

---

## First-Order Methods for Nonsmooth Convex Large-Scale Optimization, II: Utilizing Problem's Structure

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 present several state-of-the-art first-order methods for well-structured large-scale nonsmooth convex programs. In contrast to their black-box-oriented prototypes considered in Chapter 5, the methods in question utilize the problem structure in order to convert the original nonsmooth minimization problem into a saddle-point problem with a smooth convex-concave cost function. This reformulation allows us to accelerate the solution process significantly. As in Chapter 5, our emphasis is on methods which, under favorable circumstances, exhibit a (nearly) dimension-independent convergence rate. Along with investigating the general well-structured situation, we outline possibilities to further accelerate first-order methods by randomization.*

---

### 6.1 Introduction

The major drawback of the first-order methods (FOMs) considered in Chapter 5 is their slow convergence: as the number of steps  $t$  grows, the inaccuracy decreases as slowly as  $O(1/\sqrt{t})$ . As explained in Chapter 5, Section 5.1, this rate of convergence is unimprovable in the *unstructured* large-scale case;

however, convex problems usually have a lot of structure (otherwise, how could we know that the problem is convex?), and “good” algorithms should utilize this structure rather than be completely black-box-oriented. For example, by utilizing a problem’s structure, we usually can represent it as a linear/conic quadratic/semidefinite program (which usually is easy), and thus make the problem amenable to polynomial time interior-point methods for LP/CQP/SDP. Unfortunately, these algorithms, aimed at generating high accuracy solutions, can become prohibitively time-consuming in the large-scale case. A much cheaper way to exploit a problem’s structure when looking for medium-accuracy solutions was proposed by Nesterov (2005a); his main observation (although simple in the hindsight, it led to a real breakthrough) is that typical problems of nonsmooth convex minimization can be reformulated (and this is where a problem’s structure is used!) as *smooth* (often just bilinear) convex-concave saddle-point problems, and the latter can be solved by appropriate black-box-oriented FOMs with  $O(1/t)$  rate of convergence. More often than not, this simple observation allows for dramatic acceleration of the solution process, compared to the case where a problem’s structure is ignored while constantly staying within the scope of computationally cheap FOMs.

In Nesterov’s seminal paper (Nesterov, 2005a) the saddle-point reformulation of the (convex) problem of interest,  $\min_{x \in \mathcal{X}} f(x)$ , is used to construct a computationally cheap smooth convex approximation  $\tilde{f}$  of  $f$ , which further is minimized, at the rate  $O(1/t^2)$ , by Nesterov’s method for smooth convex minimization (Nesterov, 1983, 2005a). Since the smoothness parameters of  $\tilde{f}$  deteriorate as  $\tilde{f}$  approaches  $f$ , the accuracy to which the problem of interest can be solved in  $t$  iterations turns out to be  $O(1/t)$ ; from discussion in Section 5.1 (see item (c)), this is the best we can get in the large-scale case when solving a simple-looking problem such as  $\min_{\|x\|_2 \leq R} \|Ax - b\|_2$ . In what follows, we use as a “workhorse” the mirror prox (MP) saddle-point algorithm of Nemirovski (2004), which converges at the same rate  $O(1/t)$  as Nesterov’s smoothing, but is different from the latter algorithm. One of the reasons motivating this choice is a transparent structure of the MP algorithm (in this respect, it is just a simple-looking modification of the saddle-point mirror descent algorithm from Chapter 5, Section 5.6). Another reason is that, compared to smoothing, MP is better suited for accelerating by randomization (to be considered in Section 6.5).

The main body of this chapter is organized as follows. In Section 6.2, we present instructive examples of saddle-point reformulations of well-structured nonsmooth convex minimization problems, along with a kind of simple algorithmic calculus of convex functions admitting *bilinear* saddle-point representation. Our major workhorse — the mirror prox algorithm

with the rate of convergence  $O(1/t)$  for solving smooth convex-concave saddle-point problems — is presented in Section 6.3. In Section 6.4 we consider two special cases where the MP algorithm can be further accelerated. Another acceleration option is considered in Section 6.5, where we focus on bilinear saddle-point problems. We show that in this case, the MP algorithm, under favorable circumstances (e.g., when applied to saddle-point reformulations of  $\ell_1$  minimization problems  $\min_{\|x\|_1 \leq R} \|Ax - b\|_p$ ,  $p \in \{2, \infty\}$ ), can be accelerated by randomization — by passing from the precise first-order saddle-point oracle, which can be too time-consuming in the large-scale case, to a computationally much cheaper stochastic counterpart of this oracle.

The terminology and notation we use in this chapter follow those introduced in Sections 5.2.2, 5.6.1, and 5.7 of Chapter 5.

## 6.2 Saddle-Point Reformulations of Convex Minimization Problems

### 6.2.1 Saddle-Point Representations of Convex Functions

Let  $\mathcal{X} \subset E$  be a nonempty closed convex set in Euclidean space  $E_x$ , let  $f(x) : \mathcal{X} \rightarrow \mathbb{R}$  be a convex function, and let  $\phi(x, y)$  be a continuous convex-concave function on  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ , where  $\mathcal{Y} \subset E_y$  is a closed convex set, such that

$$\forall x \in \mathcal{X} : f(x) = \bar{\phi}(x) := \sup_{y \in \mathcal{Y}} \phi(x, y). \quad (6.1)$$

In this chapter, we refer to such a pair  $\phi, \mathcal{Y}$  as a *saddle-point representation* of  $f$ . Given such a representation, we can reduce the problem

$$\min_{x \in \mathcal{X}} f(x) \quad (6.2)$$

of minimizing  $f$  over  $\mathcal{X}$  (cf. (5.2)) to the convex-concave saddle-point (c.-c.s.p.) problem

$$\text{SadVal} = \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y), \quad (6.3)$$

(cf. (5.31)). Namely, assuming that  $\phi$  has a saddle-point on  $\mathcal{X} \times \mathcal{Y}$ , (6.2) is solvable and, invoking (5.32), we get, for all  $(x, y) \in \mathcal{X} \times \mathcal{Y}$ :

$$\begin{aligned} f(x) - \min_{\mathcal{X}} f &= \bar{\phi}(x) - \text{Opt}(P) = \bar{\phi}(x) - \text{SadVal} \\ &\leq \bar{\phi}(x) - \underline{\phi}(y) = \epsilon_{\text{sad}}(x, y). \end{aligned} \quad (6.4)$$

That is, the  $x$ -component of an  $\epsilon$ -solution to (6.3) (i.e., a point  $(x, y) \in \mathcal{X} \times \mathcal{Y}$  with  $\epsilon_{\text{sad}}(x, y) \leq \epsilon$ ) is an  $\epsilon$ -solution to (6.2):  $f(x) - \min_{\mathcal{X}} f \leq \epsilon$ .

