that gradient-type methods cannot be used to solve large-scale problems to within a high accuracy. On the positive side, an attractive property of these methods is that in the case of problems with "favorable geometry," the iteration complexity (the number of calls to the oracles representing the problem) for finding an ϵ-solution is dimension-independent (like $O(\epsilon^{-2})$) or nearly so (like $O(\ln(n)\epsilon^{-2})$), and an

[1] $O(n^3)$ operations, unless the problem's data possess a favorable sparsity pattern; this is usually the case in Linear Programming problems originating in decision making, and usually is *not* the case for semidefinite programs, or for LP problems originating in Signal Processing and several other areas.

iteration is "cheap" (just $O(n)$ operations on top of the computational expenses of the oracles). This makes computationally cheap gradient-type methods attractive techniques for finding low- and medium-accuracy solutions to large-scale "favorable geometry" problems with convex structure. This motivates, in particular, our goal in this chapter – extending the computational techniques for building accuracy certificates to gradient-type black box oriented algorithms. As a prototype algorithm in this development, we have chosen the state-of-the-art Non-Euclidean Restricted Memory Level method (NERML, [4]).

3.2 NERML Algorithm with Certificates

NERML is aimed at solving problems with convex structure as defined in the previous chapter and works with the variational inequalities associated with these problems, that is, finally, with convex domains X and monotone operators on these domains.

3.2.1 NERML with Certificates: Setup

The setup for the NERML algorithm is given by

1. A solid $\mathbf{B} \subset \mathbf{R}^n$ which contains the domain X of interest.

2. A norm $\|\cdot\|$ (not necessarily the Euclidean one) on the embedding space $\mathbf{R}^n$ of $\mathbf{B}$. The norm conjugate to the norm $\|\cdot\|$ is denoted by $\|\cdot\|_*$:

$$\|\xi\|_* = \max_x \{\langle \xi, x\rangle : \|x\| \le 1\}.$$

3. A strongly convex and continuously differentiable *distance generating function* $\omega(\cdot)$ on $\mathbf{B}$, strong convexity meaning that there exists $\alpha > 0$ (called *modulus of strong convexity of* ω *w.r.t.* $\|\cdot\|$) such that

$$\langle \omega'(x) - \omega'(y), x - y\rangle \ge \alpha \|x - y\|^2 \ \forall x, y \in \mathbf{B}.$$

We associate with $\mathbf{B}$ and ω the "ω-size of $\mathbf{B}$" – the quantity

$$\Omega = \max_{x,y\in\mathbf{B}} \left[\omega(y) - \omega(x) - \langle \omega'(x), y - x\rangle\right]. \tag{32}$$

Example 3.2.1. [Euclidean Setup] *The simplest setup for NERML is the one where* $\|\cdot\|$ *is the usual Euclidean norm* $\|\cdot\|_2$ *on* $\mathbf{R}^n$*, and* $\omega(x) = \frac{1}{2}x^T x$*. In this case,* $\|\cdot\|_* = \|\cdot\|_2$*,* $\alpha = 1$ *and* $\Omega = \max_{x,y\in\mathbf{B}} \frac{1}{2}\|x - y\|_2^2$*. In particular, when* $\mathbf{B}$ *is contained in the* $\|\cdot\|_2$ *ball of radius* R*, we have* $\Omega \le 2R^2$*.*

Example 3.2.2. [ℓ_1 Setup] Here $\mathbf{B}$ is contained in the standard simplex $\Delta_R = \{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i \leq R\}$ of size R, $\|\cdot\|$ is $\|\cdot\|_1$ (so that $\|\cdot\|_*$ is $\|\cdot\|_\infty$), and $\omega(x)$ is scaled and regularized entropy:

$$\omega(x) = R(1+\kappa)\sum_{i=1}^{n}(x_i + R\kappa/n)\ln(x_i + R\kappa/n),$$

where $\kappa > 0$ is a once for ever fixed small regularization parameter (say, $\kappa = 1.e-16$). In this case

$$\alpha = 1,\ \Omega \leq 2\ln(n/\kappa)R^2 \leq O(1)R^2\ln n.$$

Here is the computation proving the above facts. Due to regularization, $\omega(\cdot)$ is C^∞ on $\mathbf{B}$, and this function clearly is convex. In order to prove strong convexity with the parameter $\alpha = 1$ w.r.t. $\|\cdot\| = \|\cdot\|_1$, it suffices to verify that $h^T\omega''(x)h \geq \|h\|_1^2$ for all h and for all $x \in \mathbf{B}$. Indeed, with $x \in \mathbf{B}$, setting $a_i = \frac{x_i + R\kappa/n}{R(1+\kappa)}$, we have $a_i > 0$ and $\sum_i a_i \leq 1$ due to $x_i \geq 0$ and $\sum_i x_i \leq R$. It follows that

$$\begin{array}{l} ||h||_1^2 = \left[\sum_i |h_i|\right]^2 = \left[\sum_i [|h_i|a_i^{-1/2}]a_i^{1/2}\right]^2 \leq \left[\sum_i h_i^2/a_i\right]\left[\sum_i a_i\right] \\ \leq \sum_i h_i^2/a_i = h^T\omega''(x)h. \end{array}$$

Further, when $x, y \in \mathbf{B}$, setting $\bar{y}_i = y_i + R\kappa/n$, $\bar{x}_i = x_i + R\kappa/n$, one has

$$\begin{array}{l} \omega(y) - \omega(x) - \langle \omega'(x), y - x\rangle = R(1+\kappa)\left[\sum_i \bar{y}_i \ln(\bar{y}_i/\bar{x}_i) + \sum_i \bar{x}_i - \sum_i \bar{y}_i\right] \\ \hfill \text{[direct computation]} \\ \leq R(1+\kappa)\left[\sum_i \bar{x}_i + \sum_i \bar{y}_i \ln(\bar{y}_i/\bar{x}_i)\right]\ \ [\text{since } \bar{y}_i \geq 0] \\ \leq R(1+\kappa)\left[\sum_i \bar{x}_i + \sum_i \bar{y}_i \ln((1+\kappa)n/\kappa)\right]\ \ [\text{since } \bar{y}_i/\bar{x}_i \leq (1+\kappa)n/\kappa] \\ \leq R^2(1+\kappa)^2\left[1 + \ln((1+\kappa)n/\kappa)\right] \\ \hfill [\text{since } \sum_i \bar{x}_i \leq R(1+\kappa),\ \sum_i \bar{y}_i \leq R(1+\kappa)] \\ \leq 2R^2\ln(n/\kappa)\ [\text{since } \kappa = \text{1.e-16}] \end{array}$$

Example 3.2.3. [Spectahedron Setup] With this setup, $\mathbf{B}$ is the *spectahedron* – the set of positive semidefinite matrices of trace $\leq R$ – in the space $\mathbf{S}^\nu$ of symmetric matrices of block-diagonal structure $\nu = (\nu_1, ..., \nu_k)$ (that is, matrices from $\mathbf{S}^\nu$ are symmetric block-diagonal with k diagonal blocks of sizes $\nu_1, ..., \nu_k$). The embedding space $\mathbf{S}^\nu$ is equipped with the *Frobenius inner product* $\langle x, y\rangle = Tr(xy)$ and the *trace norm* $\|x\| = \sum_{i=1}^{|\nu|}\lambda_i(x)$, where $|\nu| = \sum_{\ell=1}^{k}\nu_\ell$ is the row size of matrices $x \in \mathbf{S}^\nu$, and $\lambda_1(x) \geq \lambda_2(x) \geq ... \geq \lambda_s(x)$ are the eigenvalues of a symmetric $s \times s$ matrix x. The distance-generating function is the *regularized matrix entropy*

$$\omega(x) = 2R(1+\kappa)\sum_{i=1}^{|\nu|}(\lambda_i(x) + R\kappa/|\nu|)\ln(\lambda_i(x) + R\kappa/|\nu|),$$

where κ plays the same role as in the case of ℓ_1 setup; we lose nothing when assuming that $\kappa = 1.e-16$. For this setup, similarly to the ℓ_1-one,

$$\alpha = 1,\ \Omega \le O(1)\ln(|\nu|/\kappa)R^2$$

see [4].

Note that when $\nu = (1, ..., 1)$, that is, $\mathbf{S}^n u$ is the space of diagonal matrices, identifying diagonal matrices and the vectors of their diagonal entries, spectahedrons become simplexes, and the Spectahedron setup essentially reduces to the ℓ_1 setup.

Example 3.2.4. [Mixed Setup] In some cases (specially when solving saddle point and Nash Equilibrium problems), the domain X in question is given as a direct product of domains $X_i \subset E_i$, $1 \le i \le k$, where X_i are convex domains, and E_i are Euclidean spaces embedding X_i. In this case, a typical way to specify NERML setup is to combine NERML setups for domains X_i, that is, to specify for $i \le k$ solids B_i, $X_i \subset B_i \subset E_i$, norms $\|\cdot\|_{(i)}$ on E_i, and distance-generating functions $\omega_i(x^i)$ for B_i, and set

$$\begin{array}{ll}(a) & \mathbf{B} = \mathbf{B}_1 \times ... \times \mathbf{B}_k \\ (b) & \omega(x^1, ..., x^k) = \sum_{i=1}^k \frac{p_i}{\alpha_i}\omega_i(x^i) \\ (c) & \|[x^1; ...; x^k]\| = \sqrt{\sum_{i=1}^k p_i \|x^i\|_{(i)}^2}\end{array}$$

where positive reals p_i are parameters of the construction, and α_i are moduli of strong convexity of $\omega_i(\cdot)$ w.r.t. $\|\cdot\|_{(i)}$. It is immediately seen that $\omega(\cdot)$ is indeed a distance-generating function for $\mathbf{B}$ with modulus of strong convexity w.r.t. $\|\cdot\|$ equal to 1, and the ω-size of $\mathbf{B}$ does not exceed the quantity

$$\Omega = \sum_{i=1}^k \frac{p_i}{\alpha_i}\Omega_i,$$

where Ω_i is the ω_i-size of $\mathbf{B}_i$.

The "free parameters" $p_1, ..., p_k$ of the construction can be chosen in order to optimize the efficiency estimate of the associated version of NERML.

Implementability. In order for NERML (with or without certificates) to be practical, $\mathbf{B}$ and ω should be "simple" and "match" each other, meaning that it should be easy to solve auxiliary optimization problems of the form

$$\min_{x \in \mathbf{B}} [\langle a, x\rangle + \omega(x)]$$

for any linear form a. From now on we assume this to be true.

For example, with the Euclidean setup, $\mathbf{B}$ matches ω when $\mathbf{B}$ is a ball, a box, a standard simplex, or a direct product of these basic domains. With the ℓ_1/Spectahedron setup, $\mathbf{B}$

matches ω when $\mathbf{B}$ is a standard simplex/spectahedron or is the intersection of a simplex with a box. Note that passing from partial setups to a mixed one (see Example 3.2.4), one preserves "matching:" if $\mathbf{B}_i$ matches ω_i for all i, then $\mathbf{B}$ matches ω.

3.2.2 NERML with Certificates: Data and Goal

Let us fix the setup $(\mathbf{B} \subset \mathbf{R}^n, \|\cdot\|, \omega(\cdot))$ of the NERML algorithm.

The data for the NERML algorithm is a vector field

$$g(x) : \mathbf{B} \mapsto \mathbf{R}^n$$

assumed to be bounded by 1:

$$\|g(x)\|_* \leq 1 \ \forall x \in \mathbf{B}. \tag{33}$$

Given the data, we

- associate with $x \in \mathbf{B}$ the affine function

$$h_x(y) = \langle g(x), x - y \rangle : \mathbf{R}^n \to \mathbf{R};$$

- associate with a finite set $S \subset \mathbf{B}$ the family $\mathcal{F}_S$ of affine functions on $\mathbf{R}^n$ which are convex combinations of functions $h_x(\cdot)$, $x \in S$.

In the sequel, the words "we have at our disposal a function $h(\cdot) \in \mathcal{F}_S$" mean that we know the functions $h_x(\cdot)$, $x \in S$, and nonnegative weights λ_x, $x \in S$, summing up to 1, such that $h(y) = \sum_{x \in S} \lambda_x h_x(y)$.

The goal of the algorithm is, given a tolerance $\delta > 0$, to find a finite set $S \subset \mathbf{B}$ and a function $h(\cdot) \in \mathcal{F}_S$ such that

$$\max_{y \in \mathbf{B}} h(y) \leq \delta. \tag{34}$$

As we shall see in Sections 3.3, 3.4, solving a problem with convex structure can be reduced to achieving the above goal for an appropriate vector field $g(\cdot)$ associated with the problem of interest.

3.2.3 NERML with Certificates: Construction

The NERML algorithm with certificates builds search sequences $x_1 \in \mathbf{B}, x_2 \in \mathbf{B}, ...$ along with the sets $S_t = \{x_1, ..., x_t\} \subset \mathbf{B}$, according to the following rules:

Initialization. We choose an arbitrary $x_1 \in \mathbf{B}$ and set $f_1 = \max\limits_{y \in \mathbf{B}} h_{x_1}(y)$. We clearly have $f_1 \geq 0$.

- In the case of $f_1 = 0$, we terminate and output $h(x) = h_{x_1}(x) \in \mathcal{F}_{S_1}$, thus ensuring (34) with $\delta = 0$.
- When $f_1 > 0$, we proceed. Our subsequent actions are split into *phases* enumerated 1,2,...

Phase $s = 1, 2, ...$ At the beginning of phase s, we have at our disposal

- the set $S^s = \{x_1, ..., x_{t_s}\} \subset \mathbf{B}$ of already built search points, and
- an affine function $h^s(\cdot) \in \mathcal{F}_{S^s}$ along with the real $f_s := \max\limits_{x \in \mathbf{B}} h^s(x) \in (0, f_1]$.

We define the *level* ℓ_s of phase s as

$$\ell_s = (1 - \gamma) f_s,$$

where $\gamma \in (0, 1)$ is a control parameter of the method. Note that $\ell_s > 0$ due to $f_s > 0$.

To save notation, we denote the search points generated at phase s as $u_1, u_2, ...$, so that $x_{t_s+t} = u_t$, $t = 1, 2, ...$

Initializing phase s**.** We somehow choose

1. A collection of m functions $h^s_{0,j}(\cdot) \in \mathcal{F}_{S^s}$, $1 \leq j \leq m$, such that the set

$$X^s_0 = \mathrm{cl}\,\{x \in \mathbf{B} : h^s_{0,j}(x) > \ell_s, 1 \leq j \leq m\}$$

   is nonempty,

2. A *prox center* $u_1 \in X^s_0$,

and set

$$\omega_s(x) = \omega(x) - \omega(u_1) - \langle x - u_1, \omega'(u_1) \rangle.$$

Here a positive integer m is a control parameter of the method.

Note that to ensure the above requirements, we can set $u_1 = \mathrm{argmax}_{x \in \mathbf{B}} h_1(x)$ and choose all $h^s_{0,j}$ equal to $h_1(\cdot)$, thus ensuring that $h^s_{0,j}(u_1) = f_1 \geq f_s > \ell_s$.

Step $t = 1, 2, ...$ of phase s:

• **At the beginning of step** t, we have at our disposal

1. The set S^s_{t-1} of all previous search points;

2. A collection of functions $\{h^s_{t-1,j}(\cdot) \in \mathcal{F}_{S^s_{t-1}}\}_{j=1}^m$ such that the set

$$X^s_{t-1} = \mathrm{cl}\,\{x \in \mathbf{B} : h^s_{t,j}(x) > \ell_s, 1 \le j \le m\}$$

is nonempty,

3. Current search point $u_t \in X^s_{t-1}$ such that

$$u_t = \operatorname*{argmin}_{x \in X^s_{t-1}} \omega_s(x) \tag{Π^s_t}$$

Note that this relation is trivially true when $t = 1$.

• **Our actions at step** t are as follows.

1. We compute $g(u_t)$ and set

$$h_{t-1,m+1}(x) = \langle g(u_t), u_t - x\rangle.$$

2. We solve the auxiliary problem

$$\mathrm{Opt} = \max_{x \in \mathbf{B}} \min_{1 \le j \le m+1} h_{t-1,j}(x) \tag{35}$$

Note that

$$\begin{aligned}\mathrm{Opt} &= \max_{x\in\mathbf{B}} \min_{\lambda_j \ge 0, \sum_j \lambda_j = 1} \sum_{j=1}^{m+1} \lambda_j h^s_{t-1,j}(x) = \min_{\lambda_j \ge 0, \sum_j \lambda_j = 1} \max_{x\in\mathbf{B}} \sum_{j=1}^{m+1} \lambda_j h^s_{t-1,j}(x) \\ &= \max_{x\in\mathbf{B}} \textstyle\sum_{j=1}^{m+1} \lambda^t_j h^s_{t-1,j}(x),\end{aligned}$$

where $\lambda^t_j \ge 0$ and $\sum_{j=1}^{m+1} \lambda^t_j = 1$. We assume that when solving the auxiliary problem, we compute the above weights λ^t_j, and thus have at our disposal the function

$$h^{s,t}(\cdot) = \sum_{j=1}^{m+1} \lambda^t_j h^s_{t-1,j}(\cdot) \in \mathcal{F}_{S^s_t}$$

such that

$$\mathrm{Opt} = \max_{x\in\mathbf{B}} h^{s,t}(x).$$

2.A: It may happen that $\mathrm{Opt} \le \delta$. In this case we terminate and output $h^{s,t}(\cdot) \in \mathcal{F}_{S^s_t}$; this function satisfies (34).

2.B: It may happen that $\mathrm{Opt} < \ell_s + \theta(f_s - \ell_s)$, where $\theta \in (0, 1)$ is method's

control parameter. In this case, we terminate phase s and start phase $s+1$ by setting $h^{s+1} = h^{s,t}$, $f_{s+1} = \text{Opt}$. Note that by construction $0 < f_{s+1} \leq [\gamma + \theta(1-\gamma)]f_s \leq f_1$, so that we have at our disposal all we need to start phase $s+1$.

2.C: When neither 2.A, nor 2.B take place, we proceed with phase s, specifically, as follows:

(a) We are in the situation when there exists a point $u \in \mathbf{B}$ such that $h^s_{t-1,j}(u) \geq \text{Opt} > \ell_s$, so that the set $Y_t = \{x \in \mathbf{B} : h^s_{t-1,j}(x) \geq \ell_s,\, 1 \leq j \leq m+1\}$, has a nonempty interior. We specify u_{t+1} as

$$u_{t+1} = \operatorname*{argmin}_{x \in Y_t} \omega_s(x). \tag{36}$$

Observe that

$$u_{t+1} \in X^s_{t-1} \tag{37}$$

due to $Y_t \subset X^s_{t-1}$.

(b) By optimality conditions for (36), for certain nonnegative μ_j, $1 \leq j \leq m+1$, such that

$$\mu_j[h^s_{t-1,j}(u_{t+1}) - \ell_s] = 0,\, 1 \leq j \leq m+1,$$

the vector

$$e := \omega_s'(u_{t+1}) - \sum_{j=1}^{m+1} \mu_j \nabla h^s_{t-1,j}(\cdot) \tag{38}$$

is such that

$$\langle e, x - u_{t+1}\rangle \geq 0 \;\forall x \in \mathbf{B}. \tag{39}$$

• In the case of $\mu = \sum_j \mu_j > 0$, we set

$$h^s_{t,1} = \frac{1}{\mu} \sum_{j=1}^{m+1} \mu_j h^s_{t-1,j},$$

so that

$$\begin{aligned} &(a) \quad h^s_{t,1} \in \mathcal{F}_{S^s_t}, \\ &(b) \quad h^s_{t,1}(u_{t+1}) = \ell_s, \\ &(c) \quad \langle \omega_s'(u_{t+1}) - \mu \nabla h^s_{t,1}, x - u_{t+1}\rangle \geq 0 \,\forall x \in \mathbf{B} \end{aligned} \tag{40}$$

We then discard from the collection $\{h^s_{t-1,j}(\cdot)\}_{j=1}^{m+1}$ two (arbitrarily chosen) elements and add to $h^s_{t,1}$ the remaining $m-1$ elements of the collection, thus getting an m-element collection $\{h^s_{t,j}\}_{j=1}^m$ of elements of $\mathcal{F}_{S^s_t}$.

Remark 3.2.1. We have ensured that the set $X^s_t = \text{cl}\,\{x \in \mathbf{B} : h^s_{t,j} > \ell_s, 1 \leq j \leq m\}$ is nonempty (indeed, we clearly have $h^s_{t,j}(\widehat{u}) > \ell_s$, $1 \leq j \leq m$, where $\widehat{u}$ is an optimal solution to (35)). Besides this, we have ensured (Π^s_{t+1}). Indeed, by

construction $u_{t+1} \in Y_t$, meaning that $h^s_{t-1,j}(u_{t+1}) \geq \ell_s$, $1 \leq j \leq m+1$; since $h^s_{t,j}$ by construction are convex combinations of the functions $h^s_{t-1,j}$, $1 \leq j \leq m+1$, it follows that $u_{t+1} \in X^s_{t+1}$. Besides this, (40.*b*-*c*) imply that

$$u_{t+1} = \operatorname*{argmin}_x \left\{\omega_s(x) : x \in \mathbf{B}, h^s_{t,1}(x) \geq \ell_s\right\},$$

and the right hand side set clearly contains X^s_{t+1}. Thus, u_{t+1} indeed is the minimizer of $\omega_s(\cdot)$ over X^s_{t+1}.

