1 First Order Methods for Nonsmooth Convex Large-Scale Optimization, I: General Purpose Methods

Anatoli Juditsky Anatoli.Juditsky@imag.fr
Laboratoire Jean Kuntzmann , Université J. Fourier
B. P. 53 38041 Grenoble Cedex, France

Arkadi Nemirovski nemirovs@isye.gatech.edu
School of Industrial and Systems Engineering, Georgia Institute of Technology
765 Ferst Drive NW, Atlanta Georgia 30332, USA

We discuss several state-of-the-art computationally cheap, as opposed to the polynomial time Interior Point algorithms, first order methods for minimizing convex objectives over "simple" large-scale feasible sets. Our emphasis is on the general situation of a nonsmooth convex objective represented by deterministic/stochastic First Order oracle and on the methods which, under favorable circumstances, exhibit (nearly) dimension-independent convergence rate.

1.1 Introduction

At present, essentially the entire Convex Programming is within the grasp of polynomial time Interior Point methods (IPMs) capable to solve convex programs to high accuracy at a low iteration count. However, the iteration cost of all known polynomial methods grows nonlinearly with problem's design dimension n (# of decision variables), something like n^3. As a result, as the design dimension grows, polynomial time methods eventually become impractical — roughly speaking, a single iteration "lasts forever." What in

fact "eventually" means, depends on problem's structure. For instance, typical Linear Programming programs of decision-making origin have extremely sparse constraint matrices, and IPMs are able to solve in reasonable time programs of this type with tens and hundreds of thousands variables and constraints. In contrast to this, Linear Programming programs arising in Machine Learning and Signal Processing often have dense constraint matrices. Such programs with "just" few thousands variables and constraints can become very difficult for an IPM. At the present level of our knowledge, the methods of choice when solving convex programs which, because of their size, are beyond the "practical grasp" of IPMs, are the *First Order methods* (FPMs) with computationally cheap iterations. In this chapter, we present several state-of-the-art FOMs for large-scale convex optimization, focusing on the most general *nonsmooth unstructured* case, where the convex objective f to be minimized can be nonsmooth and is represented by a "black box" – a routine capable to compute the values and subgradients of f.

First Order methods: limits of performance. We start with explaining what can and what can*not* be expected from FOMs, restricting ourselves for the time being with convex programs of the form

$$\mathrm{Opt}(f) = \min_{x \in \mathcal{X}} f(x), \tag{1.1}$$

where $\mathcal{X}$ is a compact convex subset of $\mathbb{R}^n$, and f is known to belong to a given family $\mathcal{F}$ of convex and (at least) Lipschitz continuous functions on $\mathcal{X}$. Formally, a FOM is an algorithm $\mathcal{B}$ which knows in advance what are $\mathcal{X}$ and $\mathcal{F}$, but does not know what exactly is $f \in \mathcal{F}$. It is restricted to "learn" f via subsequent calls to a *First Order oracle* — a routine which, given on input a point $x \in \mathcal{X}$, returns on output a value $f(x)$ and a (sub)gradient $f'(x)$ of f at x (informally speaking, this setting implicitly assumes that $\mathcal{X}$ is "simple" (like box, or ball, or standard simplex), while f can be complicated). Specifically, as applied to a particular objective $f \in \mathcal{F}$ and given on input a required accuracy $\epsilon > 0$, the method $\mathcal{B}$, after generating a finite sequence of *search points* $x_t \in \mathcal{X}$, $t = 1, 2, ...$, where the First Order oracle is called, terminates and outputs an approximate solution $\widehat{x} \in \mathcal{X}$ which should be ϵ-optimal: $f(\widehat{x}) - \mathrm{Opt}(f) \leq \epsilon$. In other words, the method itself is a collection of rules for generating subsequent search points, identifying the terminal step, and building the approximate solution. These rules, in principle, can be arbitrary, with the only limitation of being *non-anticipating*, meaning that the "output" of a rule is uniquely defined by $\mathcal{X}$ and the first order information on f accumulated before the rule is applied. As a result, for a given $\mathcal{B}$ and $\mathcal{X}$, x_1 is independent of

f, x_2 depends solely on $f(x_1), f'(x_1)$, and so on. Similarly, the decision to terminate after a particular number t of steps, same as the resulting approximate solution $\widehat{x}$, are uniquely defined by the first order information $f(x_1), f'(x_1)$, ..., $f(x_t), f'(x_t)$ accumulated in course of these t steps. Limits of performance of FOMs are given by *Information-Based Complexity Theory* which says what, for given $\mathcal{X}, \mathcal{F}, \epsilon$, may be the minimal number of steps of a FOM solving all problems (1.1) with $f \in \mathcal{F}$ within accuracy ϵ. Here are several instructive examples (see Nemirovski and Yudin, 1983):

(a) Let $\mathcal{X} \subset \{x \in \mathbb{R}^n : \|x\|_p \leq R\}$, where $p \in \{1, 2\}$, and let $\mathcal{F} = \mathcal{F}_p$ be comprised of all convex functions f which are Lipschitz continuous, with a given constant L, w.r.t. $\|\cdot\|_p$. When $\mathcal{X} = \{x \in \mathbb{R}^n : \|x\|_p \leq R\}$, the number N of steps of *any* FOM capable to solve every problem from the just outlined family within accuracy ϵ is *at least* $O(1)\min[n, L^2R^2/\epsilon^2]$ [1]. When $p = 2$, this lower complexity bound remains true when $\mathcal{F}$ is restricted to be the family of all functions of the type $f(x) = \max\limits_{1\leq i\leq n} [\epsilon_i L x_i + a_i]$ with $\epsilon_i = \pm 1$. Moreover, the bound is "nearly achievable:" whenever $\mathcal{X} \subset \{x \in \mathbb{R}^n : \|x\|_p \leq R\}$, there exist quite transparent (and simple in implementation when $\mathcal{X}$ is simple) FOMs capable to solve all problems (1.1) with $f \in \mathcal{F}_p$ within accuracy ϵ in $O(1)(\ln(n))^{2/p-1}L^2R^2/\epsilon^2$ steps.
It should be stressed that outlined nearly dimension-independent performance of FOMs heavily depends on the assumption $p \in \{1, 2\}$ [2]. With p set to $+\infty$ (i.e., when minimizing Lipschitz continuous, with constant L w.r.t. $\|\cdot\|_\infty$, convex functions over the box $\mathcal{X} = \{x \in \mathbf{R}^n : \|x\|_\infty \leq R\}$), the lower and the upper complexity bounds are $O(1)n\ln(LR/\epsilon)$ provided that $LR/\epsilon \geq 2$; these bounds heavily depend on problem's dimension.

(b) Let $\mathcal{X} = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$, and let $\mathcal{F}$ be comprised of all differentiable convex functions with Lipschitz continuous, with constant L w.r.t. $\|\cdot\|_2$, gradient. Then the number N of steps of any FOM capable to solve every problem from the just outlined family within accuracy ϵ is *at least* $O(1)\min[n, \sqrt{LR^2/\epsilon}]$. This lower complexity bound remains true when $\mathcal{F}$ is restricted to be the family of convex quadratic forms $\frac{1}{2}x^TAx + b^Tx$ with positive semidefinite symmetric matrices A of spectral norm (maximal singular value) not exceeding L. Here again the lower complexity bound is nearly achievable: whenever $\mathcal{X} \subset \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$, there exists a simple in implementation when $\mathcal{X}$ is simple (although by far not transparent) FOM – the famous *Nesterov's optimal algorithm for smooth convex minimization* (Nesterov, 1983, 2005) which allows to solve within accuracy ϵ all problems

1. From now on, all $O(1)$'s are appropriate positive *absolute* constants.
2. In fact, it can be relaxed to $1 \leq p \leq 2$.

(1.1) with $f \in \mathcal{F}$ in $O(1)\sqrt{LR^2/\epsilon}$ steps.

(c) Let $\mathcal{X}$ be as in (b), and let $\mathcal{F}$ be comprised of all functions of the form $f(x) = \|Ax - b\|_2$, where the spectral norm of A (which is not positive semidefinite anymore) does not exceed a given L. Let us slightly extend the "power" of the First Order oracle and assume that at a step of a FOM we observe b (but not A) and are allowed to carry out $O(1)$ matrix-vector multiplications involving A and A^T. In this case, the number of steps of any method capable to solve all problems in question within accuracy ϵ is at least $O(1)\min[n, LR/\epsilon]$, and there exists a method (specifically, Nesterov's optimal algorithm as applied to the quadratic form $\|Ax-b\|_2^2$) which achieves the desired accuracy in $O(1)LR/\epsilon$ steps.

The outlined results bring us both bad and good news on FOMs as applied to large-scale convex programs. The bad news is that *unless the number of steps of the method exceeds problem's design dimension* n (which is of no interest when n is really large) *and without imposing severe additional restrictions on the objectives to be minimized, a FOM can exhibit only sublinear rate of convergence*, specifically, denoting by t the number of steps, the rate $O(1)(\ln(n))^{1/p-1/2}LR/t^{1/2}$ in the case of (a) (better than nothing, but really slow), $O(1)LR^2/t^2$ in the case of (b) (much better, but alas — simple $\mathcal{X}$ along with smooth f is a rare commodity), and $O(1)LR/t$ in the case of (c) ("in-between" (a) and (b)). As a consequence, *FOMs are poorly suited for building high-accuracy solutions to large-scale convex problems.*

The good news is that *for problems with favorable geometry* (e.g., those in (a) – (c)), *good FOMs exhibit dimension-independent, or nearly so, rate of convergence*, which is of paramount importance in large-scale applications. Another good news (not declared explicitly in the above examples) is that *when $\mathcal{X}$ is simple, typical FOMs have cheap iterations — modulo computations "hidden" in the oracle, an iteration costs just $O(\dim \mathcal{X})$ a.o.* The bottom line is that *FOMs are well suited for finding medium-accuracy solutions to large-scale convex problems*, at least when the latter possess "favorable geometry."

Another conclusion of the presented results is that the limits of performance of FOMs heavily depend on the size R of the feasible domain and on the Lipschitz constant L (of f in the case of (a), and of f' in the case of (b)). This is in a sharp contrast to IPM's, where the complexity bounds depend *logarithmically* on the magnitudes of an optimal solution and of the data (the analogies or R, L, respectively), which, practically speaking, allows to handle problems with unbounded domains (one may impose an upper bound 10^6, or 10^{100} on the variables) and not to bother much of how the data is

scaled[3]. Severe dependence of the complexity of FOMs on L and R implies a number of important consequences. In particular:

• Boundedness of $\mathcal{X}$ is of paramount importance, at least theoretically. In this respect, unconstrained settings, like in Lasso: $\min_x\{\lambda\|x\|_1+\|Ax-b\|_2^2\}$ are less preferable than their "bounded domain" counterparts, like $\min\{\|Ax-b\|_2 : \|x\|_1 \le R\}$,[4] in full accordance with common sense – however difficult it is to find a needle in a haystack, a small haystack in this respect is better than a large one!

• For a given problem (1.1), the size R of the feasible domain and the Lipschitz constant L of the objective depend on the norm $\|\cdot\|$ used to quantify these quantities: $R = R_{\|\cdot\|}$, $L = L_{\|\cdot\|}$. When $\|\cdot\|$ varies, the product $L_{\|\cdot\|}R_{\|\cdot\|}$ (this product is all that matters) changes[5], and this phenomenon should be taken into account when choosing a FOM for a particular problem.