The potential benefits of saddle-point representations stem from the fact that in many important cases a nonsmooth, but well-structured, convex function  $f$  admits an explicit saddle-point representation involving smooth function  $\phi$  and simple  $\mathcal{Y}$ ; as a result, the saddle-point reformulation (6.3) of the problem (6.2) associated with  $f$  can be much better suited for processing by FOMs than problem (6.2) as it is. Let us consider some examples (where  $S_n$ ,  $S_n^+$ ,  $\Sigma_\nu$ ,  $\Sigma_\nu^+$  are the standard flat and full-dimensional simplexes/spectahedrons, see Chapter 5, Section 5.7.1):

1.  $f(x) := \max_{1 \leq \ell \leq L} f_\ell(x) = \max_{y \in S_L} [\phi(x, y) := \sum_{\ell=1}^L y_\ell f_\ell(x)]$ ; when all  $f_\ell$  are smooth, so is  $\phi$ .
2.  $f(x) := \|Ax - b\|_p = \max_{\|y\|_q \leq 1} [\phi(x, y) := y^T(Ax - b)]$ ,  $q = \frac{p}{p-1}$ . With the same  $\phi(x, y) = y^T(Ax - b)$ , and with the coordinate wise interpretation of  $[u]_+ = \max[u, 0]$  for vectors  $u$ , we have  $f(x) := \|[Ax - b]_+\|_p = \max_{\|y\|_q \leq 1, y \geq 0} \phi(x, y)$  and  $f(x) := \min_s \|[Ax - b - sc]_+\|_p = \max_{\|y\|_q \leq 1, y \geq 0, c^T y = 0} \phi(x, y)$ . In particular,

(a) Let  $\mathcal{A}(\cdot)$  be an affine mapping. The problem

$$\text{Opt} = \min_{\xi \in \Xi} [f(\xi) := \|\mathcal{A}(\xi)\|_p] \quad (6.5)$$

with  $\Xi = \{\xi \in \mathbb{R}^n : \|\xi\|_1 \leq 1\}$  (cf. Lasso and Dantzig selector) reduces to the bilinear saddle-point problem

$$\min_{x \in S_{2n}^+} \max_{\|y\|_q \leq 1} y^T \mathcal{A}(Jx) \quad [J = [I, -I], q = \frac{p}{p-1}] \quad (6.6)$$

on the product of the standard simplex and the unit  $\|\cdot\|_q$ -ball. When  $\Xi = \{\xi \in \mathbb{R}^{m \times n} : \|\xi\|_{\text{n}} \leq 1\}$ , with  $\|\cdot\|_{\text{n}}$  being the nuclear norm (cf. nuclear norm minimization) representing  $\Xi$  as the image of the spectahedron  $\Sigma_{m+n}^+$  under the linear mapping  $x = \begin{bmatrix} u & v \\ v^T & w \end{bmatrix} \mapsto \mathcal{J}x := 2v$ , (6.5) reduces to the bilinear saddle-point problem

$$\min_{x \in \Sigma_{m+n}^+} \max_{\|y\|_q \leq 1} y^T \mathcal{A}(\mathcal{J}x); \quad (6.7)$$

(b) the SVM-type problem

$$\min_{\substack{w \in \mathbb{R}^p, \|w\| \leq R, \\ s \in \mathbb{R}}} \left\| [\mathbf{1} - \text{Diag}\{\eta\}(M^T w + s\mathbf{1})]_+ \right\|_p, \quad \mathbf{1} = [1; \dots; 1]$$

reduces to the bilinear saddle-point problem

$$\min_{\|x\| \leq 1} \max_{\substack{\|y\|_q \leq 1, \\ y \geq 0, \eta^T y = 0}} \left[ \phi(x, y) := \sum_j y_j - y^T \text{Diag}\{\eta\} R M^T x \right], \quad (6.8)$$

where  $x = w/R$ .

3. Let  $\mathcal{A}(x) = A_0 + \sum_{i=1}^n x_i A_i$  with  $A_0, \dots, A_n \in \mathbf{S}^\nu$ , and let  $S_k(A)$  be the sum of the  $k$  largest eigenvalues of a symmetric matrix  $A$ . Then  $f(x) := S_k(\mathcal{A}(x)) = \max_{y \in \Sigma_\nu, y \preceq k^{-1} I} [\phi(x, y) := k \langle y, \mathcal{A}(x) \rangle]$ .

In the above examples, except for the first one,  $\phi$  is as simple as it could be — it is just bilinear. The number of examples of this type can easily be increased due to the observation that the family of convex functions  $f$  admitting explicit bilinear saddle-point representations (b.s.p.r.'s),

$$f(x) = \max_{y \in \mathcal{Y}} [\langle y, \mathcal{A}x + a \rangle + \langle b, x \rangle + c] \quad (6.9)$$

with nonempty *compact* convex sets  $\mathcal{Y}$  (with unbounded  $\mathcal{Y}$ ,  $f$  typically would be poorly defined) admits a simple calculus. Namely, it is closed w.r.t. taking the basic convexity-preserving operations: (a) affine substitution of the argument  $x \leftarrow P\xi + p$ , (b) multiplication by nonnegative reals, (c) summation, (d) direct summation  $\{f_i(x^i)\}_{i=1}^k \mapsto f(x^1, \dots, x^k) = \sum_{i=1}^k f_i(x^i)$ , and (e) taking the maximum. Here (a) and (b) are evident, and (c) is nearly so: if

$$f_i(x) = \max_{y^i \in \mathcal{Y}_i} [\langle y^i, \mathcal{A}_i x + a^i \rangle + \langle b^i, x \rangle + c_i], \quad i = 1, \dots, k, \quad (6.10)$$

with nonempty convex compact  $\mathcal{Y}_i$ , then

$$\sum_{i=1}^k f_i(x) = \max_{y=(y^1, \dots, y^k) \in \mathcal{Y}^1 \times \dots \times \mathcal{Y}^k} \left[ \overbrace{\sum_{i=1}^k [\langle y^i, \mathcal{A}_i x + a^i \rangle + \langle b^i, x \rangle + c_i]}^{\langle y, \mathcal{A}x + a \rangle + \langle b, x \rangle + c} \right].$$