• In the case of $\mu = 0$, (38) – (39) say that u_{t+1} is a minimizer of $\omega_s(\cdot)$ on $\mathbf{B}$. In this case, we discard from the collection $\{h^s_{t-1,j}\}_{j=1}^{m+1}$ one (arbitrarily chosen) element, thus getting the m-element collection $\{h^s_{t,j}\}_{j=1}^{m}$. Here, by exactly the same reasons as above, the set $X^s_{t+1} := \mathrm{cl}\{x \in \mathbf{B} : h^s_{t,j}(x) > \ell_s\}$ is nonempty and contains u_{t+1}, and of course (Π^s_{t+1}) holds true (since u_{t+1} minimizes $\omega_s(\cdot)$ on the entire $\mathbf{B}$).

In both cases, the one of $\mu > 0$ and the one of $\mu = 0$, we have built the data required to start step $t+1$ of phase s, and we proceed to this step.

The description of the algorithm is completed.

Two important remarks are in order.

Remark 3.2.2. The outlined algorithm requires solving at every step two nontrivial auxiliary optimization problems, specifically, (35) and (36). It is explained in [4] that these problems are relatively easy, provided that m is moderate (note that this parameter is under our full control) and $\mathbf{B}$ and ω are "simple and fit each other," meaning that we can easily solve problems of the form

$$\min_{x \in \mathbf{B}} \left[\omega(x) + \langle a, x\rangle\right]. \qquad (*)$$

For example, with Euclidean setup (see Example 3.2.1), problems $(*)$ are indeed easy, provided that $\mathbf{B}$ is a simple set (Euclidean ball or its intersection with the nonnegative orthant, box, standard simplex,...). In the case of ℓ_1 setup (see Example 3.2.2), problems $(*)$ are easy when $\mathbf{B}$ is the simplex $\{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i \leq R\}$. Note that the auxiliary problems arising in the presented algorithm are identical to those in the prototype NERML algorithm presented in [4], where one can find also a detailed discussion of the implementation issues.

Remark 3.2.3. By construction, the presented algorithm produces upon termination (if any)

- a *protocol* $\Pi_\tau = \{x_t, g(x_t)\}_{t=1}^{\tau}$, where τ is the step where the algorithm terminates, and x_t, $1 \leq t \leq \tau$, are the search points generated in course of the run; by construction, all these search points belong to $\mathbf{B}$;

- a collection of nonnegative weights $\lambda = (\lambda_1, \lambda_2, ..., \lambda_\tau)$ summing up to 1 such that the

output

$$h(x) = \sum_{t=1}^{\tau} \lambda_t \langle g(x_t), x_t - x \rangle$$

satisfies the relation

$$\max_{x \in \mathbf{B}} h(x) \leq \delta, \tag{41}$$

where δ is the input value of the target tolerance.

We shall see in a while that the above entities can be easily converted into an accuracy certificate.

3.2.4 NERML with Certificates: Analysis

We start with the following observation:

Proposition 3.2.1. *Given on input a target tolerance $\delta > 0$, the NERML algorithm terminates after finitely many steps, with the output described in Remark 3.2.3, and the number of steps of the algorithm does not exceed*

$$N = C(\gamma, \theta) \frac{\Omega}{\alpha \delta^2}, \text{ where } C(\gamma, \theta) = \frac{2(1+\gamma^2)}{\gamma^2[1 - [\gamma + (1-\gamma)\theta]^2]}. \tag{42}$$

Proof. Observe that the algorithm can terminate only according to 2.A, and in this case the output is indeed as claimed in Proposition. Thus, all we need to prove is the upper bound (42) on the number of steps before termination.

1^0. Let us bound from above the number of steps at an arbitrary phase s. Assume that phase s did not terminate in course of the first T steps, so that $u_1, ..., u_T$ are well defined. We claim that then

$$\|u_t - u_{t+1}\| \geq \ell_s, \ 1 \leq t < T. \tag{43}$$

Indeed, by construction $h^s_{t-1,m+1}(x) := \langle g(u_t), u_t - x \rangle$ is $\geq \ell_s = \gamma f_s$ when $x = u_{t+1}$ (due to $u_{t+1} \in Y_t$). Since $\|g(u)\|_* \leq 1$ for all $u \in \mathbf{B}$, (43) follows.

Now let us look at what happens with the quantities $\omega_s(u_t)$ as t grows. By strong convexity of ω (and thus ω_s) we have

$$\omega_s(u_{t+1}) - \omega_s(u_t) \geq \langle \omega_s'(u_t), u_{t+1} - u_t \rangle + \frac{\alpha}{2}\|u_t - u_{t+1}\|^2$$

The first term in the right hand side is ≥ 0, since u_t is the minimizer of $\omega_s(\cdot)$ over X^s_{t-1}, while $u_{t+1} \in Y_t \subset X^s_{t-1}$. The second term in the right hand side is $\geq \frac{\alpha}{2}\ell_s^2$ by (43). Since $\omega_s(u_1) = 0$ and $\omega_s(u_{t+1}) - \omega_s(u_t) \geq \frac{\alpha}{2}\ell_s^2$, we get

$$\omega_s(u_T) \geq (T-1)\frac{\alpha \ell_s^2}{2} = (T-1)\frac{\alpha \gamma^2 f_s^2}{2}.$$

Recalling the definition of Ω, the left hand side in this inequality is $\leq \Omega$. It follows that whenever phase s does not terminate in course of the first T steps, one has

$$T \leq \frac{2\Omega}{\alpha\gamma^2 f_s^2} + 1,$$

that is, the total number of steps at phase s, provided this phase exists, is at most

$$T_s = \frac{2\Omega}{\alpha\gamma^2 f_s^2} + 2.$$

Now, we have

$$f_s \leq f_1 = \max_{x\in\mathbf{B}} \langle g(x_0), x - x_0\rangle \leq \max_{x\in\mathbf{B}} \|x - x_0\|$$

(recall that $\|g(x)\|_* \leq 1$). Besides this,

$$\omega(x) \geq \omega(x_0) + \langle \omega'(x_0), x - x_0\rangle + \frac{\alpha}{2}\|x - x_0\|^2 \ \forall x \in \mathbf{B},$$

whence

$$\max_{x\in\mathbf{B}} \|x - x_0\| \leq \sqrt{2\Omega/\alpha}.$$

Putting things together, we get

$$f_s \leq f_1 \leq \sqrt{2\Omega/\alpha},$$

whence $\frac{2\Omega}{\alpha\gamma^2 f_s^2} \geq \gamma^{-2}$ and therefore

$$T_s = \frac{2\Omega}{\alpha\gamma^2 f_s^2} + 2 \leq \frac{2(1+2\gamma^2)}{\gamma^2}\frac{\Omega}{\alpha f_s^2}$$

for all s for which s-th phase exists. By construction, we have $f_s \geq \delta$ and $f_s \leq [\gamma + (1-\gamma)\theta] f_{s-1}$, whence, denoting by $\bar{s}$ the number of the phase where the method terminates, the total number of steps is bounded by

$$\begin{aligned}
&\textstyle\sum_{s=1}^{\bar{s}} \frac{2(1+2\gamma^2)}{\gamma^2}\frac{\Omega}{\alpha f_s^2} \leq \sum_{s=1}^{\bar{s}} \frac{2(1+2\gamma^2)}{\gamma^2}\frac{\Omega[\gamma+(1-\gamma)\theta]^{2(\bar{s}-s)}}{\alpha f_{\bar{s}}^2}\\
&\leq \textstyle\sum_{s=1}^{\bar{s}} \frac{2(1+2\gamma^2)}{\gamma^2}\frac{\Omega[\gamma+(1-\gamma)\theta]^{2(\bar{s}-s)}}{\alpha\delta^2}\\
&\leq \textstyle\frac{2(1+2\gamma^2)}{\gamma^2(1-[\gamma+(1-\gamma)\theta]^2)}\frac{\Omega}{\alpha\delta^2},
\end{aligned}$$

as claimed. □

We are about to explain how to utilize the above algorithm when solving problems with convex structure. To this end, we present two possible schemes, differing in the assumptions imposed on the problem to be solved.

3.3 Accuracy Certificates for Problems with Convex Structure via NERML, Case I: X is Simple

3.3.1 The Situation

Assume that we want to solve a problem with convex structure as defined in sections 2.1.1 – 2.1.4 in the situation when

A.1. We can take the domain X of the problem as the solid $\mathbf{B}$ participating in the setup of NERML.

In light of Remark 3.2.2, this assumption means that X is simple enough to be associated with an explicit distance generating function in a way which ensure that problems $(*)$ with $\mathbf{B} = X$ are easy to solve. This requirement is satisfied, for example, when X is an explicitly given ball, or box, and we use the Euclidean setup (Example 3.2.1), or when X is the standard simplex Δ_R, and we use the ℓ_1 setup. We can also "survive" the situations when X is cut off from the previously mentioned simple sets by a moderate number of explicitly given linear inequalities.

A.2. The domain of the monotone operator F associated with the problem of interest is the entire X, and associated First Order oracle "serves" the entire $\mathrm{Dom}\, F$, see Remarks 3.2.1 – 3.2.3. Moreover, we assume that there exists $c \in X = \mathbf{B}$, $\mathbf{V} < \infty$, and a $\Theta \in \mathbf{R} > 0$ such that

$$\forall (x \in \mathbf{B}, y : \|y\| \le 1) : \langle F(x), c + \Theta y - x \rangle \le \mathbf{V}. \tag{44}$$

Assumption **A.2** deserves some additional examination.

A.2.1. The simplest way to ensure **A.2** is to assume that F is bounded on its domain $\mathrm{Dom}\, F = X = \mathbf{B}$:

$$\|F\|_* := \sup_{x \in \mathrm{Dom}\, F} \|F(x)\|_* < \infty. \tag{45}$$

Denoting by $R(X)$ the radius of the smallest $\|\cdot\|$ ball containing X and choosing an arbitrary $c \in X$, we clearly have for every $\Theta > 0$:

$$\forall (y : \|y\| \le 1, x \in X) : \langle F(x), c + \Theta y - x \rangle \le \|F\|_* (2R(X) + \Theta),$$

that is, (44) holds no matter what the value is for $\Theta > 0$, provided that

$$\mathbf{V} = \mathbf{V}(\Theta) = \|F\|_* (2R(X) + \Theta). \tag{46}$$

A.2.2 In connection with **A.2.1** it should be noted that for a "functional" problem with convex structure (i.e., convex minimization problem (18), or convex-concave saddle

point problem (21), or convex Nash equilibrium problem 2.1.3), the corresponding monotone operator indeed can be chosen to be bounded, provided that the underlying function(s) are Lipschitz continuous. Specifically,

- In the case of a convex minimization problem (18), Lipschitz continuity, with a constant L, of the objective f, with respect to the norm $\|\cdot\|$, implies that at every point $x \in X$ f admits a subgradient $f'(x)$ with $\|f'(x)\|_* \leq L$, and further at all the points from int$X$ *a*ll the subgradients satisfy this bound, so that we ensure the validity of **A.1** and the relation $\|F\|_* \leq L$;

- In the case of a convex-concave saddle point problem, assuming that the norm $\|\cdot\|$ (which in this case is a norm on the direct product $\mathbf{R}^{n_1} \times \mathbf{R}^{n_2}$), satisfies the relation

  $$\|[x_1; x_2]\| \geq \max[\|[x_1; 0]\|, \|[0, x_2]\|] \; \forall x_1 \in \mathbf{R}^{n_1}, x_2 \in \mathbf{R}^{n_2},$$

  we have that Lipschitz continuity, with a constant L, of the cost function $f(x_1, x_2)$ with respect to $\|\cdot\|$ implies that the associated monotone operator can be chosen to satisfy $\|F\|_* \leq 2L$, and, moreover, whatever be the choice of partial sub-and supergradients $f_1'(x)$, $f_2'(x)$, one always has $\|F(x)\|_* \leq 2L$ when $x \in \text{int}X$;

- Similarly, in the case of a convex Nash equilibrium problem, assuming that the norm $\|\cdot\|$ (which in this case is a norm on $\mathbf{R}^{n_1} \times ... \times \mathbf{R}^{n_m}$) satisfies the relation

  $$\forall (x_i \in \mathbf{R}^{n_i},\, 1 \leq i \leq m):$$

  $$\|[x_1; ...; x_m]\| \geq \max\left[\|[x_1; 0; ...; 0]\|, \|[0, x_2; 0; ...; 0]\|, ..., \|[0; ...; 0; x_m]\|\right],$$

  we have that the Lipschitz continuity, with a constant L, of every one of the functions $f_i(x)$, $1 \leq i \leq m$, implies that the associated monotone operator can be chosen to satisfy the relation $\|F\|_* \leq mL$, and for $x \in \text{int}X$, one has $\|F(x)\|_* \leq mL$ whatever the choice of the associated partial subgradients $f_i'(x)$.

3.3.2 Construction and Result

Now we are ready to formulate out first major result on NERML.

Proposition 3.3.1. *Let assumptions* **A.1-2** *take place. Given $\epsilon > 0$, let us set*

$$\delta = \frac{\Theta\epsilon}{\mathbf{V} + \epsilon}.$$

Consider the vector field $g(x) : X = \mathbf{B} \to \mathbf{R}^n$ given by

$$g(x) = \begin{cases} 0, & F(x) = 0 \\ \frac{1}{\|F(x)\|_*} F(x), & F(x) \neq 0 \end{cases}$$

and let us apply to this vector field the NERML algorithm, the target tolerance being δ. Let $\Pi_\tau = \{x_t, g(x_t)\}_{t=1}^\tau$, λ, $h(\cdot)$ be the resulting protocol and output (see Remark 3.2.3), so that $h(x) = \sum_{t=1}^\tau \lambda_t \langle g(x_t), x_t - x\rangle$ satisfies

$$\max_{x\in\mathbf{B}} h(x) \le \delta, \tag{47}$$

while $\lambda \ge 0$ and $\sum_{t=1}^\tau \lambda_t = 1$.

In the case when there exists $t_* \in \{1, ..., \tau\}$ such that $g(x_{t_*}) = 0$, let us set $\xi_{t_*} = 1$ and $\xi_t = 0$ for all remaining $t \in \{1, ..., \tau\}$. When $g(x_t) \ne 0$ for all $t \in \{1, ..., \tau\}$, let us set

$$\xi_t = \frac{\lambda_t/\|F(x_t)\|_*}{\sum_{s=1}^\tau \lambda_s/\|F(x_s)\|_*}.$$

Then $\xi \ge 0$, $\sum_{t=1}^\tau \xi_t = 1$ and

$$\max_{x\in\mathbf{B}} \sum_{t=1}^\tau \xi_t \langle F(x_t), x_t - x\rangle \le \epsilon. \tag{48}$$

Besides this,

$$\tau \le C(\gamma,\theta)\frac{\Omega}{\alpha\delta^2} = C(\gamma,\theta)\frac{\Omega(\mathbf{V}+\epsilon)^2}{\alpha\Theta^2\epsilon^2}. \tag{49}$$

Proof. The complexity bound (49) is readily given by Proposition 3.2.1. The facts that $\xi_t \ge 0$, $\sum_t \xi_t = 1$ are evident. Relation (48) is evident when $g(x_t) = 0$ for some $t \in \{1, ..., \tau\}$. Thus, it remains to demonstrate that (48) holds true when $g(x_t) \ne 0$ for all $t \le \tau$. Let us set $h_t = F(x_t)$, $g_t = \|h_t\|_*$. Relation (47) reads

$$\forall x \in \mathbf{B} : \sum_{t=1}^\tau \frac{\lambda_t}{g_t}\langle h_t, x_t - x\rangle \le \delta. \tag{50}$$

Specifying in (44) x as x_t, we have

$$\langle h_t, c - x_t\rangle + \Theta\|h_t\|_* = \max_{y:\|y\|\le 1} \langle h_t, c - x_t + \Theta y\rangle \le \mathbf{V},$$

whence

$$\langle h_t, x_t - c\rangle \ge \Theta g_t - \mathbf{V}.$$

Multiplying these inequalities by $\frac{\lambda_t}{g_t}$ and summing up over $t = 1, ..., \tau$, we get

$$\sum_{t=1}^\tau \frac{\lambda_t}{g_t}\langle h_t, x_t - c\rangle \ge \Theta \underbrace{\sum_{t=1}^\tau \lambda_t}_{=1} - \left(\sum_{t=1}^\tau \lambda_t/g_t\right)\mathbf{V}.$$

The left hand side in this inequality is $\le \delta$ by (50) due to $c \in \mathbf{B}$. It follows that

$$\sum_{t=1}^\tau \lambda_t/g_t \ge \frac{\Theta - \delta}{\mathbf{V}} = \frac{1}{\mathbf{V}}\left[\Theta - \frac{\Theta\epsilon}{\mathbf{V}+\epsilon}\right] = \frac{\Theta}{\mathbf{V}+\epsilon} = \frac{\delta}{\epsilon}.$$

Dividing both sides in (50) by $\sum_{t=1}^{\tau}\lambda_t/g_t$, we get

$$\max_{x\in\mathbf{B}}\sum_{t=1}^{\tau}\xi_t\langle F(x_t),x_t-x\rangle\le\frac{\delta}{\sum_{t=1}^{\tau}\lambda_t/g_t}\le\epsilon,$$

as claimed in (48). ∎

3.3.3 Comments

1) In the situation of Proposition 3.3.1, the protocol $\Pi_\tau=\{x_t,g(x_t)\}_{t=1}^{\tau}$ reported by NERML can be straightforwardly converted into a "fully productive" (no non-productive steps) execution protocol $P_\tau=(\{x_t,F(x_t)\}_{t=1}^{\tau},I_\tau,J_\tau)$ for $(X=\mathbf{B},F)$, where $I_\tau=\{1,...,\tau\}$ and $J_\tau=\emptyset$. The Proposition further states that the NERML output readily provides an accuracy certificate ξ for this execution protocol, and that for this certificate one has $\varepsilon_{\rm cert}(\xi|P_\tau,\mathbf{B})\le\epsilon$. In other words, we have indeed equipped NERML with accuracy certificates.

Now let us examine the complexity of building an accuracy certificate with resolution ϵ. Assume that the monotone operator F in question is bounded. As we have seen in **A.2.1**, in this case the assumption **A.2** is satisfied for every Θ with $\mathbf{V}=\mathbf{V}(\Theta)=\|F\|_*(2R(X)+\Theta)$. Setting $\Theta=R(X)$ and assuming $\epsilon\le\|F\|_*R(X)$, we get from (49) the bound

$$\tau\le\bar{C}(\gamma,\theta)\frac{\Omega\|F\|_*^2}{\alpha\epsilon^2},\tag{51}$$

which is exactly the complexity estimate, as stated in [4], for the prototype NERML algorithm (i.e the NERML algorithm which does *not* produce accuracy certificates). Thus, as it was done in [27] for the case of Cutting Plane algorithms, we have succeeded in equipping the prototype algorithm with a computationally cheap technique for producing accuracy certificates which justifies the theoretical complexity of the prototype.

2) To illustrate the complexity results we have obtained, consider the situation when we use the ℓ_1 setup, (see Example 3.2.2). Here the complexity bound (51) (where we treat γ, and θ as absolute constants) reads

$$\tau\le O(1)\frac{\ln(n)R^2\|F\|_*^2}{\epsilon^2},$$

where R is the size of the standard simplex Δ_R containing $X=\mathbf{B}$, and $\|F\|_*=\sup\limits_{x\in X}\|F(x)\|_\infty$. Note that this complexity bound is nearly dimension-independent and, moreover, nearly optimal in the large scale case (see [4] and references therein). Similarly, in the case of the Euclidean setup (see Example 3.2.1), the complexity bound (51) becomes a fully dimension-independent (and optimal in the large-scale case) bound

$$\tau\le O(1)\frac{R^2\|F\|_2^2}{\epsilon},$$

where R is the radius of the smallest Euclidean ball containing $X = \mathbf{B}$, and $\|F\|_2 = \sup_{x \in X} \|F(x)\|_2$.

3.4 Accuracy Certificates for Problems with Convex Structure via NERML, Case II: X is given by a Separation Oracle

From a practical viewpoint, the drawback of the situation considered in section 3.3 lies in the restrictive nature of assumption **A.1** which "in reality" means that X must be really simple (otherwise the auxiliary problems arising in NERML can become too difficult). We are about to demonstrate that we can also handle the case of a "complicated" X, e.g., an X given by a Separation oracle, provided that X can be reasonably well approximated by a $\|\cdot\|$ ball.

3.4.1 Preliminaries: Semi-bounded Operators

Let $X \subset \mathbf{R}^n$ be a solid, and $F : \mathrm{Dom}\, F \to \mathbf{R}^n$ be an operator with $\mathrm{Dom}\, F \supseteq \mathrm{int} X$. Then F is said to be *bounded* on X, if the quantity $\|F\|_2 = \sup\{\|F(x)\|_2 : x \in \mathrm{int} X\}$ is finite. While F is said to be *semi-bounded* on X, if the quantity

$$\mathrm{Var}_X(F) = \sup\{\langle F(x), y - x\rangle : x \in \mathrm{int} X, y \in X\}$$

is finite. Clearly a bounded operator F is semi-bounded with $\mathrm{Var}_X(F) \le \|F\|_2 \mathrm{Diam}_2(X)$, where Diam_2 is the Euclidean diameter of X. There exists, however, semi-bounded operators which are not bounded. E.g., the monotone operator F associated with a convex minimization problem (18) with a *bounded* objective f is clearly semi-bounded: since for $x \in \mathrm{Dom}\, F$ $F(x) \in \partial_X f(x)$, we have $\langle F(x), y - x\rangle \le f(y) - f(x)$ for all $x \in \mathrm{Dom}\, F$, $y \in X$, whence

$$\mathrm{Var}_X(F) \le \sup_X f - \min_X f < \infty.$$

Moreover, the monotone operator associated with (18) can be semi-bounded for certain *unbounded* objectives f, most notably, when f is a *ϑ-self-concordant barrier for X* [30, Chapter 2]; for such a barrier, $\mathrm{Var}_X(F) \le \vartheta$ [30, Proposition 2.3.2].