What is ahead. Literature on FOMs, which always was huge, in the latest years is growing in a somehow explosive fashion — partly, due to rapidly increasing demand for large-scale optimization, and partly by endogenous reasons stemming primarily from discovering ways (Nesterov, 2005) to accelerate FOMs by exploiting problem's structure (for more details on the latter subject, see chapter 2). Even a brief overview of this literature in a single document would be completely unrealistic. Our primary "selection criteria" were (a) to focus on techniques for large-scale *nonsmooth* convex programs (these are the problems arising in most applications known to us), (b) to restrict ourselves with FOMs possessing state-of-the art (in some cases – even provably optimal) non-asymptotic efficiency estimates, and (c) possibility for self-contained presentation of the methods given space limitations. Last, but not least, we preferred to focus on the situations of which we have first-hand (or nearly so) knowledge. As a result, our presentation of FOMs being instructive (at least, so we hope), is definitely incomplete. As about "citation policy", we restrict ourselves with referring to papers directly related to what we are presenting, with no attempt to give even a "nearly exhaustive" list of references to FOM literature. We apologize in advance

3. In IPMs, scaling of the data affects stability of the methods w.r.t. rounding errors, but this is another story.

4. We believe that the desire to end up with unconstrained problems stems from the common belief that unconstrained convex minimization is simpler than the constrained one. To the best of our understanding, this belief is somehow misleading, and the actual distinction is between optimization over simple and over "sophisticated" domains; what is simple, this depends on the method in question.

5. For example, the ratio $[L_{\|\cdot\|_2}R_{\|\cdot\|_2}]/L_{\|\cdot\|_1}R_{\|\cdot\|_1}$ can be as small as $1/\sqrt{n}$ and as large as $\sqrt{n}$

for potential omissions even in this "reduced list."

In this chapter, we focus on the simplest general-purpose FOMs, *Mirror Descent* (MD) *methods* aimed at solving nonsmooth convex minimization problems, specifically, general-type problems (1.1) (section 1.2), problems (1.1) with strongly convex objectives (section 1.4), convex problems with functional constraints $\min_{x\in X}\{f_0(x) : f_i(x) \le 0, 1 \le i \le m\}$ (section 1.3), and stochastic versions of problems (1.1), where the First Order oracle is replaced with its stochastic counterpart providing unbiased random estimates of the subgradients of the objective rather than the subgradients themselves (section 1.5). Finally, section 1.6 presents extensions of the Mirror Descent scheme from problems of convex minimization to the convex-concave saddle point problems.

As it was already said, this chapter is devoted to "general purpose" FOMs, meaning that the methods in question are fully black-box-oriented – they do not assume any a priori knowledge of the structure of the objective (and the functional constraints, if any) aside of convexity and Lipschitz continuity. By itself, this generality is somehow redundant: convex programs arising in applications always possess a lot of known in advance structure, and utilizing a priori knowledge of this structure can accelerate dramatically the solution process. Acceleration FOMs by utilizing problem's structure is the subject of the subsequent chapter 2.

1.2 Mirror Descent Algorithm: Minimizing over Simple Set

1.2.1 Problem of interest

We primarily focus on solving optimization problem of the form

$$\mathrm{Opt} = \min_{x\in X} f(x), \tag{1.2}$$

where $X \subset E$ is a closed convex set in a finite-dimensional Euclidean space E, and $f : X \to \mathbb{R}$ is a Lipschitz continuous convex function represented by a *First Order oracle*. This oracle is a routine which, given on input a point $x \in X$, returns the value $f(x)$ and a subgradient $f'(x)$ of f at x. We always assume that $f'(x)$ is bounded on X. We assume also that (1.2) is solvable.

1.2.2 Mirror Descent setup

We set up the MD method with two entities:

- a norm $\|\cdot\|$ on the space E embedding X, and $\|\cdot\|_*$ the conjugate norm

on E^*: $\|\xi\|_* = \max_x \{\langle \xi, x\rangle : \|x\| \le 1\}$;

▪ a *distance-generating function* (d.-g.f. for short) *for* $\mathcal{X}$ *compatible with the norm* $\|\cdot\|$, that is, a continuous convex function $\omega(x) : \mathcal{X} \to \mathbb{R}$ such that

— $\omega(x)$ admits a selection $\omega'(x)$ of subgradient which is continuous on the set $\mathcal{X}^o = \{x \in \mathcal{X} : \partial\omega(x) \neq \emptyset\}$;

— $\omega(\cdot)$ is strongly convex, with modulus 1, w.r.t. $\|\cdot\|$:

$$\forall (x, x' \in \mathcal{X}^o) : \langle \omega'(x) - \omega'(x'), x - x'\rangle \ge \|x - y\|^2. \tag{1.3}$$

For $x \in \mathcal{X}^o$, $u \in \mathcal{X}$ let

$$V_x(u) = \omega(u) - \omega(x) - \langle \omega'(x), u - x\rangle. \tag{1.4}$$

Denote $x_c = \mathrm{argmin}_{u\in\mathcal{X}}\omega(u)$ (the existence of a minimizer is given by continuity and strong convexity of ω on $\mathcal{X}$ and by closedness of $\mathcal{X}$, and its uniqueness – by strong convexity of ω). When $\mathcal{X}$ is bounded, we define "$\omega(\cdot)$-diameter" $\Omega = \max_{u\in\mathcal{X}} V_{x_c}(u) \le \max_{\mathcal{X}} \omega(u) - \min_{\mathcal{X}} \omega(u)$ of $\mathcal{X}$. Given $x \in \mathcal{X}^o$, we define the *prox-mapping* $\mathrm{Prox}_x(\xi) : E \to \mathcal{X}^o$ as follows:

$$\mathrm{Prox}_x(\xi) = \mathrm{argmin}_{u\in\mathcal{X}} \{\langle \xi, u\rangle + V_x(u)\} \tag{1.5}$$

From now on we make the

Simplicity assumption: $\mathcal{X}$ and ω are simple and "fit" each other, specifically, given $x \in X^o$ and $\xi \in E$, it is easy to compute $\mathrm{Prox}_x(\xi)$.

1.2.3 Basic Mirror Descent algorithm

The MD algorithm associated with the outlined setup, as applied to problem (1.2), is the recurrence

$$\begin{array}{ll} (a) & x_1 = \mathrm{argmin}_{x\in\mathcal{X}}\, \omega(x) \\ (b) & x_{t+1} = \mathrm{Prox}_{x_t}(\gamma_t f'(x_t)),\ t = 1, 2, ... \\ (c) & x^t = \left[\sum_{\tau=1}^t \gamma_\tau\right]^{-1} \sum_{\tau=1}^t \gamma_\tau x_\tau \\ (d) & \widehat{x}^t = \mathrm{argmin}_{x\in\{x_1,...,x_t\}} f(x) \end{array} \tag{1.6}$$

Here x_t are subsequent *search points*, and x^t (or $\widehat{x}^t$ — the error bounds to follow work for both these choices) are subsequent *approximate solutions* generated by the algorithm; note that $x_t \in \mathcal{X}^o$ and $x^t, \widehat{x}^t \in \mathcal{X}$ for all t.

The convergence properties of MD stem from the following simple observation:

Proposition 1.1. *Suppose that f is Lipschitz continuous on $\mathcal{X}$ with $L := \sup_{x\in\mathcal{X}} \|f'(x)\|_* < \infty$. Let $\overline{f}_t = \max[f(x^t), f(\widehat{x}^t)]$. Then*

(i) *for all* $u \in \mathcal{X}$, $t \geq 1$ *one has*

$$\begin{array}{rcl} \sum_{\tau=1}^{t} \gamma_\tau \langle f'(x_\tau), x_\tau - u\rangle & \leq & V_{x_1}(u) + \frac{1}{2}\sum_{\tau=1}^{t} \gamma_\tau^2 \|f'(x_\tau)\|_*^2 \\ & \leq & V_{x_1}(u) + \frac{L^2}{2}\sum_{\tau=1}^{t}\gamma_\tau^2, \end{array} \tag{1.7}$$

As a result, for all $t \geq 1$,

$$\overline{f}_t - \mathrm{Opt} \leq \epsilon_t := \frac{V_{x_1}(x_*) + \frac{L^2}{2}\sum_{\tau=1}^{t}\gamma_\tau^2}{\sum_{\tau=1}^{t}\gamma_\tau}, \tag{1.8}$$

where x_* *is an optimal solution to* (1.2). *In particular, in the "divergent series case"* $\gamma_t \to 0$, $\sum_{\tau=1}^{t}\gamma_\tau \to +\infty$ *as* $t \to \infty$, *the algorithm converges:* $\overline{f}_t - \mathrm{Opt} \to 0$ *as* $t \to \infty$. *Moreover, with the stepsizes*

$$\gamma_t = \gamma/[\|f'(x_t)\|_*\sqrt{t}]$$

for all t *one has*

$$\overline{f}_t - \mathrm{Opt} \leq O(1)\left[\frac{V_{x_1}(x_*)}{\gamma} + \frac{\ln(t+1)\gamma}{2}\right] L t^{-1/2}. \tag{1.9}$$

(ii) *Let* $\mathcal{X}$ *be bounded, so that the "*$\omega(\cdot)$*-diameter"* Ω *of* $\mathcal{X}$ *is finite. Then for every number* N *of steps, the* N*-step MD algorithm with constant stepsizes*

$$\gamma_t = \frac{\sqrt{2\Omega}}{L\sqrt{N}}, \quad 1 \leq t \leq N. \tag{1.10}$$

ensures that

$$\begin{array}{c} \underline{f}_N = \min_{u\in\mathcal{X}} \frac{1}{N}\sum_{\tau=1}^{N}[f(x_\tau) + \langle f'(x_\tau), u - x_\tau\rangle] \leq \mathrm{Opt}, \\ \overline{f}_N - \mathrm{Opt} \leq \overline{f}_N - \underline{f}_N \leq \frac{\sqrt{2\Omega}L}{\sqrt{N}} \end{array} \tag{1.11}$$

In other words, the quality of approximate solutions (x^N *or* $\widehat{x}^N$) *can be certified by the easy-to-compute on-line lower bound* $\underline{f}_N$ *on* Opt, *and the certified level of non-optimality of the solutions can be only better than the one given by the worst-case upper bound in the right hand side of* (1.11).