(d) is an immediate consequence of (a) and (c). To verify (e), let  $f_i$  be given by (6.10), let  $E_i$  be the embedding space of  $\mathcal{Y}_i$ , and let  $\mathcal{U}_i = \{(u^i, \lambda_i) = (\lambda_i y^i, \lambda_i) : y^i \in \mathcal{Y}_i, \lambda_i \geq 0\} \subset E_i^+ = E_i \times \mathbb{R}$ . Since  $\mathcal{Y}_i$  are convex and compact, the sets  $\mathcal{U}_i$  are closed convex cones. Now let

$$\mathcal{U} = \{y = ((u^1, \lambda_1), \dots, (u^k, \lambda_k)) \in \mathcal{U}_1 \times \dots \times \mathcal{U}_k : \sum_i \lambda_i = 1\}.$$

This set clearly is nonempty, convex, and closed; it is immediately seen that

it is bounded as well. We have

$$\begin{aligned} \max_{1 \leq i \leq k} f_i(x) &= \max_{\lambda \geq 0: \sum_i \lambda_i = 1} \sum_{i=1}^k \lambda_i f_i(x) = \max_{\lambda, y^1, \dots, y^k} \left\{ \sum_{i=1}^k [\langle \overbrace{\lambda_i y^i}^{u^i}, \mathcal{A}_i x + a^i \rangle \right. \\ &\quad \left. + \langle \lambda_i b^i, x \rangle + \lambda_i c_i] : \lambda \geq 0, \sum_i \lambda_i = 1, y^i \in \mathcal{Y}_i, 1 \leq i \leq k \right\} \\ &= \max_{u = \{(u^i, \lambda_i): 1 \leq i \leq k\} \in \mathcal{U}} \left[ \sum_{i=1}^k [\langle u^i, \mathcal{A}_i x + a^i \rangle + \langle \lambda_i b^i, x \rangle + \lambda_i c_i] \right], \end{aligned}$$

and we end up with a b.s.p.r. of  $\max_i f_i$ .

### 6.3 Mirror-Prox Algorithm

We are about to present the basic MP algorithm for the problem (6.3).

#### 6.3.1 Assumptions and Setup

Here we assume that

**A.** The closed and convex sets  $\mathcal{X}$ ,  $\mathcal{Y}$  are bounded.

**B.** The convex-concave function  $\phi(x, y) : \mathcal{Z} = \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  possesses a Lipschitz continuous gradient  $\nabla \phi(x, y) = (\nabla_x \phi(x, y), \nabla_y \phi(x, y))$ .

We set  $F(x, y) = (F_x(x, y) := \nabla_x \phi(x, y), F_y(x, y) := -\nabla_y \phi(x, y))$ , thus getting a Lipschitz continuous selection for the monotone operator associated with (6.3) (see Section 5.6.1).

The setup for the MP algorithm is given by a norm  $\|\cdot\|$  on the embedding space  $E = E_x \times E_y$  of  $\mathcal{Z}$  and by a d.-g.f.  $\omega(\cdot)$  for  $\mathcal{Z}$  compatible with this norm (cf. Section 5.2.2). For  $z \in \mathcal{Z}^o$  and  $w \in \mathcal{Z}$ , let

$$V_z(w) = \omega(w) - \omega(z) - \langle \omega'(z), w - z \rangle, \quad (6.11)$$

(cf. the definition (5.4)) and let  $z_c = \operatorname{argmin}_{w \in \mathcal{Z}} \omega(w)$ . Further, we assume that given  $z \in \mathcal{Z}^o$  and  $\xi \in E$ , it is easy to compute the prox-mapping

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

and set

$$\Omega = \max_{w \in \mathcal{Z}} V_{z_c}(w) \leq \max_{\mathcal{Z}} \omega(\cdot) - \min_{\mathcal{Z}} \omega(\cdot) \quad (6.12)$$

(cf. Chapter 5, Section 5.2.2). We also assume that we have at our disposal an upper bound  $L$  on the Lipschitz constant of  $F$  from the norm  $\|\cdot\|$  to the

conjugate norm  $\|\cdot\|_*$ :

$$\forall(z, z' \in \mathcal{Z}) : \|F(z) - F(z')\|_* \leq L\|z - z'\|. \quad (6.13)$$

### 6.3.2 The Algorithm

The MP algorithm is given by the recurrence

$$\begin{aligned} (a) : & z_1 = z_c, \\ (b) : & w_\tau = \text{Prox}_{z_\tau}(\gamma_\tau F(z_\tau)), \quad z_{\tau+1} = \text{Prox}_{z_\tau}(\gamma_\tau F(w_\tau)), \\ (c) : & z^\tau = [\sum_{s=1}^\tau \gamma_s]^{-1} \sum_{s=1}^\tau \gamma_s w_s, \end{aligned} \quad (6.14)$$

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

$$\delta_\tau = \gamma_\tau \langle F(w_\tau), w_\tau - z_{\tau+1} \rangle - V_{z_\tau}(z_{\tau+1}) \quad (6.15)$$

(cf. (5.4)). The convergence properties of the algorithm are given by the following

**Proposition 6.1.** *Under assumptions **A** and **B**:*

(i) *For every  $t \geq 1$  it holds (for notation, see (6.12) and (6.15)) that*

$$\epsilon_{\text{sad}}(z^t) \leq \left[ \sum_{\tau=1}^t \gamma_\tau \right]^{-1} \left[ \Omega + \sum_{\tau=1}^t \delta_\tau \right]. \quad (6.16)$$

(ii) *If the stepsizes satisfy the conditions  $\gamma_\tau \geq L^{-1}$  and  $\delta_\tau \leq 0$  for all  $\tau$  (which certainly is so when  $\gamma_\tau \equiv L^{-1}$ ), we have*

$$\forall t \geq 1 : \epsilon_{\text{sad}}(z^t) \leq \Omega \left[ \sum_{\tau=1}^t \gamma_\tau \right]^{-1} \leq \Omega L/t. \quad (6.17)$$

*Proof.*  $1^0$ . We start with the following basic observation:

**Lemma 6.2.** *Given  $z \in \mathcal{Z}^o$ ,  $\xi, \eta \in E$ , let  $w = \text{Prox}_z(\xi)$  and  $z_+ = \text{Prox}_z(\eta)$ . Then for all  $u \in \mathcal{Z}$  it holds that*

$$\begin{aligned} \langle \eta, w - u \rangle &\leq V_z(u) - V_{z_+}(u) + \langle \eta, w - z_+ \rangle - V_z(z_+) & (a) \\ &\leq V_z(u) - V_{z_+}(u) + \langle \eta - \xi, w - z_+ \rangle - V_z(w) - V_w(z_+) & (b) \\ &\leq V_z(u) - V_{z_+}(u) + \left[ \frac{1}{2} \|\eta - \xi\|_* \|w - z_+\| - \frac{1}{2} \|z - w\|^2 - \frac{1}{2} \|z_+ - w\|^2 \right] & (c) \\ &\leq V_z(u) - V_{z_+}(u) + \frac{1}{2} [\|\eta - \xi\|_*^2 - \|w - z\|^2] & (d) \end{aligned} \quad (6.18)$$

*Proof.* By the definition of  $z_+ = \text{Prox}_z(\eta)$  we have  $\langle \eta - \omega'(z) + \omega'(z_+), u - z_+ \rangle \geq 0$ ; we obtain (6.18.a) by rearranging terms and taking into account the definition of  $V_v(u)$ , (cf. the derivation of (5.12)). By the definition of  $w = \text{Prox}_z(\xi)$  we have  $\langle \xi - \omega'(z) + \omega'(w), z_+ - w \rangle \geq 0$ , whence  $\langle \eta, w - z_+ \rangle \leq$

$\langle \eta - \xi, w - z_+ \rangle + \langle \omega'(w) - \omega'(z), z_+ - w \rangle$ ; replacing the third term in the right-hand side of (a) with this upper bound and rearranging terms, we get (b). (c) follows from (b) due to the strong convexity of  $\omega$ , implying that  $V_v(u) \geq \frac{1}{2}\|u - v\|^2$ , and (d) is an immediate consequence of (c).  $\square$

$2^0$ . Applying Lemma 6.2 to  $z = z_\tau$ ,  $\xi = \gamma_\tau F(z_\tau)$  (which results in  $w = w_\tau$ ) and to  $\eta = \gamma_\tau F(w_\tau)$  (which results in  $z_+ = z_{\tau+1}$ ), we obtain, due to (6.18.d):

$$\begin{aligned} (a) \quad & \gamma_\tau \langle F(w_\tau), w_\tau - u \rangle \leq V_{z_\tau}(u) - V_{z_{\tau+1}}(u) + \delta_\tau \quad \forall u \in \mathcal{Z}, \\ (b) \quad & \delta_\tau \leq \frac{1}{2} [\gamma_\tau^2 \|F(w_\tau) - F(z_\tau)\|_*^2 - \|w_\tau - z_\tau\|^2] \end{aligned} \quad (6.19)$$

Summing (6.19.a) over  $\tau = 1, \dots, t$ , taking into account that  $V_{z_1}(u) = V_{z_c}(u) \leq \Omega$  by (6.12) and setting, for a given  $t$ ,  $\lambda_\tau = \gamma_\tau / \sum_{\tau=1}^t \gamma_\tau$ , we get  $\lambda_\tau \geq 0$ ,  $\sum_{\tau=1}^t \lambda_\tau = 1$ , and

$$\forall u \in \mathcal{Z} : \sum_{\tau=1}^t \lambda_\tau \langle F(w_\tau), w_\tau - u \rangle \leq A := \frac{\Omega + \sum_{\tau=1}^t \delta_\tau}{\sum_{\tau=1}^t \gamma_\tau}. \quad (6.20)$$

On the other hand, setting  $w_\tau = (x_\tau, y_\tau)$ ,  $z^t = (x^t, y^t)$ ,  $u = (x, y)$ , and using (5.37), we have

$$\sum_{\tau=1}^t \lambda_\tau \langle F(w_\tau), w_\tau - u \rangle \geq \phi(x^t, y) - \phi(x, y^t),$$

so that (6.20) results in  $\phi(x^t, y) - \phi(x, y^t) \leq A$  for all  $(x, y) \in \mathcal{Z}$ . Taking the supremum in  $(x, y) \in \mathcal{Z}$ , we arrive at (6.16); (i) is proved. To prove (ii), note that with  $\gamma_t \leq L^{-1}$ , (6.19.b) implies that  $\delta_\tau \leq 0$ , see (6.13).  $\square$

### 6.3.3 Setting up the MP Algorithm

Let us restrict ourselves to the *favorable geometry* case defined completely similarly to Chapter 5, Section 5.7.2, but with  $\mathcal{Z}$  in the role of  $\mathcal{X}$ . Specifically, we assume that  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$  is a subset of the direct product  $\mathcal{Z}^+$  of  $K$  standard blocks  $\mathcal{Z}_\ell$  ( $K_b$  ball blocks and  $K_s = K - K_b$  spectahedron blocks) and that  $\mathcal{Z}$  intersects  $\text{rint } \mathcal{Z}^+$ . We assume that the representation  $\mathcal{Z}^+ = \mathcal{Z}_1 \times \dots \times \mathcal{Z}_K$  is coherent with the representation  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ , meaning that  $\mathcal{X}$  is a subset of the direct product of some of the blocks  $\mathcal{Z}_\ell$ , while  $\mathcal{Y}$  is a subset of the direct product of the remaining blocks. We equip the embedding space  $E = E_1 \times \dots \times E_K$  of  $\mathcal{Z} \subset \mathcal{Z}^+$  with the norm  $\|\cdot\|$  and a d.-g.f.  $\omega(\cdot)$  according to (5.42) and (5.43) (where, for notational consistency, we should replace  $x^\ell$  with  $z^\ell$  and  $\mathcal{X}_\ell$  with  $\mathcal{Z}_\ell$ ). Our current goal is to optimize the efficiency estimate of the associated MP algorithm over the coefficients  $\alpha_\ell$  in (5.42), (5.43). To this end assume that we have at our disposal upper bounds

$L_{\mu\nu} = L_{\nu\mu}$  on the partial Lipschitz constants of the (Lipschitz continuous by assumption **B**) vector field  $F(z = (x, y)) = (\nabla_x \phi(x, y), -\nabla_y \phi(x, y))$ , so that for  $1 \leq \mu \leq K$  and all  $u, v \in \mathcal{Z}$ , we have

$$\|F_\mu(u) - F_\mu(v)\|_{(\mu),*} \leq \sum_{\nu=1}^K L_{\mu\nu} \|u^\nu - v^\nu\|_{(\nu)},$$

where the decomposition  $F(z = (z^1, \dots, z^K)) = (F_1(z), \dots, F_K(z))$  is induced by the representation  $E = E_1 \times \dots \times E_K$ .