For similar reasons, the monotone operator associated with a convex Nash equilibrium problem (section 2.1.3) (in particular, with a convex-concave saddle point problem, section 2.1.2) is always semi-bounded:

$$\mathrm{Var}_X(F) \le \sum_{i=1}^m [\max_X f_i(x) - \min_X f_i(x)].$$

3.4.2 Situation and Goal

The situation we intend to consider is as follows: We are given a NERML setup, i.,e., a norm $\|\cdot\|$ on $\mathbf{R}^n$, a solid $\mathbf{B} \subset \mathbf{R}^n$ and a distance generating function $\omega(\cdot)$ for this solid. We assume here that $\mathbf{B}$ is *contained within* a $\|\cdot\|$-ball of a known radius $R > 0$ centered at the origin. Further, we assume that there exists a solid X, $X \subset \mathbf{B}$, represented by a Separation oracle and known *to contain* a $\|\cdot\|$-ball of a given radius $r > 0$. Finally, we assume that we are given an operator $F : \mathrm{Dom}\, F \to \mathbf{R}^n$ with convex domain such that $\mathrm{int}X \subset \mathrm{Dom}\, F \subset X$, such that F is semi-bounded on X, and we are given an upper bound $\mathbf{V} < \infty$ on the corresponding variation:

$$\sup\{\langle F(x), y - x\rangle : x \in \mathrm{int}X, y \in X\} \leq \mathbf{V}. \tag{52}$$

Finally, we assume that we are given access to a First Order oracle which, given on input $x \in \mathrm{int}X$, returns $F(x)$.

Our goal is generate, given $\epsilon > 0$, an execution protocol P_τ for (X, F) with an accuracy certificate for this protocol with resolution $\varepsilon_{\mathrm{cert}} \leq \epsilon$, that is, we desire to generate a sequence of points $x_1, ..., x_\tau$ and nonnegative reals ξ_t, $1 \leq t \leq \tau$, such that

$$\sum_{t: x_t \in \mathrm{int}X} \xi_t = 1 \quad \& \quad \max_{x \in \mathbf{B}} \sum_{t=1}^{\tau} \xi_t \langle e_t, x_t - x\rangle \leq \epsilon, \tag{53}$$

where $e_t = F(x_t)$ when $x_t \in \mathrm{int}X$ and e_t separates x_t and X:

$$e_t \neq 0,\ \langle e_t, x_t - x\rangle \geq 0\ \forall x \in X$$

when $x_t \notin \mathrm{int}X$.

3.4.3 The Construction

In order to achieve our goal, we

1. Define a vector field $g(x) : \mathbf{B} \to \mathbf{R}^n$ by the relation

$$g(x) = \begin{cases} F(x)/\|F(x)\|_*, & x \in \mathrm{int}X, F(x) \neq 0 \\ 0, & x \in \mathrm{int}X, F(x) = 0 \\ e_x/\|e_x\|_*, & x \notin \mathrm{int}X \end{cases},$$

   where for $x \notin \mathrm{int}X$ the linear form given by e_x separates x and X and is nonzero.

   Note that the Separation oracle for X and the First Order oracle for F we have at our disposal allow us to compute $g(x)$ for every $x \in \mathbf{B}$.

2. Apply to the vector field $g(\cdot)$ the NERML algorithm as described in section 3.2.2, the target tolerance being
$$\delta = \frac{\epsilon r^2}{\mathbf{V}(r+2R)+\epsilon(2r+2R)}.$$
Applying Proposition 3.2.1, in
$$\tau \le C(\gamma,\theta)\frac{\Omega}{\alpha\delta^2} \le 16C(\gamma,\theta)\frac{\Omega R^2(\mathbf{V}+\epsilon)^2}{\alpha r^4\epsilon^2} \tag{54}$$
steps we build a set $S=\{x_1,...,x_\tau\}\subset\mathbf{B}$ and a function $h(x)=\sum\limits_t \lambda_t\langle g(x_t),x_t-x\rangle$ with $\lambda_t\ge 0,\sum_t\lambda_t=1$, such that
$$\max_{x\in\mathbf{B}} h(x)\le\delta.$$
Note that as a byproduct of generating S, $h(\cdot)$, and λ, we have at our disposal the sets $I_t=\{t\le\tau: x_t\in \mathrm{int}X\}$, $J_t=\{t\le\tau: x_t\notin\mathrm{int}X\}$ and vectors e_t such that $e_t=F(x_t)$ when $t\in I_\tau$ and $e_t=g(x_t)$ separates x_t and X when $t\in J_\tau$. Thus, we have at our disposal an execution protocol $P_\tau=(\{x_t,e_t\}_{t=1}^\tau, I_\tau, J_\tau)$ for (X,F).

3. Note that two things may happen
Case A: It may happen that there exists $t_*\le\tau$ such that $g(x_t)=0$. In this case, we set $\xi_{t_*}=1$, $\xi_t=0$, $t\ne t_*$.
Case B: $g(x_t)\ne 0$ for all $t\le\tau$. In this case, we set $\mu=\sum_{t\in I}\lambda_t/\|F(x_t)\|_*$, and
$$\xi_t=\begin{cases}\frac{\lambda_t}{\mu}, & x_t\notin\mathrm{int}X\\ \frac{\lambda_t}{\|F(x_t)\|_*\mu}, & x_t\in\mathrm{int}X.\end{cases}$$

3.4.4 The Result

Our main result here is as follows:
Proposition 3.4.1. *If B is the case, the above procedure is well defined (that is, $I_\tau\ne\emptyset$ and $\mu>0$). In both cases A,B we have $\xi_t\ge 0$, $\sum_{t\in I_\tau}\xi_t=1$, and*
$$\varepsilon_{\mathrm{cert}}(\xi|P_\tau,\mathbf{B})\le\epsilon. \tag{55}$$

Proof. Since $g(x)\ne 0$ when $x\notin\mathrm{int}X$, in the case of A we have $x_{t_*}\in\mathrm{int}X$, and all required facts follow. Now assume that B is the case.

1^0. Let $P=\sum_{t\in I_\tau}\lambda_t$, and let c be such that the $\|\cdot\|$ ball of radius r centered at c is contained in X. When $t\notin I_\tau$, we have
$$\forall(y,\|y\|\le 1):\langle g(x_t),x_t-c-ry\rangle\ge 0\,\&\,\|g(x_t)\|_*=1,$$

whence

$$\langle g(x_t), x_t - c\rangle \geq r.$$

When $t \in I_\tau$, we have $\|x_t - c\| \leq 2R$ (since $x_t, c \in X \subset \mathbf{B}$ and $\mathbf{B}$ is contained in the $\|\cdot\|$ ball of radius R) and $\|g(x_t)\|_* = 1$, whence

$$\langle g(x_t), x_t - c\rangle \geq -2R.$$

We therefore have

$$\delta \geq \sum_t \lambda_t \langle g(x_t), x_t - c\rangle \geq (1-P)r - 2PR,$$

whence

$$P(r+2R) \geq r - \delta$$

and therefore

$$P \geq \frac{r-\delta}{r+2R}.$$

Since $\delta < r$, we have $P > 0$. Thus, $I_\tau \neq \emptyset$ and $\mu > 0$. It follows that $\xi_t \geq 0$ are well defined and $\sum_{t\in I_\tau} \xi_t = 1$.

2^0. For $t \in I_\tau$, let us set $g_t = \|F(x_t)\|_*$. When $x \in X$, we have $\sum_{t\in J_\tau} \lambda_t \langle g(x_t), x_t - x\rangle \geq 0$, whence

$$\forall x \in X : \sum_{t\in I_\tau} \lambda_t \langle g(x_t), x_t - x\rangle = \sum_{t\in I_\tau} \frac{\lambda_t}{g_t} \langle F(x_t), x_t - x\rangle \leq \delta.$$

Now let $t \in I_\tau$, and let v_t, $\|v_t\| = 1$, be such that $\langle F(x_t), v_t\rangle = g_t$. Since the vector $c + rv_t$ belongs to X, we have

$$\langle F(x_t), c + rv_t - x_t\rangle \leq \mathbf{V},$$

whence

$$\langle F(x_t), x_t - c\rangle \geq rg_t - \mathbf{V}.$$

Multiplying by $\frac{\lambda_t}{g_t}$ and summing up over $t \in I_\tau$, we get

$$\delta \geq \sum_{t\in I_\tau} \frac{\lambda_t}{g_t} \langle F(x_t), x_t - c\rangle \geq rP - \mu\mathbf{V},$$

whence

$$\mu \geq \frac{rP-\delta}{\mathbf{V}} \geq \frac{r\frac{r-\delta}{r+2R} - \delta}{\mathbf{V}} = \frac{\delta}{\epsilon}.$$

It follows that

$$\forall (x \in \mathbf{B}) :$$

$$\left[\sum_{t\in I_\tau} \xi_t \langle F(x_t), x_t - x\rangle + \sum_{t\in J_\tau} \xi_t \langle g(x_t), x_t - x\rangle\right] = \frac{1}{\mu} \sum_t \lambda_t \langle g(x_t), x_t - x\rangle \leq \epsilon,$$

that is, ξ is an accuracy certificate with resolution $\leq \epsilon$. ∎

CHAPTER IV

ACCURACY CERTIFICATES: ACADEMIC APPLICATIONS

In this chapter, we present several novel academic applications of the accuracy certificates.

4.1 Certifying Emptiness of the Intersection of Solids

The problem. Let $X_1, ..., X_m$ be convex solids known to belong to the centered at the origin Euclidean ball V_R of a given radius R in $\mathbf{R}^n$. We assume that these solids are given by Separation oracles. It is trivial to provide a certificate proving that $\cap X_i \neq \emptyset$, indeed to certify this it suffices to provide an $x \in \cap X_i$; the validity of such a certificate can be easily verified by calling the Separation oracles representing X_i. However it is not immediately obvious how one could certify the opposite, namely, that $\cap X_i = \emptyset$. We propose a simple certificate for this fact, namely, as follows:

Proposition 4.1.1. *Let $X = X_1 \times ... \times X_m \subset V_R^m = V_R \times ... \times V_R \subset \mathbf{R}^{mn} = \mathbf{R}^n \times ... \times \mathbf{R}^n$. Assume we can point out a finite collection of points $w_s \in \mathbf{R}^{mn}$ such that $w_s \notin \mathrm{int} X$ along with vectors $\eta_s = [\eta_s^1; ...; \eta_s^m] \in \mathbf{R}^{mn}$ and nonnegative weights ζ_s, $1 \leq s \leq S$, such that η_s separates w_s and X:*

$$\langle \eta_s, w_s \rangle \geq \sup_{w \in X} \langle \eta_s, w \rangle,$$

and the linear inequality

$$\begin{gathered} \langle P\left[\sum_{s=1}^{S} \zeta_s \eta_s\right], y \rangle \leq \sum_{s=1}^{S} \zeta_s \langle \eta_s, w_s \rangle \\ \left[P[x^1; ...; x^m] = \sum_{i=1}^{m} x^i : \mathbf{R}^{mn} \to \mathbf{R}^n\right] \end{gathered} \tag{56}$$

in variable $y \in \mathbf{R}^n$ has no solutions in the ball V_R. Then $\cap_i X_i = \emptyset$.

Proof. Assuming, on the contrary to what should be proved, that $\cap_i X_i \neq \emptyset$, we can choose $y \in \cap_i X_i$ and set $x = [y; ...; y]$, so that $x \in X$. Note that $y \in V_R$ due to $X_i \subset V_R$, $i = 1, ..., m$. We have $\langle \sum_i \eta_s^i, y \rangle = \langle \eta_s, x \rangle \leq \langle \eta_s, w_s \rangle$ for all s. It follows that $y \in V_R$ is a solution to the linear inequality in (56), which is a desired contradiction. ∎

Proposition 4.1.1 can be interpreted as follows: whenever we can specify the data S,

$\{w_s, \eta_s, \zeta_s\}_{s=1}^S$ participating in the premise of the proposition in such a way that the inequality (56) has no solutions in V_R, we can treat these data as a certificate for the emptiness of $\cap_i X_i$. Note that given the data, it is easy to verify whether (56) has or has no solutions in V_R: it has no solutions in this ball iff

$$-R\|P\left[\sum\nolimits_{s=1}^S \zeta_s \eta_s\right]\|_2 > \sum_{s=1}^S \zeta_s \langle \eta_s, w_s \rangle. \tag{57}$$

While this proposition shows that the existence of such a certificate is a *sufficient* condition for $\cap_i X_i = \emptyset$, it does not show if it is also necessary for the intersection of X_i to be empty. Similarly the proposition does not show how to build such a certificate if one does exist. We answer these questions now.

Let us define a convex function $f : X \mapsto \mathbf{R}$ as

$$f(x) = f(x^1, ..., x^m) = \frac{1}{2} \sum_{i=1}^m \|x^i - x^{i+1}\|_2^2,$$

where $x^{m+1} \equiv x^1$. Now let us consider the optimization problem

$$\mathrm{Opt} = \min_{x \in X} f(x) \tag{58}$$

By defining our problem in this manner we ensure that the sets X_i have no point in common iff $\mathrm{Opt} > 0$.

Observe that the Separation oracles for X_i induce a Separation oracle for X, and there is no problem with equipping f with a First Order oracle which serves X. Thus, we can solve the convex minimization problem (58) by an algorithm $\mathcal{B}$ with certificates, e.g., by the Ellipsoid method with certificates (Proposition 2.2.1) or by NERML with certificates (chapter 3). Assume that after τ steps the algorithm produces an execution protocol $P_\tau = (\{x_t, e_t\}_{t=1}^\tau, I_\tau, J_\tau)$ (where $x_t \in \mathrm{int}X$ and $e_t = f'(x_t)$ when $t \in I_\tau$, and $x_t \notin \mathrm{int}X$ and $e_t \neq 0$ separates x_t and X when $t \in J_\tau$) and an associated certificate ξ^τ .

It may happen that $J_\tau \neq \emptyset$; in this case, we denote by S the number of elements in J_τ and by $t(s)$, $1 \leq s \leq S$, the s-th of these elements. Then the data

$$\omega^S = \left\{w_s = x_{t(s)}, \eta_s = e_{t(s)}, \zeta_s = \xi^\tau_{t(s)}\right\}_{s=1}^S$$

can be treated as a candidate emptiness certificate as defined in Proposition 4.1.1.

Proposition 4.1.2. *Let τ be such that ξ^τ is well defined, and let*

$$\delta_\tau = 2 \sum_{t \in I_\tau} \xi^\tau_t f(x_t) - \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, V_R^m).$$

Whenever $\delta_\tau > 0$, we have $J_\tau \neq \emptyset$, so that ω^S is well defined, and ω^S is a valid emptiness certificate for $\cap_i X_i$.

Proof. Let $P_\tau = \{(x_t, e_t)\}_{t=1}^{\tau}$ be an execution protocol associated with (58) and satisfying the premise of Proposition, and let ξ be a certificate for P_τ. In order to make the upcoming calculations more transparent we set

$$\begin{aligned} a_i &= \sum_{t\in J_\tau} \xi_t \langle e_t^i, x_t^i\rangle, \\ f^i &= \sum_{t\in J_\tau} \xi_t e_t^i, \\ z^i &= \sum_{t\in I_\tau} \xi_t [2x_t^i - x_t^{i+1} - x_t^{i-1}], \end{aligned}$$

where, $x^0 := x^m$ and, same as above, $x^{m+1} := x^1$. Taking into account that f is homogeneous of degree 2, so that $\langle f'(x), x\rangle = 2f(x)$, and that $[f'(x^1, ..., x^m)]^i = 2x^i - x^{i+1} - x^{i-1}$, we have

$$\begin{aligned} & \varepsilon_{\rm cert}(\xi|P_\tau, V_R^m) \\ = & \max_{x\in V_R^m} \left[\sum_{t\in I_\tau} \xi_t \langle f'(x_t), x_t - x\rangle + \sum_{t\in J_\tau} \xi_t \sum_{i=1}^m \langle e_t^i, x_t^i - x^i\rangle\right] \\ = & \max_{x\in V_R^m} \left[\sum_{t\in I_\tau} \left[2\xi_t f(x_t) - \xi_t \sum_{i=1}^m \langle 2x_t^i - x_t^{i+1} - x_t^{i-1}, x^i\rangle\right] + \sum_{i=1}^m [a_i - \langle f^i, x^i\rangle]\right] \\ = & 2\sum_{t\in I_\tau} \xi_t f(x_t) + \sum_{i=1}^m a_i + R\sum_{i=1}^m \|z^i + f^i\|_2, \end{aligned}$$

where the last equality follows from the fact that V_R is the centered at the origin Euclidean ball of the radius R. Rearranging terms we have that

$$\sum_{i=1}^m a_i + R\sum_{i=1}^m \|z^i + f^i\|_2 = \varepsilon_{\rm cert}(\xi|P_\tau, V_R^m) - 2\sum_{t\in I_\tau} \xi_t f(x_t)$$

Now since $\sum_i z^i = 0$, and by the Triangle inequality we have

$$\sum_{i=1}^m \|z^i + f^i\|_2 \geq \|\sum_{i=1}^m (z^i + f^i)\|_2 = \|\sum_i f^i\|_2,$$

whence

$$R\|\sum_{i=1}^m f^i\|_2 \leq -\sum_i a_i + \underbrace{\varepsilon_{\rm cert}(\xi|P_\tau, V_R^m) - 2\sum_{t\in I_\tau} \xi_t f(x_t)}_{-\delta_\tau} \tag{59}$$

Now (59), and our assumption that $\delta_\tau \geq 0$ along with our definitions of $f^i = \sum_{t\in J_\tau} \xi_t e_t^i$ and $a_i = \sum_{t\in J_\tau} \xi_t \langle e_t^i, x_t^i\rangle$ immediately implies that $J_\tau \neq \emptyset$, and

$$-R\|\sum_i f^i\|_2 > \sum_{i=1}^m a_i. \tag{60}$$

While $J_\tau \neq \emptyset$ and (57) together proves that ω^S is well defined, and that ω^S is a valid emptiness certificate for $\cap_i X_i$. ∎

Note that if the algorithm with certificates $\mathcal{B}$ converges, meaning that

$$\varepsilon_{\text{cert}}(\xi^\tau | P_\tau, V_R^m) \to 0, \ \tau \to +\infty$$

(as is the case, e.g., for the Ellipsoid and NERML algorithms with certificates), and $\cap_i X_i = \emptyset$, so that Opt > 0, we will eventually have $\delta_\tau > 0$. Indeed, since $\xi_t^\tau \geq 0$, and $\sum_{t \in I_\tau} \xi_t^\tau = 1$ we have that $\sum_{t \in I_\tau} \xi_t^\tau f(x_t) \geq$ Opt, and thus $\delta_\tau = 2 \sum\limits_{t \in I_\tau} \xi_t^\tau f(x_t) - \varepsilon_{\text{cert}}(\xi^\tau | P_\tau, V_R^m) > 0$ for all large enough values of τ (namely, those for which $\varepsilon_{\text{cert}}(\xi^\tau | P_\tau, V_R^m) <$ 2Opt. When it happens for the first time that $\delta_\tau > 0$, we get an emptiness certificate for $\cap_i X_i$. We see, in particular, that the existence of an emptiness certificate is both a *necessary* and sufficient condition for $\cap_i X_i = \emptyset$. How long it takes to get such a certificate, depends on the rate of convergence of $\varepsilon_{\text{cert}}(\xi^\tau | P_\tau, V_R^m)$ to 0 as $\tau \to \infty$, and on the actual value of Opt > 0 (the "measure of closeness" of the collection $\{X_i\}$ with empty intersection to a collection with a nonempty intersection). For example, when we use the Ellipsoid algorithm with certificates and Opt > 0, the number of steps till an emptiness certificate is built does not exceed the quantity $O(1)(mn)^2 \ln\left(\frac{nR^3}{r\text{Opt}}\right)$, where $r = \min\limits_i r(X_i)$, see Proposition 2.2.1.

4.2 Minimizing Convex Function over a Solid given by a Linear Optimization Oracle

In the standard setting of a black box represented convex minimization problem, the feasible domain X of the problem is given via a Separation oracle. Another natural description of a solid $S \subset \mathbf{R}^n$ is given via a *Maximization oracle* (also known as a Linear Optimization (LO) oracle). Such an oracle, given an input vector $e \in \mathbf{R}^n$, returns a point from $\text{Argmax}_{x \in S} \langle e, x \rangle$. It is known [12] that a Separation oracle can be mimicked in polynomial time via an Optimization oracle, and vice versa, so that given an algorithm which works with a separation-type representation of X, we could convert it into an algorithm which works with a maximization-type representation of X, and vice versa. This conversion, while it preserves polynomiality of the running time of the algorithms, replaces a single step of the original algorithm with solving a nontrivial problem, which usually makes the resulting algorithm of purely academic interest with no practical value. Our goal is to develop an alternative way to solve convex minimization problems, defined over solids represented by Maximization oracles. Our approach will be based on *Fenchel-type representations* of convex functions.