Proof. From the definition of the prox-mapping,

$$x_{\tau+1} = \operatorname*{argmin}_{z\in\mathcal{X}} \left\{\langle \gamma_\tau f'(x_\tau) - \omega'(x_\tau), z\rangle + \omega(z)\right\},$$

whence, by optimality conditions,

$$\langle \gamma_\tau f'(x_\tau) - \omega'(x_\tau) + \omega'(x_{\tau+1}), u - x_{\tau+1}\rangle \geq 0 \ \forall u \in \mathcal{X}.$$

When rearranging terms, this inequality can be rewritten as

$$\begin{aligned}\gamma_\tau\langle f'(x_\tau), x_\tau - u\rangle &\le [\omega(u) - \omega(x_\tau) - \langle\omega'(x_\tau), u - x_\tau\rangle]\\ &\quad -[\omega(u) - \omega(x_{\tau+1}) - \langle\omega'(x_{\tau+1}), u - x_{\tau+1}\rangle]\\ &\quad +\gamma_\tau\langle f'(x_\tau), x_\tau - x_{\tau+1}\rangle\\ &\quad -[\omega(x_{\tau+1}) - \omega(x_\tau) - \langle\omega'(x_\tau), x_{\tau+1} - x_\tau\rangle]\\ &= V_{x_\tau}(u) - V_{x_{\tau+1}}(u) + \underbrace{[\gamma_\tau\langle f'(x_\tau), x_\tau - x_{\tau+1}\rangle - V_{x_\tau}(x_{\tau+1})]}_{\delta_\tau}. \quad (1.12)\end{aligned}$$

From strong convexity of V_{x_τ} it follows that

$$\begin{aligned}\delta_\tau &\le \gamma_\tau\langle f'(x_\tau), x_\tau - x_{\tau+1}\rangle - \tfrac{1}{2}\|x_\tau - x_{\tau+1}\|^2\\ &\le \gamma_\tau\|f'(x_\tau)\|_*\|x_\tau - x_{\tau+1}\| - \tfrac{1}{2}\|x_\tau - x_{\tau+1}\|^2\\ &\le \max_s[\gamma_\tau\|f'(x_\tau)\|_* s - \tfrac{1}{2}s^2] = \tfrac{\gamma_\tau^2}{2}\|f'(x_\tau)\|_*^2,\end{aligned}$$

and we get

$$\gamma_\tau\langle f'(x_\tau), x_\tau - x_{\tau+1}\rangle \le V_{x_\tau}(u) - V_{x_{\tau+1}}(u) + \gamma_\tau^2\|f'(x_\tau)\|_*^2/2. \quad (1.13)$$

Summing up these inequalities over $\tau = 1, ..., t$ and taking into account that $V_x(u) \ge 0$, we arrive at (1.7). With $u = x_*$, (1.7), when taking into account that $\langle f'(x_\tau), x_\tau - x_*\rangle \ge f(x_\tau) - \mathrm{Opt}$ and setting $f^t = [\sum_{\tau=1}^t \gamma_\tau]^{-1}\sum_{\tau=1}^t \gamma_\tau f(x_\tau)$, results in

$$f^t - \mathrm{Opt} \le \frac{V_{x_1}(x_*) + L^2\left[\sum_{\tau=1}^t \gamma_\tau^2\right]/2}{\sum_{\tau=1}^t \gamma_\tau}.$$

Since, clearly, $\overline{f}_t = \max[f(x^t), f(\widehat{x}^t)] \le f^t$, we have arrived at (1.8). This inequality straightforwardly implies the remaining results of (i).

To prove (ii), note that by definition of Ω and due to $x_1 = \mathrm{argmin}_{\mathcal{X}}\,\omega$, (1.7) combines with (1.10) to imply that

$$f^N - \underline{f}_N = \max_{u\in\mathcal{X}}\left[f^N - \frac{1}{N}\sum_{\tau=1}^N[f(x_\tau) + \langle f'(x_\tau), u - x_\tau\rangle\right] \le \frac{\sqrt{2\Omega}L}{\sqrt{N}}. \quad (1.14)$$

Since f is convex, the function $\frac{1}{N}\sum_{\tau=1}^N[f(x_\tau) + \langle f'(x_\tau), u - x_\tau\rangle]$ underestimates $f(u)$ everywhere on $\mathcal{X}$, that is, $\underline{f}_N \le \mathrm{Opt}$, and, as we have seen, $f^N \ge \overline{f}_N$, and therefore (ii) follows from (1.14) □

1.3 Problems with Functional Constraints

The MD algorithm can be easily extended from the case of problem (1.2) to the case of problem

$$\mathrm{Opt} = \min_{x\in\mathcal{X}} \{f_0(x) : f_i(x) \le 0,\ 1 \le i \le m\}, \tag{1.15}$$

where f_i, $0 \le f_i \le m$, are Lipschitz continuous convex functions on $\mathcal{X}$ given by First Order oracle which, given on input $x \in \mathcal{X}$, returns the values $f_i(x)$ and subgradients $f_i'(x)$ of f_i at x, with bounded on $\mathcal{X}$ selections of the subgradients $f_i'(\cdot)$. Consider the N-step algorithm as follows:

1. *Initialization:* Set $x_1 = \mathrm{argmin}_{\mathcal{X}}\, \omega$

2. *Step t, $1 \le t \le N$:* Given $x_t \in \mathcal{X}$, call the First Order oracle, x_t being the input, and check whether

$$f_i(x_t) \le \gamma \|f_i'(x_t)\|_*,\ i = 1, ..., m. \tag{1.16}$$

If it is the case ("productive step"), set $i(t) = 0$, otherwise ("non-productive step") choose $i(t) \in \{1, ..., m\}$ such that $f_{i(t)}(x) > \gamma \|f_{i(t)}'(x_t)\|_*$. Set

$$\gamma_t = \gamma / \|f_{i(t)}'(x_t)\|_*,\ x_{t+1} = \mathrm{Prox}_{x_t}(\gamma_t f_{i(t)}'(x_t)).$$

When $t < N$, loop to step $t + 1$.

3. *Termination:* After N steps are executed, output, as approximate solution $\widehat{x}^N$, the best (with the smallest value of f_0) of the points x_t associated with productive steps t; if there were no productive steps, claim (1.15) infeasible.

Proposition 1.2. *Let $\mathcal{X}$ be bounded. Given integer $N \ge 1$, let us set $\gamma = \sqrt{2\Omega}/\sqrt{N}$. Then*

(i) *If (1.15) is feasible, $\widehat{x}^N$ is well defined;*

(ii) *Whenever $\widehat{x}^N$ is well defined, one has*

$$\begin{array}{l} \max\left[f_0(\widehat{x}^N) - \mathrm{Opt}, f_1(\widehat{x}^N), ..., f_m(\widehat{x}^N)\right] \le \gamma L = \frac{\sqrt{2\Omega}L}{\sqrt{N}}, \\ L = \max_{0\le i\le m} \sup_{x\in\mathcal{X}} \|f_i'(x)\|_*. \end{array} \tag{1.17}$$

Proof. By construction, when $\widehat{x}^N$ is well defined, it is some x_t with productive t, whence $f_i(\widehat{x}^N) \le \gamma L$ for $1 \le i \le m$ by (1.16). It remains to verify that when (1.15) is feasible, $\widehat{x}^N$ is well defined and $f_0(\widehat{x}^N) \le \mathrm{Opt} + \gamma L$. Assume that it is not the case, whence at every productive step t, if any, we have $f_0(x_t) - \mathrm{Opt} > \gamma \|f_0'(x_t)\|_*$. Let x_* be an optimal solution to (1.15). Exactly the same reasoning as in the proof of Proposition 1.1 yields the following

analogy of (1.7) (with $u = x_*$):

$$\sum_{t=1}^N \gamma_t \langle f'_{i(t)}(x_t), x_t - x_*\rangle \le \Omega + \frac{1}{2}\sum_{t=1}^N \gamma_t^2 \|f'_{i(t)}(x_t)\|_*^2 = 2\Omega. \quad (1.18)$$

Now, when t is non-productive, we have $\gamma_t\langle f'_{i(t)}, x_t - x_*\rangle \ge \gamma_t f_{i(t)}(x_t) > \gamma^2$, the concluding inequality being given by the definition of $i(t)$ and γ_t. When t is productive, we have $\gamma_t\langle f'_{i(t)}(x_t), x_t - x_*\rangle = \gamma_t\langle f'_0(x_t), x_t - x_*\rangle \ge \gamma_t(f(x_t) - \mathrm{Opt}) > \gamma^2$, the concluding inequality being given by the definition of γ_t and our assumption that $f_0(x_t) - \mathrm{Opt} > \gamma\|f'_0(x_t)\|_*$ at all productive steps t. The bottom line is that the left hand side in (1.18) is $> N\gamma^2 = 2\Omega$, what contradicts (1.18). □

1.4 Minimizing Strongly Convex Functions

The MD algorithm can be modified to attain the rate $O(1/t)$ in the case where the objective f in (1.2) is *strongly convex*. The strong convexity of f *with modulus* $\kappa > 0$ means that

$$\forall (x, x' \in \mathcal{X}) \ \ \langle f'(x) - f'(x'),\, x - x'\rangle \ge \kappa\|x - x'\|^2. \quad (1.19)$$

Further, let ω be the d.-g.f. *for the entire* E (not just for $\mathcal{X}$, which may be unbounded in this case), compatible with $\|\cdot\|$. W.l.o.g. let $0 = \mathrm{argmin}_E\, \omega$, and let

$$\Omega = \max_{\|u\|\le 1} \omega(u) - \omega(0)$$

be the variation of ω on the unit ball of $\|\cdot\|$. Now, let $\omega^{R,z}(u) = \omega\big(\frac{u-z}{R}\big)$ and $V_x^{R,z}(u) = \omega^{R,z}(u) - \omega^{R,z}(x) - \langle(\omega^{R,z}(x))',\, u - x\rangle$. Given $z \in \mathcal{X}$ and $R > 0$ we define the prox-mapping

$$\mathrm{Prox}_x^{R,z}(\xi) = \operatorname*{argmin}_{u\in\mathcal{X}}[\langle\xi, u\rangle + V_x^{R,z}(u)].$$

and the recurrence (cf. (1.6))

$$\begin{array}{l} x_{t+1} = \mathrm{Prox}_{x_t}^{R,z}(\gamma_t f'(x_t)), \ \ t = 1, 2, ... \\ x^t(R, z) = \big[\sum_{\tau=1}^t \gamma_\tau\big]^{-1} \sum_{\tau=1}^t \gamma_\tau x_\tau \end{array} \quad (1.20)$$

We start with the following analogue of Proposition 1.1:

Proposition 1.3. *Let f be strongly convex on $\mathcal{X}$ with modulus $\kappa > 0$ and Lipschitz continuous on $\mathcal{X}$ with $L := \sup_{x\in\mathcal{X}} \|f'(x)\|_* < \infty$. Given $R > 0$, $t \ge 1$, suppose that $\|x_1 - x_*\| \le R$, where x_* is the minimizer of f on $\mathcal{X}$,*

and let the stepsizes γ_τ satisfy

$$\gamma_\tau = \frac{\sqrt{2\Omega}}{RL\sqrt{t}},\ 1 \le \tau \le t. \tag{1.21}$$

Then after t iterations (1.20) *one has*

$$f(x^t(R,x_1)) - \mathrm{Opt} \ \le\ \frac{1}{t}\sum_{\tau=1}^{t}\langle f'(x_\tau),\, x_\tau - x_*\rangle \le \frac{LR\sqrt{2\Omega}}{\sqrt{t}}, \tag{1.22}$$

$$\|x^t(R,x_1) - x_*\|^2 \ \le\ \frac{1}{t\kappa}\sum_{\tau=1}^{t}\langle f'(x_\tau),\, x_\tau - x_*\rangle \le \frac{LR\sqrt{2\Omega}}{\kappa\sqrt{t}}. \tag{1.23}$$