Let  $\Omega_\ell$  be defined by (5.46) with  $\mathcal{Z}_\ell$  in the role of  $\mathcal{X}_\ell$ . The choice

$$\alpha_\ell = \frac{\sum_{\nu=1}^K L_{\ell\nu} \sqrt{\Omega_\nu}}{\sqrt{\Omega_\ell} \sum_{\mu,\nu} L_{\mu\nu} \sqrt{\Omega_\mu \Omega_\nu}}$$

(cf. Nemirovski, 2004) results in

$$\Omega \leq 1 \text{ and } L \leq \mathcal{L} := \sum_{\mu,\nu} L_{\mu\nu} \sqrt{\Omega_\mu \Omega_\nu},$$

so that the bound (6.17) is

$$\epsilon_{\text{sad}}(z^t) \leq \mathcal{L}/t, \quad \mathcal{L} = \sum_{\mu,\nu} L_{\mu\nu} \sqrt{\Omega_\mu \Omega_\nu}. \quad (6.21)$$

As far as complexity of a step and dependence of the efficiency estimate on a problem's dimension are concerned, the present situation is identical to that of MD (studied in Chapter 5, Section 5.7). In particular, all our considerations in the discussion at the end of Section 5.7.2 remain valid here.

### 6.3.3.1 Illustration I

As simple and instructive illustrations, consider problems (6.8) and (6.5).

**1.** Consider problem (6.8), and assume, in full accordance with the SVM origin of the problem, that  $\|w\| = \|w\|_r$  with  $r \in \{1, 2\}$ ,  $p \in \{2, \infty\}$ , and that  $\eta$  is a  $\pm 1$  vector which has both positive and negative entries. When  $p = 2$ , (6.8) is a bilinear saddle-point problem on the product of the unit  $\|\cdot\|_r$ -ball and a simple part of  $\|\cdot\|_2$ -ball. Combining (6.21) with what was said in Section 5.7.2, we arrive at the efficiency estimate

$$\epsilon_{\text{sad}}(x^t, y^t) \leq O(1)(\ln(\dim w))^{1-r/2} R \|M\|_{2,r_*} t^{-1}, \quad r_* = r/(r-1),$$

where  $\|M\|_{2,2}$  is the spectral norm of  $M$ , and  $\|M\|_{2,\infty}$  is the maximum of the Euclidean norms of the rows in  $M$ . When  $p = 1$ , the situation becomes worse: (6.8) is now a bilinear saddle-point problem on the product of the unit  $\|\cdot\|_r$ -ball and a simple subset of the unit box  $\{y : \|y\|_\infty \leq 1\}$ , or, which is the

same, a simple subset of the Euclidean ball of the radius  $\rho = \sqrt{\dim \eta}$  centered at the origin. Substituting  $y = \rho u$ , we end up with a bilinear saddle-point problem on the direct product of the unit  $\|\cdot\|_r$  ball and a simple subset of the unit Euclidean ball, the matrix of the bilinear part of the cost function being  $\rho R \text{Diag}\{\eta\} M^T$ . As a result, we arrive at the dimension-dependent efficiency estimate

$$\epsilon_{\text{sad}}(x^t, y^t) \leq O(1)(\ln(\dim w))^{1-r/2} \sqrt{\dim \eta} R \|M\|_{2, r_*} t^{-1}, \quad r_* = r/(r-1).$$

Note that in all cases the computational effort at a step of the MP is dominated by the necessity to compute  $O(1)$  matrix-vector products involving matrices  $M$  and  $M^T$ .

**2.** Now consider problem (6.5), and let  $p \in \{2, \infty\}$ .

**2.1.** Let us start with the case of  $\Xi = \{\xi \in \mathbb{R}^n : \|\xi\|_1 \leq 1\}$ , so that  $\mathcal{A}(Jx) = A_0 + Ax$ , where  $A$  is an  $m \times 2n$  matrix. Here (6.6) is a bilinear saddle-point problem on the direct product of the standard simplex  $S_{2n}^+$  in  $\mathbb{R}^{2n}$  and the unit  $\|\cdot\|_q$ -ball in  $\mathbb{R}^m$ . Combining (6.21) with derivations in Section 5.7.2, the efficiency estimate of MP is

$$\epsilon_{\text{sad}}(x^t, y^t) \leq O(1) \sqrt{\ln(n)} (\ln(m))^{\frac{1}{2} - \frac{1}{p}} [\max_{1 \leq j \leq \dim x} \|A_j\|_p] t^{-1}, \quad (6.22)$$

where  $A_j$  are columns of  $A$ . The complexity of a step is dominated by the necessity to compute  $O(1)$  matrix-vector products involving  $A$  and  $A^T$ .

**2.2.** The next case, inspired by K. Scheinberg, is the one where  $\Xi = \{(\xi^1, \dots, \xi^k) \in \mathbb{R}^{d_1} \times \dots \times \mathbb{R}^{d_k} : \sum_j \|\xi^j\|_2 \leq 1\}$ , so that problem (6.5) is of the form arising in block Lasso ( $p = 2$ ) or block Dantzig selector ( $p = \infty$ ). Setting  $\mathcal{X} = \Xi$ , let us equip the embedding space  $E_x = \mathbb{R}^{d_1} \times \dots \times \mathbb{R}^{d_k}$  of  $\mathcal{X}$  with the norm  $\|x\|_x = \sum_{i=1}^k \|x^i[x]\|_2$ , where  $x^i[x] \in \mathbb{R}^{d_i}$  are the blocks of  $x \in E_x$ , it is easily seen that the function

$$\omega_x(x) = \frac{1}{p^\gamma} \sum_{i=1}^k \|x^i[x]\|_2^p : \mathcal{X} \rightarrow \mathbf{R},$$

$$p = \begin{cases} 2, & k \leq 2 \\ 1 + 1/\ln(k), & k \geq 3 \end{cases}, \quad \gamma = \begin{cases} 1, & k = 1 \\ 1/2, & k = 2 \\ 1/(e \ln(k)), & k > 2 \end{cases}$$