4.2.1 The Problem

We readdress the problem (18), specifically, the convex minimization problem

$$\min_{x \in X} f(x) \tag{61}$$

on a solid $X \subset \mathbf{R}^n$. However unlike our prior treatment, where X was a convex solid equipped with a separation oracle and f was represented by a First Order oracle, we now assume that X is a solid such that it is easy to maximize linear functions over X (i.e., we assume X is mated to a Linear Optimization oracle). As for f, we still assume that this is a continuous on X convex function, however we now assume that f is given by a *Fenchel-type representation*

$$f(x) = \sup_y \left[\langle x, Ay + a \rangle - h(y) \right], \tag{62}$$

where

- $y \mapsto Ay + a : \mathbf{R}^m \to \mathbf{R}^n$ is a given affine mapping, and
- $h(y) : \mathbf{R}^m \to \mathbf{R} \cup \{+\infty\}$ is a lower semicontinuous convex function,

The assumptions on X and f we have just made seem to be rather "esoteric." Indeed, usually we deal with a convex objective f given by a "closed form analytical expression" which makes it relatively easy to equip f with a computationally efficient First Order oracle, but which does not straightforwardly yield a Fenchel-type representation of f. Similarly, the feasible domains in convex minimization programs are usually given by systems of constraints

$$f_i(x) \leq 0,\ i = 1, ..., m,$$

where the f_i are continuous convex functions given by "closed form analytical expressions" allowing us to build computationally efficient First Order oracles for f_i. These oracles, in turn, can be easily combined into a Separation oracle for X, as follows. Assuming that the system of constraints in question is strictly feasible (i.e that there exists $\bar{x}$ such that $f_i(\bar{x}) < 0$ for all i), and given an $x \in \mathbf{R}^n$, we compute the values and subgradients of f_i at x. If all the values are negative, we conclude that $x \in \mathrm{int}X$. While if all the values are non-positive, but some are equal to 0, we conclude that $x \in \partial X$. Finally, if a certain constraint is violated at x (i.e some value is strictly positive), we conclude that $x \notin X$. Now if $x \notin \mathrm{int}X$, then there exists an i_* such that $f_{i_*}(x) \geq 0$. Let e be a subgradient of this function at x. Then $f_{i_*}(y) \geq f_{i_*}(x) + \langle e, y - x \rangle \geq \langle e, y - x \rangle$ (note that $f_{i_*}(x) \geq 0$). When $y \in X$, we have $f_{i_*}(y) \leq 0$, that is, $\langle e, y - x \rangle \leq 0$; this means that e separates x and X. It remains to check that $e \neq 0$, which is immediate: if $e = 0$, then the inequalities $f_{i_*}(y) \geq \langle e, y - x \rangle + f_{i_*}(x)$, $f_{i_*}(x) \geq 0$ say that $f_{i_*}(y) \geq 0$ everywhere, which contradicts the assumed strict feasibility of the system. Thus, *typically* feasible domains of convex minimization problems do admit efficient Separation oracles, which seemingly is not the case for LO oracles.

To motivate our interest in the outlined setting of a convex minimization problem, we intend to demonstrate, first, that the availability of explicit Fenchel-type representations of convex functions is more of a rule rather than an exception, and, second, that there are meaningful situations when convex solids admit efficient LO oracles which are better suited for our ultimate purpose — solving (61) numerically – compared to Separation oracles.

4.2.2 Fenchel-type Representations of Convex Functions.

A Fenchel-type representation a convex function f seems to be much less intuitive than the First Order representation of f, and one might think that the availability of such a representation is a "rare commodity." We will explain that this is not the case, and show that providing a Fenchel-type representation is no more difficult than providing a First Order representation.

Note that every function $f(x)$ which admits a Fenchel-type representation must be convex and lower semicontinuous. If the representation in question is a *Fenchel* one, meaning that $Ay + a \equiv y$, then f must also be proper, (i.e., have a nonempty domain). The *Fenchel Duality Theorem* (see, e.g., [35]) says that these conditions are not only necessary, but also sufficient for a function f to admit a Fenchel representation. Specifically, this theorem states that every proper lower semicontinuous convex function f admits a Fenchel representation

$$f(x) = \sup_y \{\langle x, y\rangle - f_*(y)\}$$

where $f_*(y)$ is the Fenchel dual of f:

$$f_*(y) = \sup_x \{\langle x, y\rangle - f(x)\}. \tag{63}$$

Moreover, there exists exactly one proper and lower semicontinuous function — the Fenchel dual of f as given by (63) — which recovers f via the Fenchel representation. It follows, then that Fenchel duality is symmetric: whenever f is a proper lower semicontinuous convex function, so is its Fenchel dual f_*, and further $(f_*)_* = f$.

We see that the assumption of the existence of a Fenchel-type representation of a convex function is not really an assumption at all. In order for it to hold we impose on a convex function f only the extremely mild restrictions of properness and lower semicontinuity — exactly the restrictions which make the problem of minimizing f well posed. From the computational viewpoint the existence of Fenchel-type representations is however insufficient. In order to use a Fenchel-type representation in a computational algorithm, like the one we intend to develop, we need an "explicit" representation of this type, one where we can equip $\mathrm{cl}\,(\mathrm{Dom}\, h)$ with a Separation oracle, and where we can equip the restriction of h onto the relative interior of its domain with a First Order oracle. From this perspective, Fenchel representations are "bad". Specifically we note that for the Fenchel representation

of f to be explicit, the Fenchel dual of f should be readily available, and functions f with this property are indeed a rare commodity. For example, there is seemingly no way to represent in a closed form the Fenchel dual of a function as simple as $\exp\{x\}+\frac{1}{2}\exp\{-x\}$. In contrast to this, there is no difficulty in pointing out an explicit *Fenchel-type* representation of the latter function (or, for that matter essentially any convex function given by a closed form analytical expression). This is the dramatic difference between Fenchel and *Fenchel-type* representations, and is precisely the reason we will be using Fenchel-*type* representations instead of Fenchel representations. This difference stems from the fact that Fenchel-type representations admit a kind of *fully algorithmic calculus*, meaning that all standard convexity-preserving operations for convex functions, (such as taking linear combinations with nonnegative coefficients, or maximum, or affine substitution of variables), can be accompanied by simple rules which build an explicit Fenchel-type representation for the result of the operation via Fenchel-type representations of the operands. Here are the most important "calculus rules:"

1. [Taking weighted sums] Let the functions f_i, $i=1,...,k$, be given by Fenchel-type representations:
$$f_i(x)=\sup_{y_i}\left\{\langle x,A_iy_i+a_i\rangle-h_i(y_i)\right\},$$
and let $\lambda_i\geq 0$. We clearly have
$$\sum_{i=1}^k\lambda_if_i(x)=\sup_{y=[y_1;...;y_k]}\Big\{\langle x,\underbrace{\sum_i\lambda_i[A_iy_i+a_i]}_{=:Ax+a}\rangle-\underbrace{\sum_i\lambda_ih_i(y_i)}_{=:h(y)}\Big\};$$
thus, *a Fenchel-type representation of $\sum_{i=1}^k\lambda_if_i(x)$ is readily the given by Fenchel-type representations of f_i.*

2. [Affine substitution of variables] Let $f(x)=\sup_y\left\{\langle x,Ay+a\rangle-h(y)\right\}$ be given by a Fenchel-type representation, and let $g(u)=f(Bu+b)$. We clearly have
$$g(u)=\sup_y\left\{\langle Bu+b,Ay+a\rangle-h(y)\right\}=\sup_y\Big\{\langle u,\underbrace{B^TAy+B^Ta}_{=:\widetilde{A}y+\widetilde{a}}\rangle-[\underbrace{h(y)-\langle b,Ay+a\rangle}_{=:\widetilde{h}(y)}]\Big\},$$
thus, *a Fenchel-type representation of g is readily given by a Fenchel-type representation of f.*

3. [Taking maximum] Let $f_i(x)=\sup_{y_i}\left\{\langle x,A_iy_i+a_i\rangle-h_i(y_i)\right\}$, $1\leq i\leq k$, be functions

given by Fenchel-type representations. Then

$$
\begin{aligned}
f(x) &:= \max_{i\le i\le k} f_i(x) = \sup_{\lambda>0,\sum_i \lambda_i=1} \textstyle\sum_i \lambda_i f_i(x) \\
&= \sup_{\lambda>0,\sum_i \lambda_i=1} \sup_{y=[y_1;...;y_k]} \{\langle x, \textstyle\sum_i \lambda_i [A_i y_i + a_i] - \sum_i \lambda_i h_i(y_i)\} \\
&= \sup_{\lambda>0,\sum_i \lambda_i=1} \sup_{w=[w_1;...;w_k]} \{\langle x, \textstyle\sum_i [A_i w_i + \lambda_i a_i]\rangle - \sum_i \lambda_i h_i(w_i/\lambda_i)\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad [\text{substitution } w_i = \lambda_i y_i] \\
&= \sup_{z=[\lambda;w]:\lambda>0,\sum_i \lambda_i=1} \Big\{ \langle x, \underbrace{\sum_i [A_i w_i + \lambda_i a_i]}_{=:Az+a} \rangle - \underbrace{\sum_i \lambda_i h_i(w_i/\lambda_i)}_{=h(z)} \Big\} \\
&= \sum_{z=[\lambda;w]} \Big\{ \langle x, Az + a\rangle - \tilde{h}(z) \Big\},
\end{aligned}
$$

where

$$
\tilde{h}([\lambda; w_1; ...; w_k]) = \begin{cases} \sum_i \widetilde{h}_i(\lambda_i, w_i), & \lambda \ge 0, \sum_i \lambda_i = 1 \\ +\infty, & \text{otherwise} \end{cases}
$$

and $\widetilde{h}_i(\lambda_i, w_i)$ is a lower semicontinuous convex extension of the function $\lambda_i h_i(w_i/\lambda_i)$ of $\lambda_i > 0$ and w_i (it is well known that this function is convex on its domain) on the domain $\{(\lambda_i, w_i) : \lambda_i \ge 0)$. We see that *taking maximum preserves explicit Fenchel-type representability.*

The above outlined "calculus rules", along with several more advanced rules, can be augmented by explicit knowledge of Fenchel representations of elementary univariate functions (given by explicit computation of their Fenchel duals) to yield a powerful "fully algorithmic" calculus of Fenchel-type representations. As a result, from a practical perspective we claim that for almost all proper convex lower semicontinuous functions, which admit a closed form representation, one can build an explicit Fenchel-type representation as well.

To aid in seeing how this can be accomplished we present three simple illustrations:
Example 4.2.1. The function $f(x) = \sum_{i=1} \exp\{a_i^T x + b_i\}$ admits an explicit Fenchel-type representation, specifically,

$$
f(x) = \sup_{y\in\mathbf{R}^k} \left\{ \langle x, \sum_i y_i a_i\rangle - \sum_i [y_i \ln y_i + (b_i - 1) y_i] \right\},
$$

where we use the convention that $s \ln s$ is zero when $s = 0$ and is $+\infty$ when $s < 0$.
Indeed, via direct computation of the Fenchel dual of the exponent we have

$$
\exp\{s\} = \sup_t \{st - [t \ln t - t]\}.
$$

Thus it only remains to apply the calculus rules on affine substitution of variables and summation. This result provides a nice counterpoint to the aforementioned fact that the explicit Fenchel representation of $\exp\{x\} + \frac{1}{2}\exp\{-x\}$ is beyond our reach.

Example 4.2.2. The convex function $f(x) = \ln\left(\sum_{i=1}^k \exp\{a_i^T x + b_i\}\right)$ admits an explicit Fenchel type representation, specifically,

$$f(x) = \sup_{y\in\mathbf{R}^k}\left\{\langle x, \sum_i y_i a_i\rangle - \left[\sum_i [y_i \ln y_i - b_i y_i] + \chi(\sum_i y_i - 1)\right]\right\},$$

where, as above, $s\ln s$ is 0 when $s = 0$, is $+\infty$ when $s < 0$ and $\chi(\cdot)$ is the proper lower semicontinuous function on the real axis equal to 0 when $s = 0$ and equal to $+\infty$ otherwise.

Indeed, via direct computation of Fenchel dual we have

$$\ln\left(\sum_i \exp\{s_i\}\right) = \sup_{y\geq 0, \sum_i y_i = 1}\left\{\sum_i s_i y_i - \sum_i y_i \ln y_i\right\},$$

and it only remains to use the calculus rule of affine substitution of variables.

Example 4.2.3. Let $B(x) = B_0 + \sum_{i=1}^k x_i B_i$, where $B_0, B_1, ..., B_k$ are symmetric $m \times m$ matrices, and let $S_p(B)$ be the sum of p largest eigenvalues of a symmetric matrix B; here $m \geq p$. The function $f(x) = S_p(B(x))$ admits an explicit Fenchel-type representation, specifically,

$$f(x) = \sup_{Y\in\mathbf{S}^m}\{\underbrace{\mathrm{Tr}(Y\sum_i x_i B_i)}_{=:\langle x, \mathcal{A}(Y)\rangle} - \underbrace{[\xi(Y) - \mathrm{Tr}(B_0 Y)]}_{=:h(Y)}\},$$

where $\mathbf{S}^m$ is the space of symmetric $m \times m$ matrices, the linear mapping $\mathcal{A} : \mathbf{S}^m \to \mathbf{R}^k$ is given by $\mathcal{A}(Y) = [\mathrm{Tr}(B_1 Y); \mathrm{Tr}(B_2 Y); ...; \mathrm{Tr}(B_k Y)]$, and

$$\xi(Y) = \begin{cases} 0, & 0 \preceq Y \preceq I, \mathrm{Tr}(Y) = p \\ +\infty, & \text{otherwise} \end{cases}.$$

Where we define the notation $P \succeq Q \Leftrightarrow Q \preceq P$ to mean that P, Q are symmetric matrices of the same size such that $P - Q$ is positive semidefinite.

Indeed, applying the rule on affine substitution of variables, it suffices to verify that

$$S_p(X) = \max_{0\preceq Y\preceq I, \mathrm{Tr}(Y)=p} \mathrm{Tr}(YX),\ \forall X \in \mathbf{S}^m.$$

While this equality can be derived in a systematic way, here we will simply demonstrate that it holds. Let us fix X and pass to the representation of all symmetric matrices in the orthonormal basis where X becomes diagonal; the statement to be proved now reads

$$s_p(x) = \max_{y\in\mathcal{Y}} x^T y,$$

where $s_p(u)$ is the sum of the p largest entries in a vector $u \in \mathbf{R}^m$, x is the vector of eigenvalues of X (recall that X is now diagonal), and $\mathcal{Y}$ is the set of all diagonals of $m \times m$ symmetric matrices which are $\succeq 0$, $\preceq I$ and have trace p. In other words, $\mathcal{Y}$ is the set of all nonnegative vectors with entries ≤ 1 and sum of entries equal to p. By elementary Linear Programming, the extreme points of $\mathcal{Y}$ are nothing but Boolean vectors with exactly p entries equal to 1, and therefore $\max_{y \in \mathcal{Y}} x^T y$ is indeed the sum of the p largest entries in x.

4.2.3 Representations of Solids by LO Oracles.

We claimed that there exist meaningfully situations when convex solids which admit Linear Optimization oracles are superior (and as will be seen, sometimes exclusive) to Separation Oracles which we have till now been solely working with. To this end we present several pseudo-academic examples of situations where LO oracles are the sole readily available oracles, or are preferable to separation oracles.

A. A simple generic example of a solid X which can be easily equipped with an LO oracle, but for which an equally simple Separation oracle can be problematic, is the situation where X is given as a convex hull of the union of convex compact sets X_i, $i = 1, ..., m$ which are "simple' (say, simplexes, boxes, balls) and thus can be equipped with easy-to-implement LO and Separation oracles. Clearly, easy-to-implement LO oracles for X_i, $i = 1, ..., m$, induce an easy-to-implement (unless m is very large) LO oracle for X: in order to maximize a linear form over $X = \mathrm{Conv}\{\cup_i X_i\}$, we maximize it over every one of the sets $X_1, ..., X_m$; a maximizer of the form over X is nothing but the best — with the largest value of the form – of the "partial maximizers" we get. Note however that in comparison there is *no* easy way to combine Separation oracles for X_i into a Separation oracle for X.

B. Another generic example of a solid X such that it is easy to maximize linear forms over X, while building a separation oracle for X is a nontrivial problem, is offered by Markov Decision processes. Imagine that we control a discrete time dynamical system over a finite time horizon $1, ..., T$, and the state of the system at every time t can be identified with a point from a given finite set which we without loss of generality identify with the "discrete segment" $S = \{1, ..., N\}$. The evolution of the system is defined by our *actions* a_t at times $1, ..., T$ taking values in another finite set, say, $\{1, ..., M\}$, according to the *state equations*

$$s_t = F_t(s_{t-1}, a_t),$$

where $s_t \in S$ is the state of the system at time t, and the *transition rules* $F_t(s, a) : \{1, ..., N\} \times \{1, ..., M\} \to \{1, ..., N\}$ are given in advance. Assuming for the sake of definiteness that the initial state s_0 of the system is once for ever fixed, we can associate with every sequence $a^{(T)} = [a_1, ..., a_T]$ of control actions the corresponding trajectory $s^{(T)} = [s_1; ...; s_T]$ of the system, which can be considered as a vector from $\mathbf{R}^T$. Denoting by X the convex

hull of all feasible (i.e., achievable via certain controls) trajectories, we get a polytope in $\mathbf{R}^T$. Note that unless M, N are astronomically large, the maximization of a linear function $e^T x$ over $x \in X$ is easy. Indeed, maximizing a linear form over X is the same as maximizing it over the feasible trajectories of the underlying dynamical system, which can be done by Dynamic Programming, that is, by generating, backward in time, the cost-to-go functions

$$C_t(s) = \max_{1 \le a \le M} \left[e^T s + C_{t+1}(F_t(s,a)) \right] : S = \{1, ..., N\} \to \mathbf{R}, \;\; t = T, T-1, ..., 1, \qquad [C_{T+1}(s) \equiv 0]$$

Note that there is no equally simple mechanism for separating a given point from the polytope X.

The following is an example of a meaningful situation where we are interested in minimizing a *nonlinear* (and non-separable) convex function over a polytope X, which is given by the above construction.

Example 4.2.4. [Protein similarity problem] One possible method for modeling a protein (a chain of amino acids folded in $\mathbf{R}^3$) is via a graph with ordered vertices $1, 2, ...T$. Where each vertex represents an amino acids, and the vertex order is derived from the order of the amino acids in the chain. In order to model the 3-D folding of the chain two vertices are linked by an arc when the corresponding amino acids are spatially close to one another. Now, given two proteins, **P** with T amino acids, and **Q** with $N > T$ amino acids, it is sometimes important to find out whether P is "similar" to a subsection of **Q**. A simple mathematical model of "similarity" is as follows: we map the vertices of P onto the vertices of Q in *some sort of order-preserving fashion*, and count how many pairs of vertices which are adjacent in P are mapped onto pairs which are adjacent in Q. We then maximize the result of this count over *all order-preserving embeddings* of P onto Q and treat the result as the "degree of similarity" of P and a part of Q. Now, order-preserving embeddings of P onto Q can be easily identified with the feasible trajectories of a discrete time dynamical system with moderate cardinalities (like $N + 1$) of the state and the action spaces on time horizon $1, ..., T$. The similarity is the maximum of an easy-to-compute function $\psi(s^{(T)})$ over all feasible trajectories $s^{(T)}$. The difficulty, however, is that this function is not separable with respect to states, which makes it impossible to maximize ψ over the feasible trajectories by Dynamic Programming. In fact, this maximization is a difficult combinatorial problem. We can, however, easily point out a simple *concave* (but nonlinear!) function $f : \mathbf{R}^T \to \mathbf{R}$ which coincides with ψ at every trajectory, and then maximize f over the convex hull X of feasible trajectories, thus obtaining an upper bound on the true similarity. Note that computing this bound is nothing more then minimizing an explicitly given convex function (namely, $-f$) over a polytope X given by the LO oracle. That is, it is exactly our current problem of interest. This bound can be used as a computable substitution of the "true" similarity; alternatively, it can be used within a branch-and-bound scheme aimed at computing similarity.

C. In **A**, and **B** we were speaking about situations where the solid of interest X can be equipped with a relatively easy-to-implement LO oracle, but not with an easy-to-implement

Separation oracle. There are also situations when both types of oracles are easy to implement, but an LO representation is more attractive computationally. Specifically, consider the case when we are solving problem (61) and have at our disposal

• LO and Separation oracles for the feasible domain X of the problem,

• a First Order oracle and a Fenchel-type representation (62) for the objective f, and, in addition,

• a Separation oracle for the set $Y = \mathrm{cl}\,(\mathrm{Dom}\, h)$ along with a First Order oracle for h, where h is the convex function participating in the Fenchel-type representation (62) of f.

In this ideal situation, where everything we might need is available, consider the case when the dimension of X is much larger than the dimension of $Y = \mathrm{cl}\,(\mathrm{Dom}\, h)$. With the approach we are about to develop, solving (61), i.e., minimizing f over X, reduces to minimizing a convex function ϕ *over* Y ("the dual problem"), where the First Order oracle for ϕ is readily given by the LO oracle for X and the First Order oracle for $h(\cdot)$. Assuming we are looking for a high accuracy solution and are implementing a black box oriented method, say, the Ellipsoid algorithm, our options are to apply this algorithm directly either to the problem of interest $\min\limits_X f$, or to the dual problem $\min\limits_Y \phi$. Since the rate of convergence of the Ellipsoid algorithm is heavily affected by dimension, in the case of $\dim Y \ll \dim X$, the second option – which uses a LO representation of X and a Fenchel-type representation of f — is highly preferable.