Proof. Observe that the modulus of strong convexity of the function $\omega^{R,x_1}(\cdot)$ w.r.t. the norm $\|\cdot\|_R = \|\cdot\|/R$ is 1, and the conjugate of the latter norm is $R\|\cdot\|_*$. Following the steps of the proof of Proposition 1.1, with $\|\cdot\|_R$ and $\omega^{R,x_1}(\cdot)$ in the roles of $\|\cdot\|$, respectively, we come to the analogue of (1.7) as follows:

$$\forall u \in \mathcal{X}: \sum_{\tau=1}^{t}\gamma_\tau\langle f'(x_\tau),\, x_\tau - u\rangle \le V_{x_1}^{R,x_1}(u) + \tfrac{1}{2}\sum_{\tau=1}^{t} R^2L^2\gamma_\tau^2 \le \Omega + \tfrac{1}{2}\sum_{\tau=1}^{t} R^2L^2\gamma_\tau^2$$

Setting $u = x_*$ (so that $V^{R,x_1}(x_*) \le \Omega$ due to $\|x_1 - x_*\| \le R$), and substituting the value (1.21) of γ_τ, we come to (1.22). Further, from strong convexity of f it follows that $\langle f'(x_\tau), x_\tau - x_*\rangle \ge \kappa\|x_\tau - x_*\|^2$, which combines with the definition of $x^t(R,x_1)$ to imply the first inequality in (1.23) (recall that γ_τ is independent of τ, so that $x^t(R,x_1) = \frac{1}{t}\sum_{\tau=1}^{t} x_\tau$). The second inequality in (1.23) follows from (1.22). □

Proposition 1.21 states that the smaller is R (i.e., the closer is the initial guess x_1 to x_*), the better will be the accuracy of the approximate solution $x^t(R,x_1)$ both in terms of f and in terms of the distance to x_*. When the upper bound on this distance, as given by (1.22), becomes small, we can restart the MD using $x^t(\cdot)$ as the improved initial point, compute new approximate solution, and so on. The algorithm below is a simple implementation this idea.

Suppose that $x_1 \in \mathcal{X}$ and $R_0 \ge \|x_* - x_1\|$ are given. The algorithm is as follows:

1. *Initialization:* Set $y_0 = x_1$.

2. *Stage $k = 1, 2, ...$:* Set $N_k = \mathrm{Ceil}(2^{k+2}\frac{L^2\Omega}{\kappa^2 R_0^2})$, where $\mathrm{Ceil}(t)$ is the smallest integer $\ge t$, and compute $y_k = x^{N_k}(R_{k-1}, y_{k-1})$ according to (1.20), with $\gamma_t = \gamma^k := \frac{\sqrt{2\Omega}}{LR_{k-1}}$, $1 \le t \le N_k$. Set $R_k^2 = 2^{-k}R_0^2$ and pass to stage $k+1$.

For the search points $x_1, ..., x_{N_k}$ of the k-th stage of the method we define

$$\delta_k = \frac{1}{N_k} \sum_{\tau=1}^{N_k} \langle f'(x_\tau), x_\tau - x_* \rangle.$$

Let k_* be the smallest integer such that $k \geq 1$ and $2^{k+2} \frac{L^2\Omega}{\kappa^2 R_0^2} > k$, and let $M_k = \sum_{j=1}^k N_j$, $k = 1, 2, ...$ Note that M_k is the total number of prox-steps carried out at the first k stages.

Proposition 1.4. *Setting $y_0 = x_1$, the points y_k, $k = 0, 1, ...$, generated by the above algorithm satisfy the following relations:*

$$\|y_k - x_*\|^2 \leq R_k^2 = 2^{-k} R_0^2, \qquad (I_k)$$

$k = 0, 1, ...,$

$$f(y_k) - \mathrm{Opt} \leq \delta_k \leq \kappa R_k^2 \ = \ \kappa 2^{-k} R_0^2, \qquad (J_k)$$

$k = 1, 2, ...$ *As a result,*

(i) *when $1 \leq k < k_*$, one has $M_k \leq 5k$ and*

$$f(y_k) - \mathrm{Opt} \leq \kappa 2^{-k} R_0^2; \qquad (1.24)$$

(ii) *when $k \geq k_*$, one has*

$$f(y_k) - \mathrm{Opt} \leq \frac{16 L^2 \Omega}{\kappa M_k}. \qquad (1.25)$$

The proposition says that when the approximate solution y_k is "far" from x_*, the method converges linearly; when approaching x_*, it slows down and switches to the rate $O(1/t)$.

Proof. Let us prove (I_k), (J_k) by induction in k. (I_0) is valid due to $y_0 = x_1$ and the origin of R_0. Assume that for some $m \geq 1$ relations (I_k) and (J_k) are valid for $1 \leq k \leq m-1$, and let us prove that then (I_m), (J_m) are valid as well. Applying Proposition 1.3 with $R = R_{m-1}$, $x_1 = y_{m-1}$ (so that $\|x_* - x_1\| \leq R$ by (I_{m-1})) and $t = N_m$, we get

$$(a): \ f(y_m) - \mathrm{Opt} \leq \delta_m \leq \frac{L R_{m-1} \sqrt{2\Omega}}{\sqrt{N_m}}, \ (b): \ \|y_m - x_*\|^2 \leq L R_{m-1} \frac{\sqrt{2\Omega}}{\kappa \sqrt{N_m}}.$$

Since $R_{m-1}^2 = 2^{1-m} R_0^2$ by (I_{m-1}) and $N_m \geq 2^{m+2} \frac{L^2\Omega}{\kappa^2 R_0^2}$, (b) implies (I_m), and (a) implies (J_m). Induction is completed.

Now let us prove that $M_k \leq 5k$ for $1 \leq k < k_*$. Indeed, for such a k and for $1 \leq j \leq k$ we have $N_j = 1$ when $2^{j+2} \frac{L^2\Omega}{\kappa^2 R_0^2} < 1$, let it be so for $j < j_*$, and $N_j \leq 2^{j+3} \frac{L^2\Omega}{\kappa^2 R_0^2}$ for $j_* \leq j \leq k$. It follows that when $j_* > k$, we have $M_k = k$.

When $j_* \le k$, we have $M := \sum_{j=j_*}^{k} N_j \le 2^{k+4}\frac{L^2\Omega}{\kappa^2 R_0^2} \le 4k$ (the concluding inequality is due to $k < k_*$), whence $M_k = j_* - 1 + M \le 5k$, as claimed. Invoking (J_k), we arrive at (i).

To prove (ii), let $k \ge k_*$, whence $N_k \ge k+1$. We have

$$2^{k+3}\frac{L^2\Omega}{\kappa^2 R_0^2} > \sum_{j=1}^{k} 2^{j+2}\frac{L^2\Omega}{\kappa^2 R_0^2} \ge \sum_{j=1}^{k}(N_j - 1) = M_k - k \ge M_k/2,$$

where the concluding $\ge$ stems from the fact that $N_k \ge k+1$, and therefore $M_k \ge \sum_{j=1}^{k-1} N_j + N_k \ge (k-1) + (k+1) = 2k$. Thus $M_k \le 2^{k+4}\frac{L^2\Omega}{\kappa^2 R_0^2}$, that is $2^{-k} \le \frac{16L^2\Omega}{M_k\kappa^2 R_0^2}$, and the right hand side of (J_k) is $\le \frac{16L^2\Omega}{M_k\kappa}$. □

1.5 Mirror Descent Stochastic Approximation

The MD algorithm can be extended to the case when the objective f in (1.2) is given by *Stochastic oracle* – a routine which at t-th call, the query point being $x_t \in \mathcal{X}$, returns a vector $G(x_t, \xi_t)$, where $\xi_1, \xi_2, ...$ are independent identically distributed "oracle noises." We assume that for all $x \in \mathcal{X}$ it holds

$$\mathbf{E}\left\{\|G(x,\xi)\|_*^2\right\} \le L^2 < \infty \ \&\ \|g(x) - f'(x)\|_* \le \mu,\ g(x) = \mathbf{E}\{G(x,\xi)\}. \quad (1.26)$$

Replacing in (1.6) the subgradients $f'(x_t)$ with their stochastic estimates $G(x_t, \xi_t)$, we arrive at *Robust Mirror Descent Stochastic Approximation* (RMDSA). The convergence properties of this procedure are presented in the following counterpart of Proposition 1.1:

Proposition 1.5. *Let $\mathcal{X}$ be bounded. Given an integer $N \ge 1$, consider N-step RMDSA with the stepsizes*

$$\gamma_t = \sqrt{2\Omega}/[L\sqrt{N}],\ 1 \le t \le N. \quad (1.27)$$

Then

$$\mathbf{E}\left\{f(x^N) - \mathrm{Opt}\right\} \le \sqrt{2\Omega}L/\sqrt{N} + 2\sqrt{2\Omega}\mu. \quad (1.28)$$

Proof. Let $\xi^t = [\xi_1; ...; \xi_t]$, so that x_t is deterministic function of ξ^{t-1}. Exactly the same reasoning as in the proof of Proposition 1.1 results in the following analogy of (1.7):

$$\sum_{\tau=1}^{N}\gamma_\tau\langle G(x_\tau,\xi_\tau), x_\tau - x_*\rangle \le \Omega + \tfrac{1}{2}\sum_{\tau=1}^{N}\gamma_\tau^2\|G(x_\tau,\xi_\tau)\|_*^2. \quad (1.29)$$

Observe that x_τ is a deterministic function of ξ^{t-1}, so that

$$\mathbf{E}_{\xi_\tau}\{\langle G_\tau(x_\tau,\xi_\tau), x_\tau - x_*\rangle\} = \langle g(x_\tau), x_\tau - x_*\rangle \geq \langle f'(x_\tau), x_\tau - x_*\rangle - \mu D,$$

where $D = \max_{x,x'\in\mathcal{X}} \|x - x'\|$ is the $\|\cdot\|$-diameter of $\mathcal{X}$. Now, when taking expectations of both sides of (1.29) we get

$$\mathbf{E}\left\{\sum\nolimits_{\tau=1}^N \gamma_\tau \langle f'(x_\tau), x_\tau - x_*\rangle\right\} \leq \Omega + \frac{L^2}{2}\sum\nolimits_{\tau=1}^N \gamma_\tau^2 + \mu D \sum\nolimits_{\tau=1}^N \gamma_\tau.$$

In the same way as in the proof of Proposition 1.1 we conclude that the left hand side in this inequality is $\geq [\sum_{\tau=1}^N \gamma_\tau]\mathbf{E}\{f(x^N) - \mathrm{Opt}\}$, so that

$$\mathbf{E}\{f(x^N) - \mathrm{Opt}\} \leq \frac{\Omega + \frac{L^2}{2}\sum_{\tau=1}^N \gamma_\tau^2}{\sum_{\tau=1}^N \gamma_\tau} + \mu D. \tag{1.30}$$

Observe that when $x \in \mathcal{X}$, we have $\omega(x) - \omega(x_1) - \langle \omega'(x_1), x - x_1\rangle \geq \frac{1}{2}\|x - x_1\|^2$ by strong convexity of ω, and $\omega(x) - \omega(x_1) - \langle \omega'(x_1), x - x_1\rangle \leq \omega(x) - \omega(x_1) \leq \Omega$ (since $x_1 = \mathrm{argmin}_{\mathcal{X}}\, \omega$ and thus $\langle \omega'(x_1), x - x_1\rangle \geq 0$). Thus, $\|x - x_1\| \leq \sqrt{2\Omega}$ for every $x \in \mathcal{X}$, whence $D := \max_{x,x'\in\mathcal{X}} \|x - x'\| \leq 2\sqrt{2\Omega}$. This relation combines with (1.30) and (1.27) to imply (1.28). □

1.6 Mirror Descent for Convex-Concave Saddle Point Problems

We are about to demonstrate that the MD scheme can be naturally extended from problems of convex minimization to the *convex-concave saddle point problems*.