is a d.-g.f. for  $\mathcal{X}$  compatible with the norm  $\|\cdot\|$ , and that the  $\omega_x(\cdot)$  diameter  $\Omega_x$  of  $\mathcal{X}$  does not exceed  $O(1) \ln(k+1)$ . Note that in the case of  $k = 1$  (where  $\mathcal{X} = \Xi$  is the unit  $\|\cdot\|_2$ -ball),  $\|\cdot\|_x = \|\cdot\|_2$ ,  $\omega_x(\cdot)$  are exactly as in the Euclidean MD setup, and in the case of  $d_1 = \dots = d_k = 1$  (where  $\mathcal{X} = \Xi$  is the unit  $\ell_1$  ball),  $\|\cdot\|_x = \|\cdot\|_1$  and  $\omega_x(\cdot)$  is, basically, the d.-g.f. from item 2b of Section 5.7.1. Applying the results of Section 6.3.3, the efficiency

estimate of MP is

$$\epsilon_{\text{sad}}(z^t) \leq O(1)(\ln(k+1))^{\frac{1}{2}}(\ln(m))^{\frac{1}{2}-\frac{1}{p}}\pi(A)t^{-1}, \quad (6.23)$$

where  $\pi(A)$  is the norm of the linear mapping  $x \mapsto Ax$  induced by the norms  $\|x\|_x = \sum_{i=1}^k \|x^i[x]\|_2$  and  $\|\cdot\|_p$  in the argument and the image spaces.

Note that the prox-mapping is easy to compute in this setup. The only nonevident part of this claim is that it is easy to minimize over  $\mathcal{X}$  a function of the form  $\omega_x(x) + \langle a, x \rangle$  or, which is the same, a function of the form  $g(x) = \frac{1}{p} \sum_{i=1}^k \|x^i[x]\|_2^p + \langle b, x \rangle$ . Here is the verification: setting  $\beta_i = \|x^i[b]\|_2$ ,  $1 \leq i \leq k$ , it is easily seen that at a minimizer  $x_*$  of  $g(\cdot)$  over  $\mathcal{X}$  the blocks  $x^i[x_*]$  are nonpositive multiples of  $x^i[b]$ , thus, all we need is to find  $\sigma_i^* = \|x^i[x_*]\|_2$ ,  $1 \leq i \leq k$ . Clearly,  $\sigma^* = [\sigma_1^*; \dots; \sigma_k^*]$  is nothing but

$$\operatorname{argmin}_{\sigma} \left\{ \frac{1}{p} \sum_{i=1}^k \sigma_i^p - \sum_{i=1}^k \beta_i \sigma_i : \sigma \geq 0, \sum_i \sigma_i \leq 1 \right\}.$$

After  $\beta_i$  are computed, the resulting “nearly separable” convex problem clearly can be solved within machine accuracy in  $O(k)$  a.o. As a result, the arithmetic cost of a step of MP in our situation is dominated by  $O(1)$  computations of matrix-vector products involving  $A$  and  $A^T$ .

Note that slightly modifying the above d.-g.f. for the unit ball  $\mathcal{X}$  of  $\|\cdot\|_x$ , we get a d.-g.f. compatible with  $\|\cdot\|_x$  on the entire  $E_x$ , namely, the function

$$\widehat{\omega}_x(x) = \frac{k^{(p-1)(2-p)/p}}{2\gamma} \left[ \sum_{i=1}^k \|x^i[x]\|_2^p \right]^{\frac{2}{p}},$$

with the same as above  $p, \gamma$ .<sup>1</sup> The associated prox-mapping is a “closed form” one. Indeed, here again the blocks  $x^i[x_*]$  of the minimizer, over the entire  $E_x$ , of  $g(x) = \frac{1}{2} \left( \sum_{i=1}^k \|x^i[x]\|_2^p \right)^{\frac{2}{p}} - \langle b, x \rangle$  are nonpositive multiples of the blocks  $x^i[b]$ ; thus, all we need is to find  $\sigma^* = [\|x^1[x_*]\|_2; \dots; \|x^k[x_*]\|_2]$ . Setting  $\beta = [\|x^1[b]\|_2; \dots; \|x^k[b]\|_2]$ , we have

$$\sigma^* = \operatorname{argmin}_{\sigma \in \mathbb{R}^k} \left\{ \frac{1}{2} \|\sigma\|_p^2 - \langle \beta, \sigma \rangle : \sigma \geq 0 \right\} = \nabla \left( \frac{1}{2} \|\beta\|_{\frac{2}{p-1}}^2 \right).$$

**2.3.** Finally, consider the case when  $\Xi$  is the unit nuclear-norm ball, so that  $\mathcal{A}(\mathcal{J}x) = a_0 + [\operatorname{Tr}(A_1x); \dots; \operatorname{Tr}(A_kx)]$  with  $A_i \in \mathbf{S}^{m+n}$ , and (6.7) is a bilinear saddle-point problem on the direct product of the spectahedron  $\Sigma_{m+n}^+$  and the unit  $\|\cdot\|_q$ -ball in  $\mathbb{R}^k$ . Applying the results of Section 6.3.3, the efficiency

1. Note that in the “extreme” cases (a):  $k = 1$  and (b):  $d_1 = \dots = d_k = 1$ , our d.-g.f. recovers the Euclidean d.-g.f. and the d.-g.f. similar to the one in item 2a of Section 5.7.1.

estimate of MP is

$$\epsilon_{\text{sad}}(x^t, y^t) \leq O(1) \sqrt{\ln(m+n)} (\ln(k))^{\frac{1}{2} - \frac{1}{p}} \left[ \max_{\|\zeta\|_2 \leq 1} \|[\zeta^T A_1 \zeta; \dots; \zeta^T A_k \zeta]\|_p \right] t^{-1}.$$

The complexity of a step is dominated by  $O(1)$  computations of the values of  $\mathcal{A}$  and of matrices of the form  $\sum_{i=1}^k y_i A_i$ , plus computing a single eigenvalue decomposition of a matrix from  $\mathbf{S}^{m+n}$ .

In all cases, the approximate solution  $(x^t, y^t)$  to the saddle-point reformulation of (6.5) straightforwardly induces a feasible solution  $\xi^t$  to the problem of interest (6.5) such that  $f(\xi^t) - \text{Opt} \leq \epsilon_{\text{sad}}(x^t, y^t)$ .

## 6.4 Accelerating the Mirror-Prox Algorithm

In what follows, we present two modifications of the MP algorithm.

### 6.4.1 Splitting

#### 6.4.1.1 Situation and Assumptions

Consider the c.-c.s.p. problem (6.3) and assume that both  $\mathcal{X}$  and  $\mathcal{Y}$  are bounded. Assume also that we are given norms  $\|\cdot\|_x, \|\cdot\|_y$  on the corresponding embedding spaces  $E_x, E_y$ , along with d.-g.f.'s  $\omega_x(\cdot)$  for  $\mathcal{X}$  and  $\omega_y(\cdot)$  for  $\mathcal{Y}$  which are compatible with the respective norms.