Example 4.2.5. As a simple example, note that in the NERML algorithm, with or without certificates, one should, at every step, solve to within a high accuracy an auxiliary problem of the form $\min\limits_{z\in\mathbf{Q}} g(z)$, where $\mathbf{Q}$ is a simple solid[1] and $g(\cdot)$ is the maximum of m linear forms $\langle z, a_s\rangle + b_s$, $1 \le s \le m$. The number m of these linear forms (the memory depth of NERML) is under our full control and usually is at most a few tens. Note that g admits a simple Fenchel-type representation

$$g(z) = \max_y \Big\{ \underbrace{\langle z, \sum_{s=1}^m y_s a_s\rangle}_{\langle z, Ay+a\rangle} - h(y)\Big\},$$

where $h(y) = -\sum_s b_s y_s$ when y belongs to the standard simplex $Y = \{y \ge 0 : \sum_s y_s = 1\}$ and is $+\infty$ outside of Y. In the large scale case, when the dimension of $\mathbf{Q}$ is of the order of thousands or more, it is incomparably easier to solve, by a black box oriented method, the low dimensional dual problem $\min\limits_Y \phi$, $\phi(y) = h(y) - \min\limits_{z\in\mathbf{Q}}\langle z, Ay + a\rangle$, rather than to directly attack, with whatever algorithm (including an interior point one) the large-scale problem $\min\limits_{\mathbf{Q}} g$.

[1] Specifically, $\mathbf{Q} = \{[x;t] : x \in \mathbf{B}, \omega(x) \le t \le \max_{\mathbf{B}} \omega\}$, where $\mathbf{B}$ and $\omega(\cdot)$ are the entities participating in NERML's setup; note that the assumption that one can easily maximize linear forms over $\mathbf{Q}$ is nothing but the assumption that $\mathbf{B}$ and $\omega(\cdot)$ are simple and match each other, made when presenting the NERML setup.

4.2.4 Construction and Main Result

Recall that our current goal is to solve the convex minimization problem (61) when X is represented by an LO oracle, and f is given by a Fenchel-type representation (62). As for the function h participating in this representation, we assume from now on that the set $Y = \mathrm{cl}\,(\mathrm{Dom}\, h)$ is a solid given by a Separation oracle, and h is given by a First Order oracle which serves either $\mathrm{int}Y$, or perhaps the entire Y.

The dual problem. Given a Fenchel-type representation (62) of f, we associate with (61) the saddle point problem

$$(\mathbf{P}) \quad \min_{x\in X}\max_{y\in Y}\left[\langle x, Ay+a\rangle - h(y)\right]$$

and consider the associated dual problem which it is convenient for us to write down in the minimization form:

$$(\mathbf{D}) \quad \min_{y\in Y}\phi(y), \text{ where } \phi(y) = -\min_{x\in X}\left[\langle x, Ay+a\rangle - h(y)\right] = \max_{x\in X}\left[h(y) - \langle x, Ay+a\rangle\right]$$

(via standard duality the dual problem would be $\max\limits_{y\in Y}[-\phi(y)]$, which is equivalent to (D)). Note that since X and Y are solids, by the Sion-Kakutani Theorem, and taking into account that the optimal value in our dual problem is minus the optimal value in the standard (i.e., the maximization) dual, the optimal values $\mathrm{Opt}(P)$ and $\mathrm{Opt}(D)$ in the problems sum up to zero:

$$\mathrm{Opt}(P) + \mathrm{Opt}(D) = 0. \tag{64}$$

Observe that the objective ϕ of the dual problem (D) admits a First Order oracle which serves either $\mathrm{int}Y$, or the entire Y, depending on whether the First Order oracle for h serves $\mathrm{int}Y$ or Y. Indeed, given y, we can call the Linear Optimization oracle for X to compute

$$x_y \in \mathop{\mathrm{Argmin}}_{x\in X}\langle x, Ay+a\rangle.$$

Given x_y, the value $h(y)$, and a subgradient $h'(y)$ of h at y (as provided by the First Order oracle for h), the value, and a subgradient of ϕ at y are readily given by

$$\phi(y) = h(y) - \langle x_y, Ay+a\rangle,\ \ \phi'(y) = h'(y) - A^T x_y. \tag{65}$$

It follows that $(\mathbf{D})$ can be solved via "usual" black box oriented algorithms for convex minimization, (i.e those which rely on a Separation oracle for the feasible domain and a First Order oracle for the objective). However, while this approach can recover the optimal value $\mathrm{Opt}(P) = -\mathrm{Opt}(D)$ of the problem of interest (P), it cannot find an approximate solution to (P), only an approximate solution to the dual problem (D). Of course, when solving (D), we get a lot of feasible solutions to (P) — specifically, all the points x_y associated with the search points y where we invoke the First Order oracle for $\phi(\cdot)$, but there is no

reason why any one of these points should be a good approximate solution to (P). (Indeed x_y typically will be extreme points of X, while the true optimal set of (P) will probably be contained in $\mathrm{int}X$.) We are about to demonstrate that when solving (D) via an algorithm *equipped with certificates*, these certificates allow us to recover good approximate solutions to (P). Towards this end we present the following construction.

The construction and the main result. When solving (D) by a black box oriented algorithm equipped with certificates, at every productive step $t \in I_\tau$ (i.e. where $y_t \in Y$ and the First Order oracle for ϕ is invoked), we get a point $x_t \in X$ which is a minimizer of the linear function $\langle x, Ay_t + a\rangle$ over $x \in X$. Thus, the execution protocol $P_\tau = (\{y_t, e_t\}_{t=1}^\tau, I_\tau, J_\tau)$, (where $e_t = \phi'(y_t)$ for $t \in I_\tau$, and e_t separates y_t and Y for $t \in J_\tau$,) is augmented by the collection $\{x_t\}_{t\in I_\tau}$. Now, assuming that the algorithm associates with P_τ an accuracy certificate ξ^τ, we can also associate with the protocol the point

$$x^\tau = \sum_{t\in I_\tau} \xi_t^\tau x_t; \tag{66}$$

this point belongs to X (as a convex combination of points $x_t \in X$; recall that ξ^τ is an accuracy certificate). Our main result here is:

Theorem 4.2.1. *In the outlined situation, let τ be such that the execution protocol P_τ is augmented with an accuracy certificate ξ^τ, so that x^τ is well defined and belongs to X. Then*

$$\varepsilon_{\mathrm{opt}}(x^\tau) := f(x^\tau) - \min_X f \le \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}), \tag{67}$$

where $\mathbf{B}$ is a solid containing Y.

Proof. For $y \in Y$, we have

$$\begin{aligned}
&\langle x^\tau, Ay + a\rangle - h(y)\\
=\ &\sum_{t\in I_\tau} \xi_t[\langle x_t, Ay + a\rangle - h(y)]\\
=\ &\sum_{t\in I_\tau} \xi_t[\langle x_t, Ay_t + a\rangle + \langle x_t, A(y - y_t)\rangle - h(y_t) + h(y_t) - h(y)] \qquad (68)\\
=\ &\sum_{t\in I_\tau} \xi_t[\langle x_t, Ay_t + a\rangle - h(y_t)] + \underbrace{\sum_{t\in I_\tau} \xi_t[\langle x_t, A(y - y_t)\rangle + h(y_t) - h(y)]}_{\delta}
\end{aligned}$$

Recalling that $e_t = \phi'(y_t)$ for $t \in I_\tau$ and e_t separates y_τ and Y when $t \notin I_\tau$, we have for $y \in Y$:

$$\begin{aligned}
\varepsilon_{\mathrm{cert}} := \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}) \ &\ge \ \sum_{t\in I_\tau} \xi_t \langle A^T x_t - h'(y_t), y - y_t\rangle\\
&= \ \sum_{t\in I_\tau} \xi_t[\langle x_t, A(y - y_t)\rangle - \langle h'(y_t), y - y_t\rangle]\\
&\ge \ \underbrace{\sum_{t\in I_\tau} \xi_t[\langle x_t, A(y - y_t)\rangle + h(y_t) - h(y)]}_{\delta}
\end{aligned}$$

Where the last inequality follows the gradient inequality on the convex function h

$$\langle h'(y_t), y - y_t\rangle \le h(y) - h(y_t)$$

Hence we have that

$$\begin{aligned}&\sum_{t\in I_\tau} \xi_t[\langle x_t, Ay_t + a\rangle - h(y_t)] + \sum_{t\in I_\tau} \xi_t[\langle x_t, A(y - y_t)\rangle + h(y_t) - h(y)]\\ &\le -\sum_{t\in I_\tau} \xi_t\phi(y_t) + \varepsilon_{\rm cert} \le -\min_{y\in Y} \phi(y) + \varepsilon_{\rm cert}\end{aligned}$$

Recalling that

$$\mathrm{Opt}(D) = \min_{y\in Y} \phi(y) = -\min_{x\in X} f(x) \equiv -\mathrm{Opt}(P)$$

we conclude that

$$\sum_{t\in I_\tau} \xi_t[\langle x_t, Ay_t + a\rangle - h(y_t)] + \sum_{t\in I_\tau} \xi_t[\langle x_t, A(y - y_t)\rangle + h(y_t) - h(y)] \le \mathrm{Opt}(P) + \varepsilon_{\rm cert},$$

which combines with (68) to imply that

$$\forall y \in Y : \langle x^\tau, Ay + a\rangle - h(y) \le \mathrm{Opt}(P) + \varepsilon_{\rm cert}.$$

The supremum of the right hand side in the latter inequality is $f(x^\tau)$, and (67) follows. ∎

Thus, when an accuracy of ϵ can be certified by an accuracy certificate ($\varepsilon_{\rm cert} \le \epsilon$), when solving the dual problem, the certificate provides us with a feasible ϵ-optimal solution to the problem of interest (61).

4.3 Minimizing Convex Function over a Solid Given by a Linear Optimization Oracle: Extensions

4.3.1 Problems with Functional Constraints

Now consider the situation where the problem of interest is a solvable problem of the form

$$\mathrm{Opt} = \min_{x\in X} \{f(x) : f_i(x) \le 0,\ 1 \le i \le m\}, \tag{69}$$

where, as above, X is a solid given by an LO oracle, and f, f_i are convex functions given by Fenchel-type representations. We can handle this problem as follows: let us set

$$f^L(x) = f(x) + L\max[0, f_1(x), ..., f_m(x)],$$

where $L > 0$ is a penalty parameter. It is immediately seen that an ϵ-optimal feasible solution $\bar{x}$ to the penalized problem

$$\min_{x\in X} f^L(x) \tag{70}$$

satisfies

$$f(\bar{x}) - \mathrm{Opt} \leq \epsilon, f_i(\bar{x}) \leq \frac{V+\epsilon}{L}, 1 \leq i \leq n,$$

where $V = \mathrm{Opt} - \min_{x \in X} f(x)$. It follows that given an "optimality tolerance" $\epsilon > 0$ *and* a feasibility tolerance $\delta > 0$, for all sufficiently large values of the penalty parameter L, an ϵ-optimal feasible solution $\bar{x}$ to (70) will be ϵ-optimal, and δ-feasible for the problem of interest (69):

$$\bar{x} \in X,\ f(\bar{x}) \leq \mathrm{Opt} + \epsilon,\ f_i(\bar{x}) \leq \delta,\ 1 \leq i \leq m. \tag{71}$$

Usually it is possible to bound V from above, and thus to point out explicitly a "large enough" value of the penalty.

Now, for any given L, the aforementioned calculus of Fenchel-type representations allows us to combine Fenchel-type representations of f and f_i, which are initially given, into an explicit Fenchel-type representation of f^L, and we can use this representation and the LO oracle for X in order to solve (70) to within a desired accuracy ϵ, as explained in section 4.2, thus solving the problem of interest (69) to within the given tolerances ϵ, and δ.

4.3.2 Minimizing over Intersection of Solids Given by Linear Maximization Oracles

Consider the following problem which is of definite academic interest. We want to solve a convex minimization problem

$$\mathrm{Opt} = \min_{x \in X} f(x), \tag{72}$$

where, as always, $X \subset \mathbf{R}^n$ is a solid and f is convex. However, now we assume that neither a Separation, nor a LO oracle for X are available; all we know is that $X = \cap_{i=1}^{m+1} X_i$ for some solids X_i which do have oracle representations, either all by Separation oracles, or all by LO oracles. The situation when all the X_i are given by Separation oracles is easy: these oracles induce straightforwardly a Separation oracle for $X = \cap_i X_i$. In contrast to this, the case when X_i are represented by LO oracles seems to be difficult: there is no simple way to combine these oracles into an LO or a Separation oracle for X. We can however use the above construction as follows.

Let $X_+ = X_1 \times X_2 \times ... \times X_{m+1}$. Denoting a point z from the linear space $\mathbf{R}^{(m+1)n} = \mathbf{R}^n \times ... \times \mathbf{R}^n$ (where $X_+ \subset \mathbf{R}^{(m+1)n}$) as $z = [x_1; ...; x_{m+1}]$ with $x_i \in \mathbf{R}^n$, we can rewrite the problem of interest as

$$\min_{z=[x_1;...;x_{m+1}] \in X_+} \{f_+(z) := f(x_1) : f_i(z) := \|x_{i+1} - x_1\|_2 \leq 0,\ 1 \leq i \leq m\} \tag{73}$$

Observe that the LO oracles for X_i clearly induce an LO oracle for X_+. Assuming that f is given by a Fenchel-type representation

$$f(x) = \sup_y \{\langle x, Ay + a \rangle - h(y)\},$$

a Fenchel-type representation of $f_+^L(z) = f_+(z) + L\max[f_1(z), ..., f_m(z)]$ is readily available, specifically:

$$f_+^L(x_1, ..., x_{m+1}) = \sup_{w=[y;y_1,...,y_m]} \Big\{ \underbrace{\langle x_1, Ay + a\rangle + L\sum_{i=1}^{m}\langle x_{i+1} - x_1, y_i\rangle}_{\langle [x_1;...;x_{m+1}], A_+w+a_+\rangle} - h(y) - H(y_1, ..., y_m) \Big\},$$

where $H(y_1, ..., y_m) = 0$ when $\|y_i\|_2 \le 1$, $i = 1, ..., m$, and $H(y_1, ..., y_m) = +\infty$ otherwise. We can now apply the above machinery to solve (73). Note that the component $\bar{x}_1$ of ϵ-optimal and δ-feasible solution x^τ to (73) satisfies $f(\bar{x}_1) \le \mathrm{Opt} + \epsilon$ and "nearly belongs" to $X = \cap_{i=1}^{m+1} X_i$, specifically, it belongs to X_1 and is at the distance at most δ from each of the sets $X_2, ..., X_{m+1}$.

4.3.3 The Case of an Approximate LO Oracle

Situation and Goal: Some *computationally intractable* convex solids X can be equipped with *approximate* LO oracles capable of approximating the maximum/minimum over X of linear forms from a specified family Ξ within a given *approximation factor*. A typical result is as follows:

> *Given a tolerance $\epsilon > 0$, a solid X from a particular family $\mathcal{X}$ and a vector ξ from a particular family Ξ such that ξ matches X (i.e., ξ belongs to the Euclidean space $\mathbf{R}_X$ such that $X \subset \mathbf{R}_X$), the algorithm in question can find efficiently a point $\bar{x} \in X$ such that*
>
> $$\langle \xi, \bar{x}\rangle \le \alpha \mathrm{Opt}_*(\xi, X) + \epsilon, \quad \mathrm{Opt}_*(\xi, X) = \min_{x\in X}\langle \xi, x\rangle, \tag{74}$$
>
> *where the approximation factor α is a specific quantity independent of the numerical values of the data and of ϵ.*

Note that unless $\alpha = 1$, such a formulation imposes implicit restrictions on the sign of $\mathrm{Opt}_*(\xi, X)$. Indeed, when $\alpha > 1$, this value should be nonnegative (otherwise $\alpha\mathrm{Opt}_*(\xi, X) < \mathrm{Opt}_*(\xi, X)$, and the required $\bar{x}$ simply cannot exist). Similarly, when $0 < \alpha < 1$, the optimal value $\mathrm{Opt}_*(\xi, X)$ must be nonpositive.

Thus, there exist four possible types of approximation results:

- *Two in the minimization framework*
  - "min-plus-definite," where the structures of $\mathcal{X}$ and Ξ ensure that $\mathrm{Opt}_*(\xi, X) \ge 0$ for all matching pairs $\xi \in \Xi$, $X \in \mathcal{X}$, and the approximation factor α is ≥ 1;

– "min-minus-definite," where the structures of $\mathcal{X}$ and Ξ ensure that $\mathrm{Opt}_*(\xi, X) \leq 0$ for all matching pairs $\xi \in \Xi$, $X \in \mathcal{X}$, and the approximation factor α is ≤ 1 and ≥ 0.

- *and two in the maximization framework,* where we want to maximize rather than minimize — to find $\bar{x} \in X$ such that

$$\langle \xi, \bar{x} \rangle \geq \alpha \mathrm{Opt}^*(\xi, X) - \epsilon, \ \mathrm{Opt}^*(\xi, X) := \max_{x \in X} \langle \xi, x \rangle$$

– "max-plus-definite," where the structures of $\mathcal{X}$ and Ξ ensure that $\mathrm{Opt}^*(\xi, X) \geq 0$ for all matching pairs $\xi \in \Xi$, $X \in \mathcal{X}$, and the approximation factor α is ≥ 0 and ≤ 1;

– "max-minus-definite," where the structures of $\mathcal{X}$ and Ξ ensure that $\mathrm{Opt}^*(\xi, X) \leq 0$ for all matching pairs $\xi \in \Xi$, $X \in \mathcal{X}$, and the approximation factor α is ≥ 1.

Clearly, when passing from the family Ξ to the family $-\Xi$, that is, from the minimization of $\langle \xi, x \rangle$ to the maximization of $\langle -\xi, x \rangle$, a min-plus-definite approximation result translates into an equivalent max-minus-definite result, and a min-minus-definite result translates into an equivalent max-plus-definite one. It follows that we can restrict our attention to "min-type" settings and results.

The problem we intend to consider now is:

> *Assume that a particular pair $(\mathcal{X}, \Xi)$ admits a min-type approximation result. To what extent can we extend the efficiency and accuracy guarantees yielded by this result when passing from minimizing linear forms $\langle \xi, \cdot \rangle$, $\xi \in \Xi$, over X to minimizing a convex function $f(\cdot)$ over X?*

Clearly, in order to hope for a meaningful answer, we need to impose on f restrictions which ensure that

– first, f is somehow "comprised" of linear forms from the family Ξ, and,

– second, the true minimum of f is of the required sign (i.e., nonnegative for a min-plus-definite result, and nonpositive for a min-minus-definite one).

To this end, it is natural to impose restrictions on a Fenchel-type-representation of f, as in the following definition:

Definition 4.3.1. (i) Let $\mathcal{X}$ be a family of solids, and Ξ be a family of vectors from Euclidean spaces. We say that these families form a

- plus-definite pair, if whenever $X \in \mathcal{X}$ and $\xi \in \Xi$ match each other (i.e., $\xi \in \mathbf{R}_X$), we have $\mathrm{Opt}_*(\xi, X) = \min_{x \in X} \langle \xi, x \rangle \geq 0$.

- minus-definite pair, if whenever $X \in \mathcal{X}$ and $\xi \in \Xi$ match each other, we have $\mathrm{Opt}_*(\xi, X) = \min\limits_{x \in X} \langle \xi, x \rangle \leq 0$.

(ii) Let Ξ be a family of vectors from Euclidean spaces, and let f be a function on a Euclidean space E given by a Fenchel-type representation $f(x) = \sup_y [\langle x, Ay + a \rangle - h(y)]$. We say that

- f plus-matches Ξ, if $Ay + a \in \Xi$ whenever $y \in \mathrm{cl}\,\mathrm{Dom}\, h$, and h is nonpositive on its domain;
- f minus-matches Ξ, if $Ay + a \in \Xi$ whenever $y \in \mathrm{cl}\,\mathrm{Dom}\, h$, and h is nonnegative on its domain.

Note that the sign, and the domain restrictions imposed on h in this definition imply that when a function f on a Euclidean space E is given by a Fenchel-type representation (62), Ξ is a family of vectors from Euclidean spaces, and $\mathcal{X}$ is a family of solids in these spaces, then

- whenever f plus-matches Ξ, and ($\mathcal{X}$,Ξ) is a plus-definite pair, then for every solid $X \in \mathcal{X}$ which is contained in E one has $\mathrm{Opt}_*(f, X) := \min\limits_X f \geq 0$;
- whenever f minus-matches Ξ, and ($\mathcal{X}$,Ξ) is a minus-definite pair, then for every solid $X \in \mathcal{X}$ which is contained in E one has $\mathrm{Opt}_*(f, X) \leq 0$.

We are about to demonstrate that if $(\mathcal{X}, \Xi)$ form a plus/minus-definite pair which admits an approximation algorithm for linear minimization with some approximation factor, then this algorithm can be extended to minimizing convex functions matching Ξ, the approximation factor being preserved. Specifically, consider a situation as follows: we are given

1. a sign-definite pair $(\mathcal{X}, \Xi)$ (i.e., either a plus-definite, or a minus-definite one),
2. a solid $X \in \mathbf{R}^n$ which belongs to $\mathcal{X}$ and is equipped with *an approximate* LO oracle. This oracle, given on input a linear form $\langle \xi, x \rangle$ with $\xi \in \Xi$, and a tolerance $\epsilon > 0$, returns a point $\bar{x}_\epsilon[\xi] \in X$ such that

$$\langle \xi, \bar{x}_\epsilon[\xi] \rangle \leq \alpha \min_{x \in X} \langle \xi, x \rangle + \epsilon, \tag{75}$$

   where α is certain approximation factor (which is ≥ 1 when $(\mathcal{X}, \Xi)$ is plus-definite, and belongs to $[0, 1]$ when $(\mathcal{X}, \Xi)$ is minus-definite)[2];
3. a function f on $\mathbf{R}^n$, given by a Fenchel-type representation (62), which sign-matches Ξ, (that is, plus-matches Ξ, when $(\mathcal{X}, \Xi)$ is a plus-definite pair, and minus-matches ξ

[2] We note that in the minus-definite case we would, equivalently to (75), have $\langle \xi, \bar{x}_\epsilon[\xi] \rangle \leq \alpha \min\limits_{x \in X} \langle \xi, x \rangle - \epsilon$

when $(\mathcal{X}, \Xi)$ is a minus-definite) pair. Assume also that the data of the Fenchel-type representation of f are such that $Y = \mathrm{cl}\,\mathrm{Dom}\, h$ is a solid given by a Separation oracle, and h is given by a First Order oracle.