1.6.1 Preliminaries

Convex-concave saddle point problem. A convex-concave saddle point (c.-c.s.p.) problem reads

$$\mathrm{SadVal} = \inf_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} \phi(x,y), \tag{1.31}$$

where $\mathcal{X} \subset E_x$, $\mathcal{Y} \subset E_y$ are nonempty closed convex sets in the respective Euclidean spaces E_x, E_y. The *cost function* $\phi(x,y)$ is a continuous function on $\mathcal{Z} = \mathcal{X}\times\mathcal{Y} \in E = E_x \times E_y$ which is convex in the variable $x \in \mathcal{X}$ and concave in the variable $y \in \mathcal{Y}$; the quantity SadVal is called the *saddle point value* of ϕ on $\mathcal{Z}$. By definition (precise) solutions to (1.31) are *saddle points* of ϕ on $\mathcal{Z}$, that is, points $(x_*, y_*) \in \mathcal{Z}$ such that $\phi(x, y_*) \geq \phi(x_*, y_*) \geq \phi(x_*, y)$ for all $(x,y) \in \mathcal{Z}$. The data of problem (1.31) give rise to a *primal-dual pair*

of convex optimization problems

$$\begin{array}{rcll} \mathrm{Opt}(P) & = & \min\limits_{x\in\mathcal{X}} \overline{\phi}(x), \quad \overline{\phi}(x) = \sup_{y\in\mathcal{Y}} \phi(x,y) & (P) \\ \mathrm{Opt}(D) & = & \max\limits_{y\in\mathcal{Y}} \underline{\phi}(y), \quad \underline{\phi}(y) = \inf\limits_{x\in\mathcal{X}} \phi(x,y) & (D) \end{array}$$

ϕ possesses saddle points on $\mathcal{Z}$ if and only if problems (P), (D) are solvable with equal optimal values. Whenever saddle points exist, they are exactly the pairs (x_*, y_*) comprised of optimal solutions x_*, y_* to the respective problems (P), (D), and for every such pair (x_*, y_*) we have

$$\begin{array}{l} \phi(x_*,y_*) = \overline{\phi}(x_*) = \mathrm{Opt}(P) = \mathrm{SadVal} := \inf\limits_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} \phi(x,y) \\ = \sup_{y\in\mathcal{Y}} \inf\limits_{x\in\mathcal{X}} \phi(x,y) = \mathrm{Opt}(D) = \underline{\phi}(y_*). \end{array}$$

From now on we assume that (1.31) is solvable.

Remark 1.6. *With our basic assumptions on ϕ (continuity and convexity-concavity on $\mathcal{X}\times\mathcal{Y}$) and on $\mathcal{X}, \mathcal{Y}$ (nonemptiness, convexity and closedness) (1.31) definitely is solvable if either $\mathcal{X}$ and $\mathcal{Y}$ are bounded, or both $\mathcal{X}$ and all level sets $\{y \in \mathcal{Y} : \underline{\phi}(y) \geq a\}$, $a \in \mathbb{R}$, of $\underline{\phi}$ are bounded; these are the only situations we are about to consider in this chapter and in chapter 2.*

Saddle point accuracy measure. A natural way to quantify the accuracy of a candidate solution $z = (x, y) \in \mathcal{Z}$ to the c.-c.s.p. problem (1.31) is given by the *gap*

$$\begin{array}{rcl} \epsilon_{\mathrm{sad}}(z) & = & \sup_{\eta\in\mathcal{Y}} \phi(x,\eta) - \inf\limits_{\xi\in\mathcal{X}} \phi(\xi,y) = \overline{\phi}(x) - \underline{\phi}(y) \\ & = & \big[\overline{\phi}(x) - \mathrm{Opt}(P)\big] + \big[\mathrm{Opt}(D) - \underline{\phi}(y)\big] \end{array} \tag{1.32}$$

where the concluding equality is given by the fact that, by our standing assumption, ϕ has a saddle point and thus $\mathrm{Opt}(P) = \mathrm{Opt}(D)$. We see that $\epsilon_{\mathrm{sad}}(x, y)$ is just the sum of non-optimalities, in terms of the respective objectives, of x as an approximate solution to (P) and y as an approximate solution to (D).

Monotone operator associated with (1.31). Let $\partial_x\phi(x, y)$ be the set of all subgradients w.r.t. $\mathcal{X}$ of (the convex function) $\phi(\cdot, y)$, taken at a point $x \in \mathcal{X}$, and $\partial_y[-\phi(x, y)]$ be the set of all subgradients w.r.t. $\mathcal{Y}$ (of the convex function) $-\phi(x, \cdot)$, taken at a point $y \in \mathcal{Y}$. We can associate with ϕ the point-to-set operator

$$\Phi(x,y) = \{\Phi_x(x,y) = \partial_x\phi(x,y)\} \times \{\Phi_y(x,y) = \partial_y[-\phi(x,y)]\}$$

The domain Dom $\Phi := \{(x,y) : \Phi(x,y) \neq \emptyset\}$ of this operator is comprised of all pairs $(x,y) \in \mathcal{Z}$ for which the corresponding subdifferentials are nonempty; it definitely contains the relative interior rint $\mathcal{Z} =$ rint $\mathcal{X} \times$ rint $\mathcal{Y}$ of $\mathcal{Z}$, and the values of Φ in its domain are direct products of nonempty closed convex sets in E_x and E_y. It is well known (and easily seen) that Φ is monotone:

$$\forall (z, z' \in \text{Dom } \Phi, F \in \Phi(z), F' \in \Phi(z')) : \langle F - F', z - z' \rangle \geq 0.$$

and the saddle points of ϕ are exactly the points z_* such that $0 \in \Phi(z_*)$. An equivalent, more convenient in our context, characterization of saddle points is as follows: z_* is a saddle point of ϕ if and only if for some (and then — for every) *selection* $F(\cdot)$ of Φ (i.e., a vector field $F(z) :$ rint $\mathcal{Z} \to E$ such that $F(z) \in \Phi(z)$ for every $z \in$ rint $\mathcal{Z}$) one has

$$\langle F(z), z - z_* \rangle \geq 0 \, \forall z \in \text{rint } \mathcal{Z}. \tag{1.33}$$

1.6.2 Saddle Point Mirror Descent

Here we assume that $\mathcal{Z}$ is bounded and ϕ is Lipschitz continuous on $\mathcal{Z}$ (whence, in particular, the domain of the associated monotone operator Φ is the entire $\mathcal{Z}$).

The setup the MP algorithm involves a norm $\|\cdot\|$ on the embedding space $E = E_x \times E_y$ of $\mathcal{Z}$ and a d.-g.f. $\omega(\cdot)$ for $\mathcal{Z}$ compatible with this norm. For $z \in \mathcal{Z}^o$, $u \in \mathcal{Z}$ let (cf. the definition (1.4))

$$V_z(u) = \omega(u) - \omega(z) - \langle \omega'(z), u - z \rangle,$$

and let $z_c = \text{argmin}_{u \in \mathcal{Z}} \omega(u)$. We assume that given $z \in \mathcal{Z}^o$ and $\xi \in E$, it is easy to compute the prox-mapping

$$\text{Prox}_z(\xi) = \operatorname*{argmin}_{u \in \mathcal{Z}} \left[\langle \xi, u \rangle + V_z(u) \right] \left(= \operatorname*{argmin}_{u \in \mathcal{Z}} \left[\langle \xi - \omega'(z), u \rangle + \omega(u) \right] \right).$$

We denote by $\Omega = \max_{u \in \mathcal{Z}} V_{z_c}(u) \leq \max_{\mathcal{Z}} \omega(\cdot) - \min_{\mathcal{Z}} \omega(\cdot)$ the "$\omega(\cdot)$-diameter" of $\mathcal{Z}$ (cf. section 1.2.2).

Let a First Order oracle for ϕ be available, so that for every $z = (x,y) \in \mathcal{Z}$ we can compute a vector $F(z) \in \Phi(z = (x,y)) := \{\partial_x \phi(x,y)\} \times \{\partial_y [-\phi(x,y)]\}$. The saddle point MD algorithm is given by the recurrence

$$\begin{array}{ll} (a): & z_1 = z_c, \\ (b): & z_{\tau+1} = \text{Prox}_{z_\tau}(\gamma_\tau F(z_\tau)), \\ (c): & z^\tau = \left[\sum_{s=1}^\tau \gamma_s\right]^{-1} \sum_{s=1}^\tau \gamma_s w_s, \end{array} \tag{1.34}$$

where $\gamma_\tau > 0$ are the stepsizes. Note that $z_\tau, \omega_\tau \in \mathcal{Z}^o$, whence $z^t \in \mathcal{Z}$.

The convergence properties of the algorithm are given by the following

Proposition 1.7. *Suppose that $F(\cdot)$ is bounded on $\mathcal{Z}$ and L is such that $\|F(z)\|_* \le L$ for all $z \in \mathcal{Z}$.*

(i) *For every $t \ge 1$ it holds*

$$\epsilon_{\rm sad}(z^t) \le \left[\sum\nolimits_{\tau=1}^t \gamma_\tau\right]^{-1} \left[\Omega + \frac{L^2}{2}\sum\nolimits_{\tau=1}^t \gamma_\tau^2\right]. \tag{1.35}$$

(ii) *As a consequence, the N-step MD algorithm with constant stepsizes $\gamma_\tau = \gamma/L\sqrt{N}$, $\tau = 1, ..., N$ satisfies*

$$\epsilon_{\rm sad}(z^N) \le \frac{L}{\sqrt{N}}\left[\frac{\Omega}{\gamma} + \frac{L\gamma}{2}\right].$$

In particular, the N-step MD algorithm with constant stepsizes $\gamma_\tau = L^{-1}\sqrt{\frac{2\Omega}{N}}$, $\tau = 1, ..., N$ satisfies

$$\epsilon_{\rm sad}(z^N) \le L\sqrt{\frac{2\Omega}{N}}.$$

Proof. By definition of $z_{\tau+1} = {\rm Prox}_{z_\tau}(\gamma_\tau F(z_\tau))$ we get

$$\forall u \in \mathcal{Z},\ \gamma_\tau\langle F(z_\tau), z_\tau - u\rangle \le V_{z_\tau}(u) - V_{z_{\tau+1}}(u) + \gamma_\tau^2\|F(z_\tau)\|_*^2/2$$

(it suffices to repeat the derivation of (1.13) in the proof of Proposition 1.1 with $f'(x_\tau)$, x_τ and $x_{\tau+1}$ substituted, respectively, with $F(z_\tau)$, z_τ and $z_{\tau+1}$). When summing up for $i = 1, ..., t$ we get for all $u \in \mathcal{Z}$:

$$\sum_{\tau=1}^t \gamma_\tau\langle F(z_\tau), z_\tau - u\rangle \le V_{z_1}(u) + \sum_{\tau=1}^t \gamma_\tau^2\|F(z_\tau)\|_*^2/2 \le \Omega + \frac{L^2}{2}\sum_{\tau=1}^t \gamma_\tau^2. \tag{1.36}$$