Main result. Our main result on approximation algorithms is as follows:

Theorem 4.3.1. *In the outlined situation, given $\epsilon > 0$, let us apply to the problem of interest* (61) *the construction from section 4.2.4, where the precise minimizers*

$$x_t \in \operatorname*{Argmin}_{x \in X} \langle x, Ay_t + a \rangle$$

are replaced with approximate minimizers $\bar{x}_t$ as reported by the approximate LO oracle associated with X. Thus, instead of the true value and subgradient of ϕ at $y_t \in \mathrm{Dom}\,\phi$ given by (65) *the algorithm with certificates "is fed" by estimates of these quantities, specifically, the estimate $h(y_t) - \langle \bar{x}_t, Ay_t + a \rangle$ of $\phi(y_t)$ and the estimate $h'(y_t) - A^T \bar{x}_t$ of $\phi'(y_t)$.*

At every step τ where the algorithm with certificates participating in our construction augments the corresponding execution protocol $P_\tau = (\{y_t, e_t\}_{t=1}^{\tau}, I_\tau, J_\tau)$ with a certificate ξ^τ, setting

$$\widehat{x}^\tau = \sum_{t \in I_\tau} \xi_t^\tau \bar{x}_t$$

we have $\widehat{x}^\tau \in X$ and

$$f(\widehat{x}^\tau) \leq \alpha \min_{x \in X} f(x) + \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}) + \epsilon, \tag{76}$$

where $\mathbf{B}$ is a solid containing Y.

Proof is completely similar to the proof of Theorem 4.2.1.

For ease of terminology we will assume we are dealing with the plus matching, plus definite case. The minus matching, minus definite case is exactly similar. Now let τ be as in the premise of Theorem, and let (P) and (D) be the primal and the dual problem associated with f, X, Y and (62) in exactly the same way as in section 4.2.4. Finally, let

$$\phi(y) = h(y) - \max_{x \in X} \langle x, Ay + a \rangle$$

be the objective in (D). Observe that since h is nonpositive on its domain (since f plus-matches Ξ and $Ay + a \in \Xi$ for $y \in \mathrm{cl}\,\mathrm{Dom}\, h$), we have

$$\bar{\phi}(y) := h(y) - \langle \bar{x}_\epsilon[Ay + a], Ay + a \rangle \geq \alpha\phi(y) - \epsilon \tag{77}$$

for all $y \in \mathrm{Dom}\,\phi = \mathrm{Dom}\,h$ (see (75) and take into account that $\alpha \geq 1$). We now have

$$
\begin{aligned}
&\forall y \in \mathrm{Dom}\,h:\\
&[\langle \widehat{x}^\tau, Ay + a\rangle - h(y)] = \textstyle\sum_{t\in I_\tau} \xi_t^\tau\,[\langle x_t, Ay + a\rangle - h(y)] \ \text{[definition of } \widehat{x}^\tau]\\
&= \textstyle\sum_{t\in I_\tau} \xi_t^\tau \Big[[\langle x_t, Ay - Ay_t\rangle + \langle h'(y_t), y_t - y\rangle] - [h(y_t) - \langle x_t, Ay_t + a\rangle]\\
&+[h(y_t) + \langle h'(y_t), y - y_t\rangle - h(y)]\Big]\\
&\leq \textstyle\sum_{t\in I_\tau} \xi_t^\tau \Big[[\langle x_t, Ay - Ay_t\rangle + \langle h'(y_t), y_t - y\rangle] - [h(y_t) - \langle x_t, Ay_t + a\rangle]\Big]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[by gradient inequality for convex } h]\\
&\leq \textstyle\sum_{t\in I_\tau} \xi_t \Big[\underbrace{[\langle x_t, Ay - Ay_t\rangle + \langle h'(y_t), y_t - y\rangle]}_{\langle g(y_t), y_t - y\rangle} - \underbrace{[h(y_t) - \langle x_t, Ay_t + a\rangle]}_{\bar\phi(y_t)}\Big]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[definitions of } g \text{ and } \bar\phi]\\
&\leq \left[\textstyle\sum_{t\in I_\tau} \langle g(y_t), y_t - y\rangle + \sum\limits_{t\in J_\tau} \langle e_t, y_t - y\rangle\right] - \textstyle\sum_{t\in I_\tau} \xi_t^\tau\, \bar\phi(y_t)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[since } y \in \mathrm{Dom}\,h \subset Y \text{ and thus } \langle e_t, y_t - y\rangle \geq 0,\ t \in J_\tau]\\
&\leq \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}) - \textstyle\sum_{t\in I_\tau} \xi_t^\tau\, \bar\phi(y_t) \ \text{[since } y \in \mathrm{Dom}\,h \subset Y \subset \mathbf{B}]\\
&\leq \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}) + \textstyle\sum_{t\in I_\tau} \xi_t^\tau[-\alpha\phi(y_t) + \epsilon] \ \text{[by (77)]}\\
&\leq \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}) + \textstyle\sum_{t\in I_\tau} \xi_t^\tau[-\alpha\mathrm{Opt}(D) + \epsilon] \ \text{[since } y_t \in Y]\\
&= \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}) + \alpha\mathrm{Opt}(P) + \epsilon \ \text{[by (64)]}
\end{aligned}
$$

Thus,

$$\forall y \in \mathrm{Dom}\,h : \langle \widehat{x}^\tau, Ay + a\rangle - h(y) \leq \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}) + \alpha\mathrm{Opt}(P) + \epsilon,$$

since $f(\widehat{x}^\tau) = \sup_{y\in \mathrm{Dom}\,h}[\langle \widehat{x}^\tau, Ay + a\rangle - h(y)]$ by (62), (76) follows. ∎

Observe that in the convex minimization case, the convergence properties (i.e the rate at which the resolution of the generated certificates ξ^τ goes to 0 as τ grows) of the algorithms with certificates presented so far (the Ellipsoid algorithm with certificates, and the NERML algorithm with certificates), depends solely on the (semi)boundedness properties of the operator F the algorithm is applied to, and not on whether F is, or is not, the subgradient vector field of the objective we intend to minimize. In the situation we are considering now, where the operator we are processing is yielded by an approximate LO algorithm for X, these boundedness properties, as it is immediately seen, are independent of whether the LO oracle in question is a precise or an approximate one. As a result, in our present situation, the rate at which the resolution of ξ^τ goes to 0 as τ grows is essentially the same as in the case of a precise LO oracle. In particular, when the Ellipsoid algorithm with certificates is used, it takes a polynomial, in the $\dim Y$ and $\ln(1/\epsilon)$, number of steps to make the resolution $\leq \epsilon$. Looking at (76) we conclude that given the ability to efficiently minimize linear forms,

from the sign-matched family Ξ, over X to within a given approximation factor, we can then efficiently minimize *all convex functions*, sign-matching Ξ, over X to within that same approximation factor.

Example 4.3.1. Consider a unit box $B_n = \{u \in \mathbf{R}^n : \|u\|_\infty \leq 1\}$, and let X_n be the convex hull of all dyadic (rank 1) matrices uu^T, $u \in B_n$. Then the space $\mathbf{S}^n$ is the set of all $n \times n$ symmetric matrices (which contains X_n) equipped with the Frobenius inner product $\langle p, q\rangle = \mathrm{Tr}(pq)$. [3] The set X_n is "severely computationally intractable." Indeed, to minimize a linear form $\langle \xi, x\rangle$ over $x \in X_n$ given by a matrix $\xi \in \mathbf{S}^n$ is exactly the same as maximizing the quadratic form $u^T\xi u$ over the unit box B_n; it is known that the latter problem is NP-hard even in the case when ξ is a general-type positive semidefinite matrix, and when a relative accuracy of around 4% is sought. At the same time, when ξ is positive semidefinite, the problem

$$\max_{u\in B_n} u^T\xi u = \max_{x\in X_n} \langle \xi, x\rangle$$

admits a polynomial time approximation algorithm, based on semidefinite relaxation of the problem, with approximation ratio $2/\pi$ (Nesterov's $\pi/2$ Theorem [32]). In other words, setting $\mathcal{X} = \{X_n\}_{n=1}^\infty$ and $\Xi = \bigcup_n \{\xi \in \mathbf{S}^n : \xi \preceq 0\}$, we get a minus-definite pair which admits an efficient approximation algorithm with approximation factor $\alpha = 2/\pi$. Now consider the function $f(x)$ which is minus the sum of the k smallest eigenvalues of a symmetric matrix x, or, equivalently, the sum of the k largest eigenvalues of the matrix $-x$. Function f admits the explicit Fenchel representation

$$f(x) = \sup_w \{\langle x, w\rangle : w \in \mathbf{S}^n, 0 \succeq w \succeq -I, \mathrm{Tr}(w) = -k\}.$$

Observe that the feasible set of the right hand side problem can be easily represented as the affine image of a simple solid $Y = Y_n$ in a Euclidean space (the subspace of $\mathbf{S}^n$ comprised of matrices with zero trace) under an explicitly given affine mapping $y \mapsto Ay + a$, and that the image of Y under this mapping belongs to the cone of negative semidefinite matrices, i.e., belongs to our Ξ. Setting $h(y) = 0$ when $y \in Y$ and $h(y) = +\infty$ otherwise, we get a Fenchel-type representation

$$f(x) = \sup_y [\langle x, Ay + a\rangle - h(y)]$$

which shows that f minus-matches Ξ. Invoking Theorem 4.3.1, we conclude that, say, the Ellipsoid algorithm with certificates induces a polynomial time approximation algorithm for minimizing f over X_n, the approximation factor being $> 2/\pi$. Passing from minimizing f to maximizing $-f$, this result can be reformulated in a more natural form, specifically, as follows:

> *For every $\alpha' > 2/\pi$, there exists an efficient approximation algorithm with approximation factor α' which allows us to maximize within the factor $2/\pi$ the sum of the k smallest eigenvalues of a symmetric matrix running over the (heavily computationally intractable!) convex set X_n — the convex hull of dyadic matrices uu^T with $\|u\|_\infty \leq 1$.*

[3] in order to be consistent with our previous notation, we are forced now to denote matrices by lowercase letters.

CHAPTER V

DECOMPOSITION OF LINEAR PROGRAMS

5.1 Motivation

Consider a large scale Linear Programming program

$$\mathrm{Opt} = \min_{x=[x_1;x_2]} \left\{ c_1^T x_1 + c_2^T x_2 : \begin{array}{r} A^{11}x_1 + A^{12}x_2 \le b_1 \\ A^{21}x_1 + A^{22}x_2 \le b_2 \\ \|x\|_\infty \le R \end{array} \right\} \tag{78}$$

with $n = n_1 + n_2$ variables $x = [x_1; x_2]$, and $m = m_1 + m_2$ linear inequality constraints, where the sizes of A^{11}, A^{12} and A^{21} are, $m_1 \times n_1$, $m_1 \times n_2$ and $m_2 \times n_1$ respectively. Note that the bounds on variables $\|x\|_\infty \le R$, which we add for technical reasons, are of no real importance from the practical viewpoint, since R can be arbitrarily large. We assume that

(1) It is relatively easy to solve Linear Programming problems of the form

$$\min_{x_1} \left\{ c^T x_1 : A^{11}x_1 \le b, \|x_1\|_\infty \le R \right\} \tag{79}$$

For example, A^{11} can be a block-diagonal matrix with a large number N of relatively small diagonal blocks, so that (79) is just a collection of independent small LP programs. [1]

(2) The number of *linking constraints* m_2 is $\ll$ the total number of constraints m in (78), and the number of *linking variables* n_2 is $\ll$ the total number of variables n in (78).

The question is, how can we exploit the specific structure of (78) in order to accelerate its solution. This is the question we intend to address in this chapter. In section 5.2 we address two well known cases of the outlined situation where we do know how to act — one where there are no linking constraints , and one where there are no linking variables. The general case where both linking variables and linking constraints are present is considered in section 5.3.

[1] By no means is this the only situation which meets our criteria of an *easy to solve* Linear Programming problem. In fact the only general criteria is that both the Primal and the Dual LPs can be solved quickly.

5.2 Two Well Known Cases

5.2.1 Case A: No Linking Constraints

Consider the case where there is no linking constraints, that is, $m_2 = 0$. In this case, we can define the convex function

$$f(x_2) = \min_{x_1} \left\{ c_1^T x_1 + c_2^T x_2 : A^{11} x_1 + A^{12} x_2 \le b_1, \|x_1\|_\infty \le R \right\}$$

and replace the problem of interest with an equivalent problem

$$\min_{x_2 : \|x_2\|_\infty \le R} f(x_2). \tag{80}$$

According to assumption (1), it is relatively easy to check, given x_2, whether $x_2 \in \mathrm{Dom}\, f$, and if so, to compute $f(x_2)$ and a subgradient $f'(x_2)$ of f_2 at x_2. Indeed, to this end one should solve the LP program

$$\min_{x_1} \left\{ c_1^T x_1 : A^{11} x_1 \le b_1 - A^{12} x_2, \|x_1\|_\infty \le R \right\}, \tag{81}$$

which by our assumption is easy. If this problem is solvable, then $x_2 \in \mathrm{Dom}\, f$, and the optimal solution $x_1(x_2)$ to the LP (81), along with an optimal solution $y_1 = y_1(x_2)$ to its associated dual problem

$$\max_{y_1} \left\{ [A^{12} x_2 - b_1]^T y_1 : [A^{11}]^T y_1 + c_1 = 0, y_1 \ge 0 \right\}$$

allows us to compute $f(x_2)$, and $f'(x_2)$ according to

$$f(x_2) = c_1^T x_1(x_2) + c_2^T x_2 = y_1^T [A^{12} x_2 - b_1] + c_2^T x_2, \;\; f'(x_2) = [A^{12}]^T y_1 + c_2.$$

Conversely if the problem (81) is unsolvable (which, since the feasible set of the problem clearly is bounded, can happen only if the problem (81) is infeasible), we will receive, via the General Theorem on Alternative, an infeasibility certificate $y_1 = y_1(x_2)$ such that

$$y_1 \ge 0 \;\&\; y_1^T [A^{12} x_2 - b_1] > \|[A^{11}]^T y_1\|_\infty;$$

in this case, the linear form $\langle [A^{12}]^T y_1, \cdot \rangle$ is nonzero and separates x_2 from the feasible domain X_2 of (80). Thus, (80) is naturally equipped with both a Separation oracle, and a First Order oracle, and as such can be solved by a black box oriented method for convex minimization; by assumption, the dimension of this problem n_2 is $\ll$ the dimension of (78), which can make this outlined approach highly preferable to a direct attack on (78) with an LP solver.

Note that after a feasible ϵ-optimal solution x_2 to (80) is found, we automatically have at our disposal an ϵ-optimal feasible solution $[x_1(x_2); x_2]$ to the problem of interest.

Note that one of the most popular implementations of this scheme is the classical *Benders decomposition*, see, e.g., [5].

5.2.2 Case B: No Linking Variables

Now consider the case where there are no linking variables, that is, $n_2 = 0$. One well-known way to utilize the resulting structure is to use the well-known Dantzig-Wolfe decomposition which is intrinsically linked to the revised Primal Simplex Method, see, e.g., [5]. Another standard way to handle (78) with no linking variables is to use *Lagrangian decomposition*, that is, to associate with (78) its partial dual problem (where we *dualize* the linking constraints) – which we write down in the minimization form

$$\min_{y_2\geq 0} f(y_2),\ f(y_2) = \max_{x_1}\left\{-[c_1 + [A^{21}]^T y_2]^T x_1 : A^{11}x_1 \leq b_1, \|x_1\|_\infty \leq R\right\} + b_2^T y_2. \quad (82)$$

Observe that $f(y_2)$ is an everywhere finite convex function which can be easily equipped with a First Order oracle, since by our assumption the parametric LP problem specifying f is relatively easy to solve. Indeed the LP associated with (82) can be rewritten as $\min\{d^T x_1 : A^{11}x_1 \leq b_1, \|x_1\|_1 \leq R\}$ where $d = c_1 + [A^{12}]^T y_2$. Denoting by $x_1(y_2)$ an optimal solution to the optimization problem

$$\max_{x_1}\left\{-[c_1 + [A^{21}]^T y_2]^T x_1 : A^{11}x_1 \leq b_1, \|x_1\|_\infty \leq R\right\}, \quad (83)$$

we have

$$f(y_2) = -[c_1 + [A^{21}]^T y_2]^T x_1(y_2) + b_2^T y_2,\ f'(y_2) = b_2 - A^{21}x_1(y_2). \quad (84)$$

In many cases we can find a finite upper bound L on, say, the $\|\cdot\|_1$-norm of an optimal solution to (82) and thus reduce this problem to a convex minimization on a simple solid, which we can solve by a black box oriented method (see e.g. section 2.1.1). Since the dimension m_2 of (82) is small when compared to the sizes of the problem of interest (78), this approach can be more attractive than a direct attack on (78) with an LP solver.

Note that the Lagrange decomposition approach requires us to solve the nontrivial problem of recovering a good approximate solution to the problem of interest (78) from a good approximate solution to the problem

$$-\mathrm{Opt}(L) := \min_{y_2\in\mathbf{B}} f(y_2),\ \mathbf{B} = \mathbf{B}_L = \left\{y_2 : y_2 \geq 0, \sum_{i=1}^{m_2}[y_2]_i \leq L\right\} \quad (85)$$

(which is nothing but the problem (82) with an added bound on variables). Observe that by Weak Duality $-\mathrm{Opt}(L) \leq \mathrm{Opt}$ for all L. One of the ways to resolve this problem is to use accuracy certificates; to this end, we have proved the following result:

Proposition 5.2.1. *Assume that $n_2 = 0$ and that (85) is solved by an algorithm with accuracy certificates. By construction of the First Order oracle for the objective of this problem, at every productive step t of this algorithm, the search point being $y_2^t \in \mathbf{B}_L$, we have at our disposal a point*

$$x_1^t = x_1(y_2^t) \in \mathrm{Argmax}_{x_1}\left\{-[c_1 + [A^{21}]^T y_2^t]^T x_1 : A^{11}x_1 \leq b_1, \|x_1\|_\infty \leq R\right\}.$$

Now let τ be a step such that the accuracy certificate ξ^τ associated with the corresponding execution protocol $P_\tau = \left\{\{y_2^t, e_t\}_{t=1}^\tau, I_\tau, J_\tau\right)$ is well defined. Setting

$$\widehat{x}_1^\tau = \sum_{t\in I_\tau} \xi_t^\tau x_1^t,$$

we get an approximate solution to the problem of interest (78) *such that*

$$\begin{aligned} &A^{11}\widehat{x}_1^\tau \le b_1 \ \&\ \|x_1^\tau\|_\infty \le R, \\ &A^{21}\widehat{x}_1^\tau \le b_2 + L^{-1}[\mathrm{Opt} + \varepsilon_{\mathrm{cert}}(\xi^\tau|P_\tau, \mathbf{B}_L) + R\|c_1\|_1]\mathbf{1} \\ &c_1^T\widehat{x}_1^\tau \le \mathrm{Opt} + \varepsilon_{\mathrm{cert}}(\xi^\tau|P_\tau, \mathbf{B}_L), \end{aligned} \tag{86}$$

where $\mathbf{1}$ is the all-ones vector of dimension m_2.

In addition, let $\widetilde{y} = [\widetilde{y}_1; \widetilde{y}_2; \widetilde{y}_+; \widetilde{y}_-] \ge 0$ be the vector of optimal Lagrange multipliers for (78)*, so that*

$$c_1^T x_1 + \widetilde{y}_1^T[A^{11}x_1 - b_1] + \widetilde{y}_2^T[A^{21}x_1 - b_2] + \widetilde{y}_+^T[x_1 - R\mathbf{1}] + \widetilde{y}_-^T[-x_1 - R\mathbf{1}] \equiv \mathrm{Opt}\,\forall x_1. \tag{87}$$

When $\ell := L - \sum_{i=1}^{m_2}[\widetilde{y}_2]_i > 0$, then, in addition to the second relation in (86)*, we have*

$$A^{21}\widehat{x}_1^\tau \le b_2 + \ell^{-1}\varepsilon_{\mathrm{cert}}(\xi^\tau|P_\tau, \mathbf{B}_L)\mathbf{1} \tag{88}$$

Proof. Let $\widehat{x}_1^\tau$ be as stated in Proposition. Then

i. Clearly $A^{11}\widehat{x}_1^\tau \le b_1 \ \&\ \|\widehat{x}_1^\tau\|_\infty \le R$ holds since by construction $\widehat{x}_1^\tau$ is a convex combination of feasible solutions to (83).

ii. We have for every $y_2 \in \mathbf{B}_L$:

$$\begin{aligned} &\varepsilon_{\mathrm{cert}}(\xi^\tau|P_\tau, \mathbf{B}_L) \ge \textstyle\sum_{t=1}^\tau \xi_t\langle e_t, y_2^t - y_2\rangle \ [\text{definition of } \varepsilon_{\mathrm{cert}}] \\ &\ge \textstyle\sum_{t\in I_\tau} \xi_t\langle f'(y_2^t), y_2^t - y_2\rangle \ [\text{since } \langle e_t, y_2^t - y_2\rangle \ge 0 \text{ for } t \notin I_\tau] \\ &= \textstyle\sum_{t\in I_\tau} \xi_t\langle b_2 - A^{21}x_1^t, y_2^t - y_2\rangle \ [\text{see (84)}] \\ &= \textstyle\sum_{t\in I_\tau} \xi_t\left[\langle b_2 - A^{21}x_1^t, y_2^t\rangle - \langle b_2 - A^{21}x_1^t, y_2\rangle\right] \\ &= \textstyle\sum_{t\in I_\tau} \xi_t\left[f(y_2^t) + c_1^T x_1^t - \langle b_2 - A^{21}x_1^t, y_2\rangle\right] \ [\text{see (84)}] \\ &= \textstyle\sum_{t\in I_\tau} \xi_t f(y_2^t) + c_1^T\widehat{x}_1^\tau - \langle b_2 - A^{21}\widehat{x}_1^\tau, y_2\rangle \end{aligned}$$

whence, taking into account that $f(y_2^t) \ge -\mathrm{Opt}(L)$ for $t \in I_\tau$,

$$\langle A^{21}\widehat{x}_1^\tau - b_2, y_2\rangle \le \mathrm{Opt}(L) + \varepsilon_{\mathrm{cert}}(\xi^\tau|P_\tau, \mathbf{B}_L) - c_1^T\widehat{x}_1^\tau. \tag{!}$$

Relation (!) holds true for all $y_2 \in \mathbf{B}_L$; recalling what $\mathbf{B}_L$ is and maximizing the left hand side in (!) over $y_2 \in \mathbf{B}_L$ with $\sum_i (y_2)_i = L$, we get

$$\begin{aligned} &L \max_i [A^{21}\widehat{x}_1^\tau - b_2]_i \le \mathrm{Opt}(L) + \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}_L) - c_1^T \widehat{x}_1^\tau \\ &\le \mathrm{Opt}(L) + \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}_L) - \|c_1\|_1 R \end{aligned}$$

which is nothing but the second relation in (86). Setting in (!) $y_2 = 0$, we get

$$c_1^T \widehat{x}_1^\tau \le \mathrm{Opt}(L) + \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}_L).$$

Taking into account that by weak duality $\mathrm{Opt} \ge \sup\limits_{y_2 \ge 0} [-f(y_2)] \ge \sup\limits_{y_2 \in \mathbf{B}_L} [-f(y_2)] = \mathrm{Opt}(L)$, we arrive at the third relation in (86).

iii. To prove (88), we can assume w.l.o.g. that the set $K = \{i : [A^{21}\widehat{x}_1^\tau - b_2]_i > 0\}$ is nonempty, since otherwise (88) is evident. Setting in (87) $x_1 = \widehat{x}_1^\tau$ and taking into account the first relation in (86), we get

$$c_1^T \widehat{x}_1^\tau + \widetilde{y}_2^T [A^{21}\widehat{x}_1^\tau - b_2] \ge \mathrm{Opt} \ge \mathrm{Opt}(L),$$

so that

$$\mathrm{Opt}(L) - c_1^T \widehat{x}_1^\tau \le \widetilde{y}_2^T [A^{21}\widehat{x}_1^\tau - b_2],$$

which combines with (!) to imply that for all $y_2 \in \mathbf{B}_L$ one has

$$\begin{aligned} &\langle y_2, A^{21}\widehat{x}_1^\tau - b_2 \rangle \le \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}_L) + \widetilde{y}_2^T [A^{21}\widehat{x}_1^\tau - b_2] \\ &\le \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}_L) + [\textstyle\sum_{i=1}^{m_2} [\widetilde{y}_2]_i] \max_i [A^{12}\widehat{x}_1^\tau - b_2]_i. \end{aligned}$$

Let $\mu := \max_i [A^{12}\widehat{x}_1^\tau - b_2]_i = [A^{12}\widehat{x}_1^\tau - b_2]_{i_*}$, so that the above relation reads

$$\langle y_2, b_2 - A^{21}\widehat{x}_1^\tau \rangle \le \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}_L) + \mu \left[\sum_{i=1}^{m_2} [\widetilde{y}_2]_i \right].$$

Setting the i_*-th coordinate of y_2 to the value L, and the remaining coordinates – to the value 0, we get from the latter inequality

$$\mu L \le \varepsilon_{\mathrm{cert}}(\xi^\tau | P_\tau, \mathbf{B}_L) + \mu \left[\sum_{i=1}^{m_2} [\widetilde{y}_2]_i \right],$$

and (88) follows. ∎

5.3 General Case

Now consider the general case, when there are both linking variables and linking constraints. To the best of our knowledge, the only decomposition scheme proposed for this case is "cross

decomposition" originating from T.J. Van Roy [42, 43], see also [45, 13, 14, 15, 16, 36] and references therein. In this scheme, essentially, one alternates iteratively between the Benders and the Lagrange decompositions. To the best of our understanding, no complexity results for this scheme are known. [2]

5.3.1 Assumptions and Approach

Consider the Lagrange function of problem (78), assuming that both linking constraints and linking variables exist. This bilinear function is:

$$\Psi(x_1, x_2, y_1, y_2) = \langle c_1, x_1\rangle + \langle c_2, x_2\rangle + \langle y_1, A^{11}x_1 + A^{12}x_2 - b_1\rangle + \langle y_2, A^{21}x_1 + A^{22}x_2 - b_2\rangle \quad (89)$$

This function leads us naturally to consider the bilinear convex-concave saddle point problem equivalent to problem (78).

$$\inf_{x=[x_1;x_2]\in X_1\times X_2} \sup_{y=[y_1;y_2]\geq 0} \Psi(x_1, x_2, y_1, y_2), \quad (90)$$

where $X_1 = \{x_1 : \|x_1\|_\infty \leq R\}$, and $X_2 = \{x_2 : \|x_2\|_\infty \leq R\}$ are simple solids. Assume, as is usually the case, that we can shrink the y-domain of this saddle point problem, which is a nonnegative orthant, to a direct product $Y = Y_1 \times Y_2$ of two simple solids which are "large enough" to ensure that the saddle point problem

$$\min_{x=[x_1;x_2]\in X_1\times X_2} \max_{y=[y_1;y_2]\in Y_1\times Y_2} \Psi(x_1, x_2, y_1, y_2) \quad (91)$$

is equivalent to (90) and thus – to the LP of interest (or approximates (90) "well enough").

Assumption. At the beginning of this chapter we have assumed that it is easy to solve the LPs of the form $\min_x\{c^T x_1 : A^{11}x_1 \leq b, \|x\|_\infty \leq R\}$. From now on we modify this assumption, specifically, assume that

A: *Given $x_2 \in X_2, y_2 \in Y_2$, it is easy to solve the saddle point problem*

$$\min_{x_1\in X_1} \max_{y_1\in Y_1} \left\{\Psi(x_1, x_2, y_1, y_2) = \langle c + [A^{11}]^T y_1, x_1\rangle - \langle b, y_1\rangle\right\}, \quad \left[c = c_1 + [A^{12}]^T y_2,\ b = b_1 - A^{12}x_2\right] \quad (92)$$

Note that the original assumption is nothing but the modified one with $Y_1 = \mathbf{R}^{m_1}_+$. For our machinery to work, we need Y_1 to be bounded, this is where the modification comes from. Note also that in typical situations where the unmodified assumption holds true, so

[2] This being said, note that the primary motivation and application of cross decomposition is the Mixed Integer version of (78), where the linking variables x_2 are subject to additional constraints of integrality.

is the modified one, provided that Y_1 is chosen properly. For example, this is so in the case considered in section 5.1, where the problem $\min_{x_1\in X_1}\{c^Tx_1 : A^{11}x_1 \le b, \|x_1\|_\infty \le R\}$ is just a collection of a large number of independent of each other LPs, or, which is the same, $x_1 = [x_{11}; ...; x_{1K}]$ is a collection of a large number K of components x_{1k} with dimensions low as compared to n_1, and A^{11} is block-diagonal: $A^{11}x_1 = [A^{11}_1x_{11}; A^{11}_2x_{12}; ...; A^{11}_Kx_{1K}]$, with the dimensions of the blocks $A^{11}_kx_{1k}$ low as compared to m_1. In this case, choosing Y_1 as a large box:

$$Y_1 = \{y = [y_{11}; ...; y_{1K}] : 0 \le y_{1k} \le R_k[1; ...; 1]\}$$

with the blocks y_{1k} in y_1 corresponding to the blocks $A^{11}_kx_{1k}$ in $A^{11}x_1$, the saddle point problem in **A** decomposes into K independent low dimensional bilinear saddle point problems of the form

$$\min_{x_{1k},\|x_{1k}\|_\infty\le R} \max_{y_{1k}:0\le y_{1k}\le R_k[1;...;1]} \left[\langle p_k, x_{1k}\rangle + \langle q_k, y_{1k}\rangle + \langle y_{1k}, A^{11}_kx_{1k}\rangle\right],$$

thus making assumption **A** quite realistic.

Approach. On a closer inspection, Assumption **A** says that it is easy to build the First Order oracle for the (clearly convex-concave) function

$$\widetilde{\Psi}(x_2, y_2) = \min_{x_1\in X_1} \max_{y_1\in Y_1} \Psi(x_1, x_2, y_1, y_2),$$

so that the saddle point problem

$$\min_{x_2\in X_2} \max_{y_2\in Y_2} \widetilde{\Psi}(x_2, y_2) \tag{93}$$

(which, by assumption, is of dimensions much smaller than those of the problem of interest) is well suited for solving by a black-box-oriented method, like NERML or perhaps even the Ellipsoid method[3].

When the sizes of (93) are small as compared to those of (91) and elimination of x_1, y_1, that is, computing the first order information for $\widetilde{\Psi}$, is cheap, the outlined approach can be computationally much more attractive than a direct attack on (90) (or, equivalently, on (78)). Note that with this approach, we in fact *simultaneously eliminate both linking variables and linking constraints* (and do it just once), in a sharp contrast to cross decomposition, where we iteratively eliminate, in an alternating fashion, either linking variables, or linking constraints.

The crucial (and nontrivial) question underlying the outlined approach is how to recover a good solution to the problem of actual interest (91) (or, which is the same, (78)) from

[3]The latter is quite realistic, provided that the sizes n_2 and m_2 are in the range of tens; note that there are meaningful applications where n_2 and m_2 are indeed in the range of tens, while m_1 and n_1 are in the range of many thousands.

a good solution to the induced saddle point problem (93). Our major goal in the sequel is to demonstrate that *when solving* (93) *by an algorithm with accuracy certificates, these certificates allow to convert ϵ-solutions to* (93) *into ϵ-solutions of the saddle point problem of actual interest* (91) *and thus – to ϵ-solutions to the original problem* (78).

5.3.2 Induced Pairs of Saddle Point Problems

We intend to consider a situation which is an extension of the one considered in the previous section. Specifically, we intend to investigate saddle point problems *induced* by larger saddle point problems. The general setting is as follows.

Let $X_i \subset \mathbf{R}^{n_{x_i}}$, and $Y_i \subset \mathbf{R}^{n_{y_i}}$, $i = 1, 2$, be solids. We associate with these solids the sets

$$\begin{gathered} X = X_1 \times X_2 \subset \mathbf{R}^{n_x} = \mathbf{R}^{n_{x_1}} \times \mathbf{R}^{n_{x_2}}, \quad Y = Y_1 \times Y_2 \subset \mathbf{R}^{n_y} = \mathbf{R}^{n_{y_1}} \times \mathbf{R}^{n_{y_2}}, \\ Z = X \times Y \subset \mathbf{R}^{n_z} = \mathbf{R}^{n_x} \times \mathbf{R}^{n_y}. \end{gathered} \tag{94}$$

Let $\Psi(z) : Z \to \mathbf{R}$ be a continuous function; with slight abuse of notation, we shall denote this function also $\Psi(x, y)$, $\Psi(x_1, x_2, y_1, y_2)$, etc. From now on, we make the following

Assumption B. *Ψ is convex in $x \in X$ and concave in $y \in Y$.*

Note that Assumption **B** is automatically satisfied when Ψ is bilinear in x, y, as it is the case in (89) – (91).

Function Ψ gives rise to two functions as follows:

$$\begin{aligned} \Psi_1(x_1, y_1) &= \min_{x_2 \in X_2} \max_{y_2 \in Y_2} \Psi(x_1, x_2, y_1, y_2) : X_1 \times Y_1 \to \mathbf{R}, \\ \Psi_2(x_2, y_2) &= \min_{x_1 \in X_1} \max_{y_1 \in Y_1} \Psi(x_1, x_2, y_1, y_2) : X_2 \times Y_2 \to \mathbf{R} \end{aligned}$$

Since Z is a compact set and Ψ is continuous on Z, these functions are continuous on the respective solids $Z_i = X_i \times Y_i$.

Lemma 5.3.1. *$\Psi_i(x_i, y_i)$ is convex in $x_i \in X_i$ and is concave in $y_i \in Y_i$, $i = 1, 2$.*

Proof. Indeed, we have, that $\Psi_1(x_1, y_1) = \min_{x_2 \in X_2} [\max_{y_2 \in Y_2} \Psi(x_1, x_2, y_1, y_2)]$ is concave in y_1, since the function in brackets $[\,]$ is concave in y_1 due to the concavity of $\Psi(x, y)$ in y. The convexity of $\Psi_1(x_1, y_1)$ in x_1 follows via similar argument from the representation $\Psi_1(x_1, y_1) = \max_{y_2 \in Y_2} [\min_{x_2 \in X_2} \Psi(x_1, x_2, y_1, y_2)]$, where the interchange of min and max is legitimate due to the fact that $\Psi(x_1, x_2, y_1, y_2)$ is convex-concave and continuous in x_2, y_2 and X_2, Y_2 are convex compact sets (from now on, we skip similar justifications of swapping the order of min and max). The proof for $\Psi_2(x_2, y_2)$ follows a symmetric line of arguments.

■

We are about to consider two convex-concave saddle point problems

$$\begin{array}{l}(P): \mathrm{SadVal}_{X_1\times Y_1}(\Psi_1) = \min\limits_{x_1\in X_1}\max\limits_{y_1\in Y_1}\Psi_1(x_1,y_1)\\ (D): \mathrm{SadVal}_{X_2\times Y_2}(\Psi_2) = \min\limits_{x_2\in X_2}\max\limits_{y_2\in Y_2}\Psi_2(x_2,y_2)\end{array} \tag{95}$$

which we call the problems *induced* by the master saddle point problem

$$(M): \mathrm{SadVal}_{X\times Y}(\Psi) = \min_{x=(x_1,x_2)\in X=X_1\times X_2}\ \max_{y=(y_1,y_2)\in Y=Y_1\times Y_2}\Psi(x,y). \tag{96}$$

We start with the following observation:

Proposition 5.3.1. *One has*

$$\mathrm{SadVal}_{X_1\times Y_1}(\Psi_1) = \mathrm{SadVal}_{X_2\times Y_2}(\Psi_2) = \mathrm{SadVal}_{X\times Y}(\Psi) \tag{97}$$

and

$$\mathrm{SadSet}_{X\times Y}(\Psi) \subset \mathrm{SadSet}_{X_1\times Y_1}(\Psi_1)\times \mathrm{SadSet}_{X_2\times Y_2}(\Psi_2), \tag{98}$$

where $\mathrm{SadSet}_{U\times V}(\Psi)$ *is the set of saddle points of the function* $\Psi(\cdot)$ *on* $U\times V$.

Moreover, for every $(x_1,x_2,y_1,y_2)\in Z$ *one has*

$$\varepsilon_{\mathrm{sad}}((x_i,y_i)|\Psi_i,X_i,Y_i) \le \varepsilon_{\mathrm{sad}}((x,y)|\Psi,X,Y),\ i=1,2. \tag{99}$$

Proof. Let (x^*,y^*) be a saddle point of Ψ on $X\times Y$. We have for Ψ_1, that

$$\begin{array}{l}\Psi_1(x_1^*,y_1) = \min_{x_2\in X_2}\max_{y_2\in Y_2}\Psi(x_1^*,x_2,y_1,y_2) = \max_{y_2\in Y_2}\min_{x_2\in X_2}\Psi(x_1^*,x_2,y_1,y_2)\\ \le \max_{y_2\in Y_2}\Psi(x_1^*,x_2^*,y_1,y_2) \le \Psi(x_1^*,x_2^*,y_1^*,y_2^*),\end{array}$$

and

$$\Psi_1(x_1,y_1^*) = \min_{x_2\in X_2}\max_{y_2\in Y_2}\Psi(x_1,x_2,y_1^*,y_2) \ge \min_{x_2\in X_2}\Psi(x_1,x_2,y_1^*,y_2^*) \ge \Psi(x_1^*,x_2^*,y_1^*,y_2^*),$$

so that

$$\forall(x_1\in X_1,y_1\in Y_1): \Psi_1(x_1^*,y_1)\le\Psi(x^*,y^*)\le\Psi_1(x_1,y_1^*). \tag{100}$$

By the standard definition of saddle points as applied to Ψ_1 it follows that $\Psi_1(x_1^*,y_1^*) = \Psi(x^*,y^*)$, (and thus that $\mathrm{SadVal}(\Psi_1) = \mathrm{SadVal}(\Psi)$) which combines with (100) to imply that $(x_1^*,y_1^*)\in\mathrm{SadSet}(\Psi_1)$. By "symmetric" reasoning, $(x_2^*,y_2^*)\in\mathrm{SadSet}(\Psi_2)$ and $\mathrm{SadVal}(\Psi_2) = \mathrm{SadVal}(\Psi)$. Thus (97) and (98) are proved.

Now let $(x_1, x_2, y_1, y_2) \in Z$. We have by (22)

$$\begin{aligned}
&\varepsilon_{\rm sad}((x_1, y_1)|\Psi_1, X_1, Y_1) = \max_{\eta_1 \in Y_1} \Psi_1(x_1, \eta_1) - \min_{\xi_1 \in X_1} \Psi_1(\xi_1, y_1) \\
&= \max_{\eta_1 \in Y_1} \min_{\xi_2 \in X_2} \max_{\eta_2 \in Y_2} \Psi(x_1, \xi_2, \eta_1, \eta_2) \\
&\quad - \min_{\xi_1 \in X_1} \min_{\xi_2 \in X_2} \max_{\eta_2 \in Y_2} \Psi(\xi_1, \xi_2, y_1, \eta_2) \\
&\leq \max_{\eta_1 \in Y_1} \max_{\eta_2 \in Y_2} \Psi(x_1, x_2, \eta_1, \eta_2) \\
&\quad - \min_{\xi_1 \in X_1} \min_{\xi_2 \in X_2} \Psi(\xi_1, \xi_2, y_1, y_2) \\
&= \varepsilon_{\rm sad}((x, y)|\Psi, X, Y),
\end{aligned}$$

and similarly for Ψ_2. (99) is proved. ∎

5.3.3 Recovering Approximate Solutions to the Master Problem: Goal and Assumptions

The goal. Consider a master saddle point problem (96) along with the induced problems (95). By Proposition 5.3.1, (specifically by (99)) we can easily extract good approximate solutions to each of the induced problems from a good approximate solution to the master problem. The question is: To what extent is the opposite true?

Specifically, assume we have at our disposal a first-order method capable of solving to within a desired accuracy one of the induced problems, say problem (P), and our goal is to extract from this solution a good approximate solution to the master problem (and thus, by Proposition 5.3.1, to problem (D) as well). When and how could we achieve this goal?

We will demonstrate that this goal is achievable, provided that

- the first order information used by the algorithm in question satisfies some not too restrictive technical assumptions, and that
- we have at our disposal, not only the approximate solution to (P) to be converted into approximate solutions to (M) and (D), but also an accuracy certificate for this solution.

Preliminaries. We start with some technical issues. From now on, we make

> **Assumption C.** *$\Psi(x_1, x_2, y_1, y_2)$ is not only continuous convex-concave on $Z = X \times Y$, but*
>
> - *Ψ is differentiable in $x_2 \in X_2$ whenever $x_1 \in {\rm int}X_1$, $y_1 \in {\rm int}Y_1$ and $y_2 \in Y_2$, the derivative being continuous in $x \in {\rm int}X_1 \times X_2$ for every $y \in {\rm int}Y_1 \times Y_2$;*

- Ψ *is differentiable in* $y_2 \in Y_2$ *whenever* $x_1 \in \mathrm{int}X_1$, $y_1 \in \mathrm{int}Y_1$ *and* $x_2 \in X_2$, *the derivative being continuous in* $y \in \mathrm{int}Y_1 \times Y_2$ *for every* $x \in \mathrm{int}X_1 \times X_2$.

Under Assumption **C**, for every point $(x = (x_1, x_2), y = (y_1, y_2)) \in Z$ with $(x_1, y_1) \in \mathrm{int}X_1 \times \mathrm{int}Y_1$, function $\Psi(\cdot, y)$ at the point x admits a *regular* subgradient $\Psi'_x(x, y)$ – that is a subgradient whose projection onto the subspace $\mathbf{R}^{n_{x_2}}$ equals the gradient of the continuously differentiable function $\Psi(x_1, \cdot, y)$ at the point x_2.