Let $z_\tau = (x_\tau, y_\tau)$, $z^t = (x^t, y^t)$ and $\lambda_\tau = \left[\sum_{s=1}^t \gamma_s\right]^{-1}\gamma_\tau$. Note that $\sum_{s=1}^t \lambda_s = 1$, and for

$$\sum_{\tau=1}^t \lambda_\tau\langle F(z_\tau), z_\tau - u\rangle = \sum_{\tau=1}^t \lambda_\tau\left[\langle\nabla_x\phi(x_\tau, y_\tau), x_\tau - x\rangle + \langle\nabla_y\phi(x_\tau, y_\tau), y - y_\tau\rangle\right]$$

we have

$$\begin{array}{ll} \sum_{\tau=1}^t \lambda_\tau\left[\langle\nabla_x\phi(x_\tau, y_\tau), x_\tau - x\rangle + \langle\nabla_y\phi(x_\tau, y_\tau), y - y_\tau\rangle\right] & \\ \ge \sum_{\tau=1}^t \lambda_\tau\left[[\phi(x_\tau, y_\tau) - \phi(x, y_\tau)] + [\phi(x_\tau, y) - \phi(x_\tau, y_\tau)]\right] & (a) \\ = \sum_{\tau=1}^t \lambda_\tau[\phi(x_\tau, y) - \phi(x, y_\tau)] & \\ \ge \phi(\sum_{\tau=1}^t \lambda_\tau x_\tau, y) - \phi(x, \sum_{\tau=1}^t \lambda_\tau y_\tau) = \phi(x^t, y) - \phi(x, y^t) & (b) \end{array} \tag{1.37}$$

(inequalities in $(a), (b)$ are due to the convexity-concavity of ϕ). We come to so that (1.36) results in

$$\phi(x^t, y) - \phi(x, y^t) \leq \frac{\Omega + \frac{L^2}{2}\sum_{\tau=1}^t \gamma_\tau^2}{\sum_{\tau=1}^t \gamma_\tau} \quad \forall (x, y) \in \mathcal{Z}.$$

Taking supremum in $(x, y) \in \mathcal{Z}$, we arrive at (1.35). □

1.7 Setting up a Mirror Descent method

An advantage of the Mirror Descent scheme is that its "degrees of freedom" (the norm $\|\cdot\|$ and the d.-g.f. $\omega(\cdot)$) allow to adjust, to some extent, the method to the geometry of the problem under consideration. This is the issue we are focusing on in this section. For the sake of definiteness, we restrict ourselves with the minimization problem (1.2); the saddle point case (1.31) is completely similar, with $\mathcal{Z}$ in the role of $\mathcal{X}$.

1.7.1 Building blocks

The "basic" MD setups are as follows:

1. *"Euclidean setup:"* $\|\cdot\| = \|\cdot\|_2$, $\omega(x) = \frac{1}{2}x^T x$.

2. *"ℓ_1 setup:"* For this setup, $E = \mathbb{R}^n$, $n > 1$, and $\|\cdot\| = \|\cdot\|_1$. As about $\omega(\cdot)$, there could be several choices, depending on what $\mathcal{X}$ is:

   (a) when $\mathcal{X}$ is unbounded, seemingly the only good choice is $\omega(x) = C\ln(n)\|x\|^2_{p(n)}$ with $p(n) = 1 + \frac{1}{2\ln(n)}$, where an *absolute constant* C is chosen in a way which ensures (1.3) (one can take $C = \mathrm{e}$);

   (b) when $\mathcal{X}$ is bounded, assuming w.l.o.g. that $\mathcal{X} \subset B^{n,1} := \{x \in \mathbb{R}^n : \|x\|_1 \leq 1\}$, one can set $\omega(x) = C\ln(n)\sum_{i=1}^n |x_i|^{p(n)}$ with the same as above $p(n)$ and $C = 2\mathrm{e}$;

   (c) when $\mathcal{X}$ is a part of the simplex $\mathrm{S}_n^+ = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i \leq 1\}$ (or the "flat" simplex $\mathrm{S}_n = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$) intersecting int $\mathbb{R}^n_+$, a good choice of $\omega(x)$ is the entropy:

   $$\omega(x) = \mathrm{Ent}(x) := \sum_{i=1}^n x_i \ln(x_i) \tag{1.38}$$

3. *"Matrix setup:"* This is the "matrix analogy" of the ℓ_1 setup. Here the embedding space E of $\mathcal{X}$ is the space $\mathbf{S}^\nu$ of block-diagonal symmetric matrices with fixed block-diagonal structure $\nu = [\nu_1; ...; \nu_k]$ (k diagonal blocks of row sizes $\nu_1, ..., \nu_k$). $\mathbf{S}^\nu$ is equipped with the Frobenius inner product $\langle X, Y\rangle = \mathrm{Tr}(XY)$ and the trace norm $|X|_1 = \|\lambda(X)\|_1$, where

$\lambda(X)$ is the vector of eigenvalues (taken with their multiplicities in the non-ascending order) of a symmetric matrix X. The d.-g.f.'s are the "matrix analogies" of those for ℓ_1 setup. Specifically,

(a) when $\mathcal{X}$ is unbounded, we set $\omega(X) = C\ln(|\nu|)\|\lambda(X)\|^2_{p(|\nu|)}$, where $|\nu| = \sum_{\ell=1}^k \nu_\ell$ is the total row size of matrices from $\mathbf{S}^\nu$, and C is an appropriate absolute constant which ensures (1.3) (one can take $C = 2\mathrm{e}$);

(b) when $\mathcal{X}$ is bounded, assuming w.l.o.g. that $\mathcal{X} \subset B^{\nu,1} = \{X \in \mathbf{S}^\nu : |X|_1 \le 1\}$, we can take $\omega(X) = 4\mathrm{e}\ln(|\nu|)\sum_{i=1}^{|\nu|}|\lambda_i(X)|^{p(|\nu|)}$;

(c) when $\mathcal{X}$ is a part of the *spectahedron* $\Sigma_\nu^+ = \{X \in \mathbf{S}^\nu : X \succeq 0, \mathrm{Tr}(X) \le 1\}$ (or the "flat" spectahedron $\Sigma_\nu = \{X \in \mathbf{S}^\nu : X \succeq 0, \mathrm{Tr}(X) = 1\}$) intersecting the interior $\{X \succ 0\}$ of the positive semidefinite cone $\mathbf{S}^\nu_+ = \{X \in \mathbf{S}^\nu : X \succeq 0\}$, one can take as $\omega(X)$ the matrix entropy: $\omega(X) = 2\mathrm{Ent}(\lambda(X)) = 2\sum_{i=1}^{|\nu|}\lambda_i(X)\ln(\lambda_i(X))$.

Note that ℓ_1-setup can be viewed as a particular case of the matrix one, corresponding to the case when the block-diagonal matrices in question are diagonal, and we identify a diagonal matrix with the vector of its diagonal entries.

With the outlined setups, the Simplicity assumption holds provided that $\mathcal{X}$ is simple enough, specifically:

- Within the Euclidean setup, $\mathrm{Prox}_x(\xi)$ is the metric projection of the vector $x - \xi$ onto $\mathcal{X}$, that is, the point of $\mathcal{X}$ which is the closest to $x - \xi$ in ℓ_2-norm. Examples of sets $\mathcal{X} \subset \mathbb{R}^n$ for which metric projection is easy include, among others, $\|\cdot\|_p$-balls and intersections of centered at the origin $\|\cdot\|_p$-balls with the nonnegative orthant $\mathbb{R}^n_+$;
- Within ℓ_1-setup, computing the prox-mapping is reasonably easy
— in the case of 2a — when $\mathcal{X}$ is the entire $\mathbb{R}^n$ or $\mathbb{R}^n_+$,
— in the case of 2b — when $\mathcal{X}$ is the entire $B^{n,1}$ or the intersection of $B^{n,1}$ with $\mathbb{R}^n_+$,
— in the case of 2c — when $\mathcal{X}$ is the entire S^+_n or S_n.
With the indicated sets $\mathcal{X}$, in the cases of 2a – 2b computing the prox-mapping requires solving auxiliary one- or two-dimensional convex problems (which can be done within machine accuracy by, e.g., the Ellipsoid algorithm in $O(n)$ operations, cf. Nemirovski and Yudin (1983), Chapter II). In the

case of 2c, the prox-mappings are given by the explicit formulas:

$$\begin{array}{l}\mathcal{X}=\mathrm{S}_n^+\Rightarrow \mathrm{Prox}_x(\xi)=\begin{cases}[x_1\mathrm{e}^{\xi_1-1};...;x_n\mathrm{e}^{\xi_n-1}], & \sum_i \mathrm{e}^{\eta_i-1}\le 1\\ \left[\sum_i x_i\mathrm{e}^{\xi_i}\right]^{-1}\left[x_1\mathrm{e}^{\eta_1};...;\,x_n\mathrm{e}^{\eta_n}\right], & \text{otherwise}\end{cases}\\ \mathcal{X}=\mathrm{S}_n\Rightarrow \mathrm{Prox}_x(\xi)=\left[\sum_i x_i\mathrm{e}^{\xi_i}\right]^{-1}\left[x_1\mathrm{e}^{\eta_1};...;\,x_n\mathrm{e}^{\eta_n}\right].\end{array} \tag{1.39}$$

▪ Within the Matrix setup, computing the prox-mapping is relatively easy
— in the case of 3a — when $\mathcal{X}$ is the entire $\mathbf{S}^\nu$ or the positive semidefinite cone $\mathbf{S}^\nu_+=\{X\in\mathbf{S}^\nu: X\succeq 0\}$,
— in the case of 3b — when $\mathcal{X}$ is the entire $B^{\nu,1}$ or the intersection of $B^{\nu,1}$ with $\mathbf{S}^\nu_+$,
— in the case of 3c — when $\mathcal{X}$ is the entire spectahedron Σ^+_ν or Σ_ν.
Indeed, in the outlined cases computing $W=\mathrm{Prox}_X(\Xi)$ reduces to computing the eigenvalue decomposition of the matrix X (which allows to get $\omega'(X)$) and subsequent eigenvalue decomposition of the matrix $H=\Xi-\omega'(X)$: $H=U\,\mathrm{Diag}\{h\}U^T$ (here $\mathrm{Diag}(A)$ stands for the diagonal matrix with the same diagonal as A). It is easily seen that in the cases in question $W=U\,\mathrm{Diag}\{w\}U^T$, $w=\operatorname*{argmin}_{z:\,\mathrm{Diag}\{z\}\in\mathcal{X}}\{\langle\mathrm{Diag}\{h\},\mathrm{Diag}\{z\}\rangle+\omega(\mathrm{Diag}\{z\})\}$, and the latter problem is exactly the one arising in the ℓ_1 setup.

1.7.1.1 Illustration: Euclidean setup vs. ℓ_1 setup

To illustrate the ability of the MD scheme to adjust, to some extent, the method to problem's geometry, consider problem (1.2) when $\mathcal{X}$ is the unit $\|\cdot\|_p$-ball in $\mathbb{R}^n$, where $p=1$ or $p=2$, and let us compare the respective "performances" of the Euclidean and the ℓ_1 setups (to make optimization over the unit Euclidean ball $B^{n,2}$ available for ℓ_1 setup, we pass from $\min_{\|x\|_2\le 1}f(x)$ to the equivalent problem $\min_{\|u\|_2\le n^{-1/2}}f(n^{1/2}u)$ and use the setup from item 2b, section 1.7.1). The ratio of the corresponding efficiency estimates (the right hand sides in (1.11)) within an absolute constant factor is

$$\Theta:=\frac{\mathrm{EffEst(Eucl)}}{\mathrm{EffEst}(\ell_1)}=\underbrace{\frac{1}{n^{1-1/p}\sqrt{\ln(n)}}}_{A}\cdot\underbrace{\frac{\sup_{x\in\mathcal{X}}\|f'(x)\|_2}{\sup_{x\in\mathcal{X}}\|f'(x)\|_{1\infty}}}_{B};$$

note that $\Theta\ll 1$ means that the MD with Euclidean setup significantly outperforms the MD with ℓ_1 setup, while $\Theta\gg 1$ witnesses exactly opposite. Now, A is ≤ 1 and thus is always "in favor" of the Euclidean setup, and is as small as $1/\sqrt{n\ln(n)}$ when $\mathcal{X}$ is the Euclidean ball ($p=2$), while

the factor B is in favor of the ℓ_1 setup — it is ≥ 1 and $\leq \sqrt{n}$ and well can be of order of $\sqrt{n}$ (look what happens when all entries in $f'(x)$ are of the same order of magnitude). Which one of the factors "overweights," it depends on f; however, reasonable choice can be made independently of the "fine structure" of f. Specifically, when $\mathcal{X}$ is the Euclidean ball, the factor $A = 1/\sqrt{n \ln n}$ is so small that the product AB definitely is ≤ 1, that is, the situation definitely is in favor of the Euclidean setup. In contrast to this, when $\mathcal{X}$ is the ℓ_1 ball ($p = 1$), A is "nearly constant" — just $O(1/\sqrt{\ln(n)})$; since B can be as large as $\sqrt{n}$, the situation is definitely in favor of the ℓ_1 setup — it can be outperformed by the Euclidean setup only marginally (by factor $\leq \sqrt{\ln n}$), with reasonable chances to outperform its adversary quite significantly — by factor $O(\sqrt{n/\ln(n)})$. Thus, there are all reasons to select the Euclidean setup when $p = 2$ and the ℓ_1 setup when $p = 1$.[6]

1.7.2 "Favorable Geometry" case

Consider the case when the domain $\mathcal{X}$ of (1.2) is bounded and, moreover, is a *subset* of the direct product $\mathcal{X}^+$ of "standard blocks:"

$$\mathcal{X}^+ = \mathcal{X}_1 \times ... \times \mathcal{X}_K \in E_1 \times ... \times E_K, \tag{1.40}$$

where for every $\ell = 1, ..., K$ the pair $(\mathcal{X}_\ell, E_\ell \supset \mathcal{X}_\ell)$ is
— either a *ball block*, that is, $E_\ell = \mathbb{R}^{n_\ell}$ and $\mathcal{X}_\ell$ is either the unit Euclidean ball $B^{n_\ell,2} = \{x \in \mathbb{R}^{n_\ell} : \|x\|_2 \leq 1\}$ in E_ℓ, or the intersection of this ball with $\mathbb{R}^{n_\ell}_+$;
— or a *spectahedron block*, that is, $E_\ell = \mathbf{S}^{\nu^\ell}$ is the space of block-diagonal, with block-diagonal structure ν^ℓ, symmetric matrices, and $\mathcal{X}_\ell$ is either the unit trace-norm ball $\{X \in \mathbf{S}^{\nu^\ell} : |X|_1 \leq 1\}$, or the intersection of this ball with $\mathbf{S}^{\nu^\ell}_+$, or the spectahedron $\Sigma^+_{\nu^\ell} = \{X \in \mathbf{S}^{\nu^\ell}_+ : \mathrm{Tr}(X) \leq 1\}$, or the "flat" spectahedron $\Sigma_{\nu^\ell} = \{X \in \mathbf{S}^{\nu^\ell}_+ : \mathrm{Tr}(X) = 1\}$.

Note that according to our convention to identify vectors with diagonals of diagonal matrices, we allow for some of $\mathcal{X}_\ell$ to be the unit ℓ_1 balls, or their nonnegative parts, or simplices — they are nothing but spectahedron blocks with purely diagonal structure ν^ℓ.

We equip the embedding spaces E_ℓ of blocks with the natural inner

6. In fact, with this recommendation we get *theoretically unimprovable*, in terms of the Information-Based Complexity Theory, methods for large-scale nonsmooth convex optimization on Euclidean and ℓ_1 balls (for details, see Nemirovski and Yudin, 1983; Ben-Tal et al., 2001); numerical experiments reported in (Ben-Tal et al., 2001; Nemirovski et al., 2009) seem to fully support the advantages of ℓ_1 setup when minimizing over large-scale simplices.

products (the standard inner products when $E_\ell = \mathbb{R}^{n_\ell}$ and the Frobenius inner product when $E_\ell = \mathbf{S}^{\nu^\ell}$) and norms $\|\cdot\|_{(\ell)}$ (the standard Euclidean norm when $E_\ell = \mathbb{R}^{n_\ell}$ and the trace-norm when $E_\ell = \mathbf{S}^{\nu^\ell}$), and the standard blocks $\mathcal{X}_\ell$ — with d.-g.f.'s

$$\omega_\ell(x^\ell) = \begin{cases} \frac{1}{2}[x^\ell]^T x^\ell, & \mathcal{X}_\ell \text{ is a ball block} \\ 4\mathrm{e}\ln(|\nu^\ell|)\sum_i |\lambda_i(X^\ell)|^{p(|\nu^\ell|)}, & \mathcal{X}_\ell \text{ is the unit } |\cdot|_1 \text{ ball } B^{\nu^\ell,1} \text{ in } E_\ell = \mathbf{S}^{\nu^\ell}, \text{ or } B^{\nu^\ell,1} \cap \mathbf{S}^{\nu^\ell}_+ \\ 2\mathrm{Ent}(\lambda(X^\ell)), & \mathcal{X}_\ell \text{ is the spectahedron } (\Sigma^+_{\nu^\ell} \text{ or } \Sigma_{\nu^\ell}) \text{ in } E_\ell = \mathbf{S}^{\nu^\ell} \end{cases} \tag{1.41}$$

cf. section 1.7.1. Finally, the embedding space $E = E_1 \times ... \times E_K$ of $\mathcal{X}^+$ (and thus — of $\mathcal{X} \subset \mathcal{X}^+$) is equipped with the direct product type Euclidean structure induced by the inner products on $E_1, ..., E_K$ and with the norm

$$\|(x^1, ..., x^K)\| = \sqrt{\sum\nolimits_{\ell=1}^K \alpha_\ell \|x^\ell\|^2_{(\ell)}} \tag{1.42}$$

where $\alpha_\ell > 0$ are construction's parameters. $\mathcal{X}^+$ is equipped with the d.-g.f.

$$\omega(x^1, ..., x^K) = \sum\nolimits_{\ell=1}^K \alpha_\ell \omega_\ell(x^\ell) \tag{1.43}$$

which, as it is easily seen, is compatible with the norm $\|\cdot\|$.
Assuming from now on that $\mathcal{X}$ intersects the relative interior rint $\mathcal{X}^+$, the restriction of $\omega(\cdot)$ onto $\mathcal{X}$ is a d.-g.f. for $\mathcal{X}$ compatible with the norm $\|\cdot\|$ on the space E embedding $\mathcal{X}$, and we can solve (1.2) by the MD algorithm associated with $\|\cdot\|$ and $\omega(\cdot)$. Let us optimize the efficiency estimate of this algorithm over the parameters α_ℓ of our construction. For the sake of definiteness, consider the case where f is represented by a deterministic First Order oracle (the "tuning" of the MD setup in the case of Stochastic oracle being completely similar). To this end assume that we have at our disposal upper bounds $L_\ell < \infty$, $1 \le \ell \le K$, on the quantities $\|f'_{x^\ell}(x^1, ..., x^K)\|_{(\ell),*}$, $x = (x^1, ..., x^K) \in \mathcal{X}$, where $f'_{x^\ell}(x)$ is the projection of $f'(x)$ onto E_ℓ, and $\|\cdot\|_{(\ell),*}$ is the norm on E_ℓ conjugate to $\|\cdot\|_{(\ell)}$ (that is, $\|\cdot\|_{(\ell),*}$ is the standard Euclidean norm $\|\cdot\|_2$ on E_ℓ when $E_\ell = \mathbb{R}^{n_\ell}$, and $\|\cdot\|_{(\ell),*}$ is the standard matrix norm (maximal singular value) when $E_\ell = \mathbf{S}^{\nu^\ell}$). The norm $\|\cdot\|_*$ conjugate to the norm $\|\cdot\|$ on E is

$$\begin{aligned} \|(\xi^1, ..., \xi^K)\|_* &= \sqrt{\textstyle\sum_{\ell=1}^K \alpha_\ell^{-1} \|\xi^\ell\|^2_{(\ell),*}} \\ \Rightarrow (\forall x \in \mathcal{X}) : \|f'(x)\|_* &\le L := \sqrt{\textstyle\sum_{\ell=1}^K \alpha_\ell^{-1} L_\ell^2}, \end{aligned} \tag{1.44}$$

and the quantity we need to minimize in order to get as efficient MD method as possible within our framework, is $\sqrt{\Omega}L$, see, e.g. (1.11). We clearly have $\Omega \le \Omega[\mathcal{X}^+] \le \sum_{\ell=1}^K \alpha_\ell \Omega_\ell[\mathcal{X}_\ell]$, where $\Omega_\ell[\mathcal{X}_\ell]$ is the variation (maximum minus minimum) of ω_ℓ on $\mathcal{X}_\ell$. These variations are upper-bounded by the quantities

$$\Omega_\ell = \begin{cases} \frac{1}{2} & \text{for ball blocks } \mathcal{X}_\ell \\ 4\mathrm{e}\ln(|\nu^\ell|) & \text{for spectahedron blocks } \mathcal{X}_\ell \end{cases}. \tag{1.45}$$

Assuming that we have K_b ball blocks $\mathcal{X}_1, ..., \mathcal{X}_{K_b}$ and K_s spectahedron blocks $\mathcal{X}_{K_b+1}, ..., \mathcal{X}_{K=K_b+K_s}$, we get

$$\Omega L \le \Omega[\mathcal{X}^+]L \le \left[\frac{1}{2}\sum_{\ell=1}^{K_b} \alpha_\ell + 4\mathrm{e}\sum_{\ell=K_b+1}^{K_b+K_s} \alpha_\ell \ln(|\nu^\ell|)\right] \sqrt{\sum_{\ell=1}^K \alpha_\ell^{-1} L_\ell^2}.$$