Indeed, let $x_2^t \in \mathrm{int}X_2$ be such that $x_2^t \to x_2$ as $t \to \infty$, and let g^t be a subgradient of $\Psi(\cdot, y)$ at the point $x^t = (x_1, x_2^t) \in \mathrm{int}X$. By evident reasons, such a subgradient is automatically regular. Besides this, the vectors g^t form a bounded sequence, since $\Psi(\xi_1, \xi_2, y)$ is uniformly in $\xi_1 \in V$ Lipschitz continuous in $\xi_2 \in X_2$, V being a neighbourhood of x_1 with the closure belonging to $\mathrm{int}X_1$, and is uniformly in $\xi_2 \in X_2$ Lipschitz continuous in $\xi_1 \in V$. Passing to a subsequence. we may assume that g^t has a limit g as $t \to \infty$; by construction, g is a subgradient of $\Psi(\cdot, y)$ at the point x, and its projection on $\mathbf{R}^{n_{x_2}}$ is $\lim_{t\to\infty} \nabla_s\big|_{s=x_2^t} \Psi(x_1, s, y) = \nabla_s\big|_{s=x_2} \Psi(x_1, s, y)$, as required from a regular subgradient.

We define similarly the notion of a *regular* supergradient $\Psi'_y(x, y)$ in y (a supergradient of the concave function $\Psi(x, \cdot)$ at the point y such that the projection of this supergradient onto $\mathbf{R}^{n_{y_2}}$ is the gradient of $\Psi(x, y_1, \cdot)$ at the point y_2). Such a supergradient also exists, provided that $(x_1, y_1) \in \mathrm{int}X_1 \times \mathrm{int}Y_1$, and $(x_2, y_2) \in X_2 \times Y_2$.

Remark 5.3.1. *Note that Assumption* **C** *is automatically satisfied when* Ψ *is convex-concave and continuously differentiable (as it is the case, e.g., when* Ψ *is bilinear in* x, y, *cf.* (89), (91)*). In this case, choosing as subgradients of* Ψ *w.r.t.* x *the corresponding partial gradients, and similarly for supergradients of* Ψ *in* y, *we automatically end up with regular sub- and supergradients.*

Lemma 5.3.2. *Given* $(\bar{x}_1, \bar{y}_1) \in \mathrm{int}X_1 \times \mathrm{int}Y_1$, *let* $(\bar{x}_2, \bar{y}_2)$ *be a saddle point of the convex-concave continuous function* $\Psi_{\bar{x}_1,\bar{y}_1}(x_2, y_2) = \Psi(\bar{x}_1, x_2, \bar{y}_1, y_2)$ *on* $X_2 \times Y_2$, *and let* $\Psi'_x(\bar{z})$, $\Psi'_y(\bar{z})$ *be regular sub- and supergradients of* Ψ *in* x *and in* y, *respectively, computed at the point* $\bar{z} = (\bar{x}_1, \bar{x}_2, \bar{y}_1, \bar{y}_2)$. *Let, further,* $\Psi'_{1,x}(\bar{x}_1, \bar{y}_1)$ *be the projection of* $\Psi'_x(\bar{z})$ *onto* $\mathbf{R}^{n_{x_1}}$, *and* $\Psi'_{1,y}(\bar{x}_1, \bar{y}_1)$ *be the projection of* $\Psi'_y(\bar{z})$ *onto* $\mathbf{R}^{n_{y_1}}$. *Then* $\Psi'_{1,x}(\bar{x}_1, \bar{y}_1)$ *is a subgradient of the convex function* $\Psi_1(x_1, \bar{y}_1)$ *of* $x_1 \in X_1$ *at the point* $x_1 = \bar{x}_1$, *and* $\Psi'_{1,y}(\bar{x}_1, \bar{y}_1)$ *is a supergradient of the concave function* $\Psi_1(\bar{x}_1, y_1)$ *of* $y_1 \in Y_1$ *at the point* $y_1 = \bar{y}_1$.

Proof. For $x_1 \in X_1$, we have

$$\begin{aligned}
&\Psi_1(x_1, \bar{y}_1) = \min_{x_2 \in X_2} \max_{y_2 \in Y_2} \Psi(x_1, x_2, \bar{y}_1, y_2) \geq \min_{x_2 \in X_2} \Psi(x_1, x_2, \bar{y}) \\
&\geq \min_{x_2 \in X_2} \left[\Psi(\bar{x}, \bar{y}) + \langle \Psi'_x(\bar{z}), x - \bar{x} \rangle\right] \text{ [since } \Psi(\cdot, \bar{y}) \text{ is convex]} \\
&= \underbrace{\Psi_1(\bar{x}_1, \bar{y}_1)}_{=\Psi(\bar{x}, \bar{y})} + \min_{x_2 \in X_2} \left[\langle \Psi'_{1,x}(\bar{x}_1, \bar{y}_1), x_1 - \bar{x}_1 \rangle + \langle \nabla_s \big|_{s=\bar{x}_2} \Psi(\bar{x}_1, s, \bar{y}), x_2 - \bar{x}_2 \rangle\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[since } \Psi'_x(\bar{z}) \text{ is regular]} \\
&\geq \Psi_1(\bar{x}_1, \bar{y}_1) + \min_{x_2 \in X_2} \left[\langle \Psi'_{1,x}(\bar{x}_1, \bar{y}_1), x_1 - \bar{x}_1 \rangle\right] \\
&\qquad \left[\begin{array}{l} \text{since } \bar{x}_2 \in \operatorname{Argmin}_{x_2 \in X_2} f(x_2),\ f(x_2) = \Psi(\bar{x}_1, x_2, \bar{y}_1, \bar{y}_2), \text{ and } f(x_2) \text{ is differentiable} \\ \text{at } \bar{x}_2, \text{ so that } \langle \nabla_s \big|_{s=\bar{x}_2} \Psi(\bar{x}_1, s, \bar{y}), x_2 - \bar{x}_2 \rangle \geq 0 \text{ whenever } x_2 \in X_2 \end{array}\right] \\
&= \Psi_1(\bar{x}_1, \bar{y}_1) + \langle \Psi'_{1,x}(\bar{x}_1, \bar{y}_1), x_1 - \bar{x}_1 \rangle;
\end{aligned}$$

The concluding inequality says that $\Psi'_{1,x}(\bar{x}_1, \bar{y}_1)$ indeed is a subgradient of $\Psi_1(x_1, \bar{y}_1)$ in $x_1 \in X_1$ evaluated at $x_1 = \bar{x}_1$. The "symmetric" reasoning proves the "supergradient" part of the statement. ∎

5.3.4 Recovering Approximate Solutions to the Master Problem: Construction and Main Result

In the above described situation, let us assume that we have access to Separation oracles for X_1 and Y_1 (and thus – to a Separation oracle for $Z_1 = X_1 \times Y_1$). We also assume that we have access to a Φ-oracle, where $\Phi : \mathrm{int} Z_1 \to \mathbf{R}^{n_{x_1}} \times \mathbf{R}^{n_{y_1}}$ is the monotone mapping associated with the convex-concave saddle point problem

$$\max_{x_1 \in X_1} \min_{y_1 \in Y_1} \Psi_1(x_1, y_1), \tag{101}$$

specifically, as follows:

Given on input $(x_1, y_1) \in \mathrm{int} X_1 \times \mathrm{int} Y_1$, the Φ-oracle

- solves the saddle point problem

$$\min_{\xi_2 \in X_2} \max_{\eta_2 \in Y_2} \Psi(x_1, \xi_2, y_1, \eta_2) \tag{102}$$

  and computes a saddle point (x_2, y_2) of this problem, along with

  - the projection e_{x_1} of the regular subgradient of $\Psi(\xi, y_1, y_2)$ in $\xi \in X$ computed at the point $\xi = (x_1, x_2)$, onto the space $\mathbf{R}^{n_{x_1}}$;

  - the projection $-e_{y_1}$ of the regular supergradient of $\Psi(x_1, x_2, \eta)$ in $\eta \in Y$ computed at the point $\eta = (y_1, y_2)$, onto the space $\mathbf{R}^{n_{y_1}}$.

- returns the pair (x_2, y_2) and the vector $\Phi(x_1, y_1) = (e_{x_1}, e_{y_1})$, thus, by Lemma 5.3.2, reporting the value at (x_1, y_1) of the monotone mapping associated with the saddle point problem (101).

Assume that we have built a τ-point execution protocol $P_\tau = \{(z_1^t, e^t)\}_{t=1}^\tau$, where $z_1^t = (x_1^t, y_1^t)$ are the search points, partitioned into those which are strictly feasible ($z_1^t \in \mathrm{int} Z_1 \Leftrightarrow t \in I_\tau$) and all the remaining search points ($z_1^t \notin \mathrm{int} Z_1 \Leftrightarrow t \in J_\tau$). Also assume that e^t is either $\Phi(z_1^t)$ (this is so when $t \in I_\tau$), or e^t is a nonzero separator of z_1^t and Z_1 (this is so when $t \in J_\tau$). According to the construction of the Φ-oracle, this protocol can be augmented with pairs $z_2^t = (x_2^t, y_2^t)$, $t \in I_\tau$, reported by the Φ-oracle at the productive steps (those from I_τ). Our main result is as follows:

Theorem 5.3.1. *Let $\mathbf{B}$ be a solid containing Z_1, and let P_τ be an execution protocol which admits an accuracy certificate ζ. Given this certificate, let us set*

$$\widehat{x}_i^\tau = \sum_{t \in I_\tau} \zeta_t x_i^t, \ \widehat{y}_i^\tau = \sum_{t \in I_\tau} \zeta_t y_i^t, \ i = 1, 2.$$

Then $(\widehat{x}^\tau, \widehat{y}^\tau) \in Z$ and

$$\varepsilon_{\mathrm{sad}}((\widehat{x}^\tau, \widehat{y}^\tau) | \Psi, X, Y) \le \varepsilon_{\mathrm{cert}}(\zeta | P_\tau, \mathbf{B}), \tag{103}$$

whence, by Proposition 5.3.1, also

$$\varepsilon_{\mathrm{sad}}((\widehat{x}_i^\tau, \widehat{y}_i^\tau) | \Psi_i, X_i, Y_i) \le \varepsilon_{\mathrm{cert}}(\zeta | P_\tau, \mathbf{B}), \ i = 1, 2. \tag{104}$$

Proof. For $t \in I_\tau$, let $z^t = (x_1^t, x_2^t, y_1^t, y_2^t) = (x^t, y^t)$. Recall that for $t \in I_\tau$, $e_{x_1}^t$ is the projection onto $\mathbf{R}^{n_{x_1}}$ of a subgradient $\Psi'_x(x^t, y^t)$ of the function $\Psi(\cdot, y^t)$ computed at the point x^t, and the projection $e_{x_2}^t$ of this subgradient onto $\mathbf{R}^{n_{x_2}}$ is the vector $\nabla_{x_2}\big|_{x_2 = x_2^t} \Psi(x_1^t, x_2, y_1^t, y_2^t)$, whence

$$\langle e_{x_2}^t, x_2 - x_2^t \rangle \ge 0 \ \forall x_2 \in X_2, \tag{105}$$

due to the fact that $\Psi(x_1^t, x_2, y_1^t, y_2^t)$ attains its minimum over $x_2 \in X_2$ at the point x_2^t. Similarly, for $t \in I_\tau$, $e_{y_1}^t$ is the projection onto $\mathbf{R}^{n_{y_1}}$ of a subgradient $-\Psi'_y(x^t, y^t)$ of the function $-\Psi(x^t, \cdot)$ computed at the point y^t, and the projection $e_{y_2}^t$ of this subgradient onto $\mathbf{R}^{n_{y_2}}$ is the vector $-\nabla_{y_2}\big|_{y_2 = y_2^t} \Psi(x_1^t, x_2^t, y_1^t, y_2)$, whence

$$\langle e_{y_2}^t, y_2 - y_2^t \rangle \ge 0 \ \forall y_2 \in Y_2. \tag{106}$$

Let $z = (x_1, x_2, y_1, y_2) = (x, y) \in Z$. We have

$$
\begin{aligned}
&-\varepsilon_{\text{cert}}(\zeta|P_\tau, \mathbf{B}) \le \textstyle\sum_{t=1}^{\tau} \zeta_t \langle e^t, z - z^t\rangle \\
&\le \textstyle\sum_{t\in I_\tau} \zeta_t \langle e^t, z - z^t\rangle \text{ [since } e^t \text{ separates } Z \text{ and } z^t \text{ for } t \in J_\tau] \\
&= \textstyle\sum_{t\in I_\tau} \zeta_t \left[\langle e^t_{x_1}, x_1 - x_1^t\rangle + \langle e^t_{y_1}, y_1 - y_1^t\rangle\right] \\
&\le \textstyle\sum_{t\in I_\tau} \zeta_t \left[\langle \Psi'_x(x^t, y^t), x - x^t\rangle + \langle -\Psi'_y(x^t, y^t), y - y^t\rangle\right] \text{ [by (105), (106)]} \\
&\le \textstyle\sum_{t\in I_\tau} \zeta_t \left[[\Psi(x, y^t) - \Psi(x^t, y^t)] + [\Psi(x^t, y^t) - \Psi(x^t, y)]\right] \text{ [since } \Psi \text{ is convex-concave]} \\
&= \textstyle\sum_{t\in I_\tau} \zeta_t \left[\Psi(x, y^t) - \Psi(x^t, y)\right] \\
&\le \Psi(x, \widehat{y}^\tau) - \Psi(\widehat{x}^\tau, y) \text{ [since } \Psi \text{ is convex-concave]}
\end{aligned}
$$

Thus,

$$\forall (x \in X, y \in Y) : \Psi(\widehat{x}^\tau, y) - \Psi(x, \widehat{y}^\tau) \le \varepsilon_{\text{cert}}(\zeta|P_\tau, \mathbf{B}).$$

Taking the supremum of the left hand side in $(x, y) \in Z$, we arrive at (103), see (23). ∎

REFERENCES

[1] Auslender, A., *Optimization Methodes Numeriques*, Mason, Paris, 1976.

[2] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization*, SIAM, Philadelphia, 2001.

[3] Ben-Tal, A., Margalit, T., and Nemirovski, A., The Ordered Subsets Mirror Descent optimization method with applications to Tomography. *SIAM J. Optim.* **12** (2001), 79-108.

[4] A. Ben-Tal, A. Nemirovski, Non-Euclidean restricted memory level method for large-scale convex optimization. *Math. Progr.* **102**, 407–456, 2005.

[5] D. Bertsimas, J.N. Tsitsiklis, *Introduction to Linear Optimization*. Athena Scientific, 1997.

[6] S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[7] V.P. Bulatov and L.O. Shepot'ko, Method of centers of orthogonal simpexes for solving convex programming problems (in Russian), In: *Methods of Optimization and Their Application*, Nauka, Novosibirsk 1982 .

[8] D. Gabelev (2003), Polynomial time cutting plane algorithms associated with symmetric cones, M.Sc. Thesis in OR and Systems Analysis, Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Technion City, Haifa 32000, Israel
E-print: http://www2.isye.gatech.edu/ nemirovs/Dima.pdf

[9] Francisco Facchinei and Jong-Shi Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems I*, Springer-Verlag, New York, 2003

[10] Fritz John, Extremum problems with inequalities as subsidiary conditions, in: *Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948*, Interscience Publishers, Inc., New York, N. Y., 1948, pp. 187–204.

[11] E.G. Gol'shtein, Primal-dual vlock method for Linear Programming. *Automation and Remote Control* No. 11, 1996.

[12] M. Grötschel, L. Lovasz and A. Schrijver, *The Ellipsoid Method and Combinatorial Optimizatoon*. Springer-Verlag, 1986.

[13] K. Holmberg, On the convergence of cross decomposition, *Math. Progr.* **47:2** (1990), 269-296.

[14] K. Holmberg, Linear mean value cross decomposition: A generalization of the Kornai-Liptak method, *European Journal of Operational Research* **62:1** (1992), 55-73.

[15] K. Holmberg, A convergence proof for linear mean value cross decomposition, *Zeitschrift fur Operations Research* **39:2** (1994), 157-186.

[16] K. Holmberg, Mean value cross decomposition applied to integer programming problems, *European Journal of Operational Research* **97:1** (1997), 124-138.

[17] C. Lemaréchal, Nonsmooth optimization and descent methods. Research Report 78-4, IIASA, Laxenburg, Austria, 1978.

[18] K. Kiwiel, Proximal level bundle method for convex nondifferentable optimization, saddle point problems and variational inequalities. *Math. Progr. Series B* **69** (1995), 89-109.

[19] K.C. Kiwiel, T. Larson, P.O. Lindberg, The efficiency of ballstep subgradient level methods for convex optimization. *Math. of Oper. Res.* **24** (1999), 237-254.

[20] C. Lemaréchal, J.J. Strodiot, and A. Bihain, On a bundle algorithm for nonsmooth optimization. In: O.L. Mangasarian, R.R. Meyer, S.M. Robinson, Eds., *Nonlinear Programming 4* (Academic Press, NY, 1981), 245-282.

[21] C. Lemaréchal, A. Nemirovski, and Yu. Nesterov, New variants of bundle methods. *Math. Progr.* 69:1, 111-148, 1995.

[22] A.Yu. Levin, On an algorithm for the minimization of convex functions (in Russian), *Doklady Akad. Nauk SSSR* **160:6** (1965), 1244–1247. (English translation: Soviet Mathematics Doklady, 6:286–290, 1965.)

[23] R. Mifflin, A modification and an extension of Lemaréchal's algorithm for nonsmooth minimization. *Math. Progr. Study* **17** (1982), 77-90.

[24] A. Nemirovskii and D. Yudin, Efficient methods for large-scale convex problems. (in Russian) - *Ekonomika i Matematicheskie Metody*, **15 :1** (1979) (the journal is translated into English as *Matekon*)

[25] A. Nemirovskii, Efficient iterative algorithms for variational inequalities with monotone operators. (in Russian) - *Ekonomika i Matematicheskie Metody*, **17:2** (1981), 344-359 (the journal is translated into English as *Matekon*)

[26] A. Nemirovskii and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley & Son, 1983.

[27] A. Nemirovski, S. Onn and U.G. Rothblum, Accuracy certificates for computational problems with convex structure. *Math. of Oper. Res.* **35** (2010), 52-78.

[28] A. Nemirovski, Polynomial time methods in Convex Programming, in: J. Renegar, M. Shub and S. Smale, Eds., *The Mathematics of Numerical Analysis*, AMS-SIAM Summer Seminar on Mathematics in Applied Mathematics, July 17– August 11, 1995, Park City, Utah. Lectures in Applied Mathematics, AMS, Providence, 32:543-589, 1996.

[29] A. Nemirovski, Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems, *SIAM J. on Optim.* 15, 229–251, 2004.

[30] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.

[31] Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course*. Kluwer Academic Press, 2004

[32] Yu. Nesterov, Semidefinite relaxation and nonconvex quadratic minimization – *Optimization and Software* **9** (1998), 141-160.

[33] Yu. Nesterov, Smooth minimization of non-smooth functions, *Math. Progr.* 103:1, 127–152, 2005.

[34] D.J. Newman, Location of maximum on unimodal surfaces, *Journ. of the Assoc. for Computing Machinery* 12:11-23, 1965.

[35] R.T. Rockafellar, *Convex Analysis*, Princeton University press, Princeton, New Jersey, 1970.

[36] Han-Suk Sohn and D. Bricker, Cross-Decomposition of the Degree-Constrained Minimum Spanning Tree Problem, *Journal of Systemics, Cybernetics and Informatics* **5:1** (2007).

[37] B.T. Polyak, A general method for solving extremal problems. *Soviet Math. Doklady* **174** (1967), 33-36.

[38] N.Z. Shor, Generalized gradient descent with application to block programming. *Kibernetika* 1967 No. 3 (in Russian).

[39] H. Schramm and J. Zowe, A version of bundle idea for minimizing a non-smooth function: conceptual idea, convergence analysis, numerical results. *SIAM J. Optim.* **2** (1992), 121-152.

[40] N.Z. Shor, Cutting plane method with space dilation for the solution of convex programming problems (in Russian), *Kibernetika*, 1977 No. 1, 94-95.

[41] S.P. Tarasov, L.G. Khachiyan and I.I. Erlikh, The method of inscribed ellipsoids, *Soviet Mathematics Doklady* **37** (1988), 226–230.

[42] T. J. Van Roy, Cross decomposition for mixed integer programs, *Mathematical Programming* **25** (1983), 46-63.

[43] T. J. Van Roy, A Cross decomposition algorithm for capacitated facility location, *Operations Research* **34** (1986), 145-163.

[44] B. Yamnitsky and L. Levin, An Old Linear Programming Algorithm Runs in Polynomoial Time, In: 23rd Annual Symposium on Foundations of Computer Science, IEEE, New York, 327-328, 1982.

[45] Chun-Beon Yooa and Dong-Wan Tchaa, A cross decomposition procedure for the facility location problem with a choice of facility type, *Computers & Industrial Engineering* **10:4** (1986), 283-290.

[46] D. Yudin and A. Nemirovskii, Informational complexity and effective methods of solution for convex extremal problems (in Russian), *Ekonomika i Matematicheskie Metody* **12:2** (1976), 357–369 (translated into English as *Matekon*: Transl. Russian and East European Math. Economics **13** 25–45, Spring '77).

VITA

Bruce Cox graduated from the Worcester Polytechnic Institute, with a Bachelor of Science in Mathematics 1999. Upon graduation he was commissioned into the United States Air Force. He attended the University of Colorado at Colorado Springs for three semesters prior to moving to Richmond, Virgina to finish his Masters in Mathematics at Virginia Commonwealth University in 2004. Shortly afterwards he was selected to teach at the Air Force Institute of Technology and attended Georgia Tech from 2006 to 2011 graduating with a PhD in Operations Research.