When optimizing the right hand side bound in $\alpha_1, ..., \alpha_L$, we get

$$\alpha_\ell = \frac{L_\ell}{\sqrt{\Omega_\ell}\sum_{i=1}^K L_i\sqrt{\Omega_i}},\ \Omega[\mathcal{X}^+] = 1,\ L = \mathcal{L} := \sum_{\ell=1}^K L_\ell \sqrt{\Omega_\ell}. \tag{1.46}$$

The associated with our "optimized setup" efficiency estimate (1.11) reads

$$\begin{array}{l} \overline{f}_N - \mathrm{Opt} \le O(1)\mathcal{L}N^{-1/2} \\ = O(1)[\max_{1\le\ell\le K} L_\ell]\left[K_b + \sum_{\ell=K_b+1}^{K_b+K_s}\sqrt{\ln(|\nu^\ell|)}\right] N^{-1/2}. \end{array} \tag{1.47}$$

We see that *if we consider* $\max_{1\le\ell\le K} L_\ell$, K_b *and* K_s *as given constants, the rate of convergence of the MD algorithm is* $O(1/\sqrt{N})$, N *being the number of steps, with the factor hidden in* $O(\cdot)$ *completely independent of the dimensions of the ball blocks and nearly independent of the sizes of the spectahedron blocks.* In other words, *when the total number* K *of standard blocks in* $\mathcal{X}^+$ *is* $O(1)$*, the MD algorithm exhibits nearly dimension-independent* $O(N^{-1/2})$ *rate of convergence,* which definitely is good news when solving large-scale problems. Needless to say, the rate of convergence is not the only entity of interest; what matters is the arithmetic cost of an iteration. The latter, modulo the computational effort for obtaining the first order information on f, is dominated by the computational complexity of the prox-mapping. This complexity, let us denote it $\mathcal{C}$, depends on what exactly is $\mathcal{X}$. As it was explained in section 1.7.1, *in the case of* $\mathcal{X} = \mathcal{X}^+$, $\mathcal{C}$ is $O(\sum_{\ell=1}^{K_b} \dim \mathcal{X}_\ell)$ plus the complexity of the eigenvalue decomposition of a matrix from $\mathbf{S}^{\nu^1} \times ... \times \mathbf{S}^{\nu^{K_s}}$. In particular, *when all spectahedron blocks are* ℓ_1 *balls and simplices,* $\mathcal{C}$ *is just linear in the dimension of* $\mathcal{X}^+$. Further, when $\mathcal{X}$ is cut off $\mathcal{X}^+$ by $O(1)$ linear inequalities, $\mathcal{C}$ is, essentially, the same as when $\mathcal{X} = \mathcal{X}_+$. Indeed, here computing the prox-mapping for $\mathcal{X}$ reduces

to solving the problem

$$\min_{z\in\mathcal{X}^+}\left\{\langle a,z\rangle+\omega(z):z\in\mathcal{X}^+,Az\le b\right\},\ \dim b=k=O(1),$$

or, which is the same, by duality, to solving the problem

$$\max_{\lambda\in\mathbb{R}^k_+} f_*(\lambda),\ f_*(\lambda)=\left[-b^T\lambda+\min_{z\in\mathcal{X}^+}\left[\langle a+A^T\lambda,z\rangle+\omega(z)\right]\right].$$

We are in the situation of $O(1)$ λ-variables, and thus the latter problem can be solved to machine precision in $O(1)$ steps of a simple first order algorithm, like the Ellipsoid method. The required by this method first order information for f_* "costs" computing a single prox-mapping for $\mathcal{X}^+$, so that computing the prox-mapping for $\mathcal{X}_+$ is, for all practical purposes, just by an absolute constant factor more costly than computing this mapping for $\mathcal{X}^+$.

When $\mathcal{X}$ is a "sophisticated" subset of $\mathcal{X}^+$, computing the prox-mapping for $\mathcal{X}$ may become more involved, and the outlined setup could become difficult to implement. One of potential remedies is to rewrite the problem (1.2) in the form of (1.15) with $\mathcal{X}$ extended to $\mathcal{X}^+$, with f in the role of f_0 and the constraints which cut $\mathcal{X}$ off $\mathcal{X}^+$ in the role of the functional constraints $f_1(x)\le 0$,..., $f_m(x)\le 0$ of (1.15).

1.8 Notes and Remarks

1. The very first Mirror Descent method – the *Subgradient Descent* – originates from Shor (1967) and Polyak (1967); SD is nothing but the MD algorithm with Euclidean setup: $x_{t+1}=\mathrm{argmin}_{u\in\mathcal{X}}\|(x_t-\gamma_t f'(x_t))-u\|_2$. "Non-Euclidean extensions" (i.e., the general MD scheme) originate from (Nemirovskii, 1979; Nemirovski and Yudin, 1983); the contemporary form of this scheme used in our presentation is due to Beck and Teboulle (2003). An ingenious version of the method, which also allows to recover dual solutions is proposed by Nesterov (2009). The construction presented in section 1.3 originates from (Nemirovski and Yudin, 1983), for recent version, see (Beck et al., 2009).

2. Practical performance of FOMs of the type we have considered can be improved significantly by passing to their *bundle* versions explicitly utilizing both the latest and the past first order information (in MD, only the latest first order information is used explicitly, while the past one is "loosely summarized" in the current iterate). The "Euclidean" bundle methods originate from Lemaréchal (1978) and are the subject of numerous papers

(see, e.g., Lemaréchal et al., 1981; Mifflin, 1982; Kiwiel, 1983, 1995, 1997; Schramm and Zowe, 1992; Lemaréchal et al., 1995; Kiwiel et al., 1999, and references therein). For a MD version of the bundle scheme, see (Ben-Tal and Nemirovski, 2005).

3. "Classical" Stochastic Approximation (the Euclidean setup version of the algorithm from Proposition 1.5 without averaging: $x^t = x_t$) originates from Robbins and Monro (1951) and assumes the objective f to be smooth and strongly convex; there is a huge related literature, see (Nevelson and Hasminskii, 1976; Benveniste et al., 1987, and references therein). The averaging of the trajectory which allows to extend the method to the case of nonsmooth convex minimization and plays the crucial role in FOMs for saddle point problems and variational inequalities, was introduced, in the Euclidean setup, in (Bruck, 1977; Nemirovskii and Yudin, 1978). For more results on "classical" and robust Stochastic Approximation, see, e.g., (Nemirovski and Yudin, 1983; Polyak, 1990; Polyak and Juditsky, 1992; Nemirovski and Rubinstein, 2002; Kushner and Yin, 2009; Nemirovski et al., 2009) and references therein.

4. The extensions of the MD scheme from convex minimization to convex-concave saddle point problems and variational inequalities with monotone operators originate from Nemirovskii (1981); Nemirovski and Yudin (1983). For a comprehensive presentation, see, e.g., Ben-Tal and Nemirovski (2005).

Acknowledgment. The research of the second author was partly supported by ONR grant N000140811104 and NSF grants DMI-0619977, DMS-0914785.

References

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31:167–175, 2003.

A. Beck, A. Ben-Tal, N. Guttmann-Beck, and L. Tetruashvili. The comirror algorithm for solving nonsmooth constrained convex problems. *Minerva Optimization Center, Tecnion - Israel Institute of Technology*, 2009.

A. Ben-Tal and A. Nemirovski. Non-Euclidean restricted memory level method for large-scale convex optimization. *Math. Progr.*, 102:407–456, 2005.

A. Ben-Tal, T. Margalit, and A. Nemirovski. The ordered subsets mirror

descent optimization method with applications to tomography. *SIAM J. Optim.*, 12:79–108, 2001.

A. Benveniste, M. Métivier, and P. Priouret. *Algorithmes Adaptatifs et Approximations Stochastiques*. Mason, 1987.

R. Bruck. On weak convergence of an ergodic iteration for the solution of variational inequalities with monotone operators in Hilbert space. *J. Math. Anal. Appl.*, 61(1), 1977.

K. C. Kiwiel. An aggregate subgradient method for nonsmooth convex minimization. *Math. Progr.*, 27:320–341, 1983.

K. C. Kiwiel. Proximal level bundle method for convex nondifferentiable optimization, saddle point problems and variational inequalities. *Math. Progr. Ser. B*, 69:89–109, 1995.

K. C. Kiwiel. Proximal minimization methods with generalized bregman distances. *SIAM J. on Control and Optim.*, 35:1142–1168, 1997.

K. C. Kiwiel, T. Larson, and P. O. Lindberg. The efficiency of ballstep subgradient level methods for convex optimization. *Mathematics of Oper. Res.*, 24:237–254, 1999.

H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2009.

C. Lemaréchal. Nonsmooth optimization and descent methods. *Research Report 78-4, IIASA, Laxenburg, Austria*, 1978.

C. Lemaréchal, J. J. Strodiot, and A. Bihain. On a bundle algorithm for nonsmooth optimization. O.L. Mangasarian, R.R.Meyer, and S.M. Robinson, Eds. *Nonlinear Programming 4,* Academic Press, pages 245–282, 1981.

C. Lemaréchal, A. Nemirovski, and Yu. Nesterov. New variants of bundle methods. *Math. Progr. Ser. B*, 69:111–148, 1995.

R. Mifflin. A modification and an extension of lemaréchal's algorithm for nonsmooth minimization. *Mathematical Programming Study*, 17:77–90, 1982.

A. Nemirovski and R. Rubinstein. An efficient stochastic approximation algorithm for stochastic saddle point problems. M. Dror, P. L'Ecuyer, and F. Szidarovszky, Eds. *Modeling Uncertainty: Examination of Stochastic Theory, Methods, and Applications,* Kluwer Academic Publishers, pages 155–184, 2002.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4): 1574–1609, 2009.

A.S. Nemirovski and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. J. Wiley & Sons, 1983.

A. Nemirovskii. Efficient methods for large-scale convex problems. *Ekonomika i Matematicheskie Metody (in Russian)*, 15, 1979.

A. Nemirovskii. Efficient iterative algorithms for variational inequalities with monotone operators. *Ekonomika i Matematicheskie Metody (in Russian)*, 17(2):344–359, 1981.

A. Nemirovskii and D. Yudin. On Cezari's convergence of the steepest descent method for approximating saddle points of convex-concave functions. *Soviet Math. Doklady*, 19(2), 1978.

Yu. Nesterov. A method for solving a convex programming problem with rate of convergence $o(1/k^2)$. *Soviet Math. Dokl.*, 27(2):372–376, 1983.

Yu. Nesterov. Smooth minimization of non-smooth functions. *Math. Progr.*, 103:127–152, 2005.

Yu. Nesterov. Primal-dual subgradient methods for convex problems. *Math. Progr., Ser. B*, 120(1), 2009.

M. B. Nevelson and R. Z. Hasminskii. *Stochastic Approximation and Recursive Estimation*. (Translations of Mathematical Monographs) American Mathematical Society, 1976.

B. T. Polyak. A general method for solving extremal problems. *Soviet Math. Doklady*, 174:33–36, 1967.

B. T. Polyak. New stochastic approximation type procedures. *Automatika i Telemelhanika (in Russian)*, 7:98–107, 1990.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control and Optim.*, 30:838–855, 1992.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Stat.*, 22:400–407, 1951.

H. Schramm and J. Zowe. A version of bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results. *SIAM J. Optim.*, 2:121–152, 1992.

N. Z. Shor. Generalized gradient descent with application to block programming. *Kibernetika (in Russian)*, (3), 1967.