We already know that if the convex-concave cost function  $\phi$  is smooth (i.e., possesses a Lipschitz continuous gradient), the problem can be solved at the rate  $O(1/t)$ . We are about to demonstrate that the same holds true when, roughly speaking,  $\phi$  can be represented as a sum of a “simple” part and a smooth parts. Specifically, let us assume the following:

**C.1.** The monotone operator  $\Phi$  associated with (6.3) (see Section 5.6.1) admits splitting: we can point out a *Lipschitz continuous on  $\mathcal{Z}$*  vector field  $G(z) = (G_x(z), G_y(z)) : \mathcal{Z} \rightarrow E = E_x \times E_y$ , and a point-to-set monotone operator  $\mathcal{H}$  with the same domain as  $\Phi$  such that the sets  $\mathcal{H}(z), z \in \text{Dom } \mathcal{H}$ , are convex and nonempty, the graph of  $\mathcal{H}$  (the set  $\{(z, h) : z \in \text{Dom } \mathcal{H}, h \in \mathcal{H}(z)\}$ ) is closed, and

$$\forall z \in \text{Dom } \mathcal{H} : \mathcal{H}(z) + G(z) \subset \Phi(z). \quad (6.24)$$

**C.2.**  $\mathcal{H}$  is simple, specifically, it is easy to find a weak solution to the variational inequality associated with  $\mathcal{Z}$  and a monotone operator of the form  $\Psi(x, y) = \alpha \mathcal{H}(x, y) + [\alpha_x \omega'_x(x) + e; \alpha_y \omega'_y(y) + f]$  (where  $\alpha, \alpha_x, \alpha_y$  are

positive), that is, it is easy to find a point  $\hat{z} \in \mathcal{Z}$  satisfying

$$\forall (z \in \text{rint } \mathcal{Z}, F \in \Psi(z)) : \langle F, z - \hat{z} \rangle \geq 0. \quad (6.25)$$

It is easily seen that in the case of **C.1**, (6.25) has a unique solution  $\hat{z} = (\hat{x}, \hat{y})$  which belongs to  $\text{Dom } \Phi \cap \mathcal{Z}^o$  and in fact is a strong solution: there exists  $\zeta \in \mathcal{H}(\hat{z})$  such that

$$\forall z \in \mathcal{Z} : \langle \alpha \zeta + [\alpha_z \omega'_x(\hat{x}) + e; \alpha_y \omega'_y(\hat{y}) + f], z - \hat{z} \rangle \geq 0. \quad (6.26)$$

We assume that when solving (6.25), we get both  $\hat{z}$  and  $\zeta$ .

We intend to demonstrate that under assumptions **C.1** and **C.2** we can solve (6.3) as if there were no  $\mathcal{H}$ -component at all.

### 6.4.2 Algorithm MPa

#### 6.4.2.1 Preliminaries

Recall that the mapping  $G(x, y) = (G_x(x, y), G_y(x, y)) : \mathcal{Z} \rightarrow E$  defined in **C.1** is Lipschitz continuous. We assume that we have at our disposal nonnegative constants  $L_{xx}, L_{yy}, L_{xy}$  such that

$$\forall (z = (x, y) \in \mathcal{Z}, z' = (x', y') \in \mathcal{Z}) : \begin{cases} \|G_x(x', y) - G_x(x, y)\|_{x,*} \leq L_{xx} \|x' - x\|_x, \\ \|G_y(x, y') - G_y(x, y)\|_{y,*} \leq L_{yy} \|y' - y\|_y \\ \|G_x(x, y') - G_x(x, y)\|_{x,*} \leq L_{xy} \|y' - y\|_y, \\ \|G_y(x', y) - G_y(x, y)\|_{y,*} \leq L_{xy} \|x' - x\|_x \end{cases} \quad (6.27)$$

where  $\|\cdot\|_{x,*}$  and  $\|\cdot\|_{y,*}$  are the norms conjugate to  $\|\cdot\|_x$  and  $\|\cdot\|_y$ , respectively. We set

$$\begin{aligned} \Omega_x &= \max_{\mathcal{X}} \omega_x(\cdot) - \min_{\mathcal{X}} \omega_x(\cdot), \quad \Omega_y = \max_{\mathcal{Y}} \omega_y(\cdot) - \min_{\mathcal{Y}} \omega_y(\cdot), \\ \mathcal{L} &= L_{xx} \Omega_x + L_{xy} \Omega_y + 2L_{xy} \sqrt{\Omega_x \Omega_y}, \\ \alpha &= [L_{xx} \Omega_x + L_{xy} \sqrt{\Omega_x \Omega_y}] / \mathcal{L}, \quad \beta = [L_{yy} \Omega_y + L_{xy} \sqrt{\Omega_x \Omega_y}] / \mathcal{L}, \\ \omega(x, y) &= \frac{\alpha}{\Omega_x} \omega_x(x) + \frac{\beta}{\Omega_y} \omega_y(y) : \mathcal{Z} \rightarrow \mathbb{R}, \\ \|(x, y)\| &= \sqrt{\frac{\alpha}{\Omega_x} \|x\|_x^2 + \frac{\beta}{\Omega_y} \|y\|_y^2} \end{aligned} \quad (6.28)$$

so that the conjugate norm is  $\|(x, y)\|_* = \sqrt{\frac{\Omega_x}{\alpha} \|x\|_{x,*}^2 + \frac{\Omega_y}{\beta} \|y\|_{y,*}^2}$  (cf. Section 6.3.3). Observe that  $\omega(\cdot)$  is a d.-g.f. on  $\mathcal{Z}$  compatible with the norm  $\|\cdot\|$ . It is easily seen that  $\Omega := 1 \geq \max_{z \in \mathcal{Z}} \omega(z) - \min_{z \in \mathcal{Z}} \omega(z)$  and

$$\forall (z, z' \in \mathcal{Z}) : \|G(z) - G(z')\|_* \leq \mathcal{L} \|z - z'\|. \quad (6.29)$$

### 6.4.2.2 Algorithm MPa

Our new version, MPa, of the MP algorithm is as follows:

1. *Initialization:* Set  $z_1 = \operatorname{argmin}_Z \omega(\cdot)$ .
2. *Step*  $\tau = 1, 2, \dots$ : Given  $z_\tau \in Z^o$  and a stepsize  $\gamma_\tau > 0$ , we find  $w_\tau$  that satisfies

$$(\forall u \in \operatorname{rint} Z, F \in \mathcal{H}(u)) : \langle \gamma_\tau(F + G(z_\tau)) + \omega'(u) - \omega'(z_\tau), u - w_\tau \rangle \geq 0$$

and find  $\zeta_\tau \in \mathcal{H}(w_\tau)$  such that

$$\forall (u \in Z) : \langle \gamma_\tau(\zeta_\tau + G(z_\tau)) + \omega'(w_\tau) - \omega'(z_\tau), u - w_\tau \rangle \geq 0; \quad (6.30)$$

by assumption **C.2**, computation of  $w_\tau$  and  $\zeta_\tau$  is easy. Next, we compute

$$\begin{aligned} z_{\tau+1} &= \operatorname{Prox}_{z_\tau}(\gamma_\tau(\zeta_\tau + G(w_\tau))) \\ &:= \operatorname{argmin}_{z \in Z} [\langle \gamma_\tau(\zeta_\tau + G(w_\tau)), z \rangle + V_{z_\tau}(z)], \end{aligned} \quad (6.31)$$

where  $V(\cdot)$  is defined in (6.11). We set

$$z^\tau = \left[ \sum_{s=1}^{\tau} \gamma_s \right]^{-1} \sum_{s=1}^{\tau} \gamma_s w_s$$

and loop to step  $\tau + 1$ .

Let

$$\delta_\tau = \langle \gamma_\tau(\zeta_\tau + G(w_\tau)), w_\tau - z_{\tau+1} \rangle - V_{z_\tau}(z_{\tau+1})$$

(cf. (6.27)). The convergence properties of the algorithm are given by

**Proposition 6.3.** *Under assumptions **C.1** and **C.2**, algorithm MPa ensures that*

- (i) *For every  $t \geq 1$  it holds that*

$$\epsilon_{\text{sad}}(z^t) \leq \left[ \sum_{\tau=1}^t \gamma_\tau \right]^{-1} \left[ 1 + \sum_{\tau=1}^t \delta_\tau \right]. \quad (6.32)$$

- (ii) *If the stepsizes satisfy the condition  $\gamma_\tau \geq \mathcal{L}^{-1}$ ,  $\delta_\tau \leq 0$  for all  $\tau$  (which certainly is so when  $\gamma_\tau \equiv \mathcal{L}^{-1}$ ), we have*

$$\forall t \geq 1 : \epsilon_{\text{sad}}(z^t) \leq \left[ \sum_{\tau=1}^t \gamma_\tau \right]^{-1} \leq \mathcal{L}/t. \quad (6.33)$$

*Proof.* Relation (6.30) exactly expresses the fact that  $w_\tau = \operatorname{Prox}_{z_\tau}(\gamma_\tau(\zeta_\tau +$

$G(z_\tau)$ ). With this in mind, Lemma 6.2 implies that

$$\begin{aligned} (a) \quad & \gamma_\tau \langle \zeta_\tau + G(w_\tau), w_\tau - u \rangle \leq V_{z_\tau}(u) - V_{z_{\tau+1}}(u) + \delta_\tau \quad \forall u \in \mathcal{Z}, \\ (b) \quad & \delta_\tau \leq \frac{1}{2} [\gamma_\tau^2 \|G(w_\tau) - G(z_\tau)\|_*^2 - \|w_\tau - z_\tau\|^2] \end{aligned} \quad (6.34)$$

(cf. (6.19)). It remains to repeat word by word the reasoning in items 2<sup>0</sup>–3<sup>0</sup> of the proof of Proposition 6.1, keeping in mind (6.29) and the fact that, by the origin of  $\zeta_\tau$  and in view of (6.24), we have  $\zeta_\tau + G(w_\tau) \in \Phi(w_\tau)$ .  $\square$

### 6.4.2.3 Illustration II

Consider a problem of the Dantzig selector type

$$\text{Opt} = \min_{\|x\|_1 \leq 1} \|A^T(Ax - b)\|_\infty \quad [A : m \times n, m \leq n] \quad (6.35)$$

(cf. (6.5)) along with its saddle-point reformulation:

$$\text{Opt} = \min_{\|x\|_1 \leq 1} \max_{\|y\|_1 \leq 1} y^T [Bx - c], \quad B = A^T A, \quad c = A^T b. \quad (6.36)$$

As already mentioned, the efficiency estimate for the basic MP as applied to this problem is  $\epsilon_{\text{sad}}(z^t) \leq O(1) \sqrt{\ln(n)} \|B\|_{1,\infty} t^{-1}$ , where  $\|B\|_{1,\infty}$  is the maximum of magnitudes of entries in  $B$ . Now, in typical large-scale compressed sensing applications, columns  $A_i$  of  $A$  are of nearly unit  $\|\cdot\|_2$ -norm and are nearly orthogonal: the *mutual incoherence*  $\mu(A) = \max_{i \neq j} |A_i^T A_j| / |A_i^T A_i|$  is  $\ll 1$ . In other words, the diagonal entries in  $B$  are of order 1, and the magnitudes of off-diagonal entries do not exceed  $\mu \ll 1$ . For example, for a typical randomly selected  $A$ ,  $\mu$  is as small as  $O(\sqrt{\ln(n)}/m)$ . Now, the monotone operator associated with (6.36) admits an affine selection  $F(x, y) = (B^T y, c - Bx)$  and can be split as

$$F(x, y) = \overbrace{(Dy, -Dx)}^{\mathcal{H}(x,y)} + \overbrace{(\widehat{B}^T y, c - \widehat{B}x)}^{G(x,y)},$$

where  $D$  is the diagonal matrix with the same diagonal as in  $B$ , and  $\widehat{B} = B - D$ . Now, the domains  $\mathcal{X} = \mathcal{Y}$  associated with (6.36) are unit  $\ell_1$ -balls in the respective embedding spaces  $E_x = E_y = \mathbb{R}^n$ . Equipping  $E_x = E_y$  with the norm  $\|\cdot\|_1$ , and the unit  $\|\cdot\|_1$  ball  $\mathcal{X} = \mathcal{Y}$  in  $\mathbb{R}^n$  with the d.-g.f. presented in item 2b of Chapter 5, Section 5.7.1, we clearly satisfy **C.1** and, on a closest inspection, satisfy **C.2** as well. As a result, we can solve the problem by MPa, the efficiency estimate being  $\epsilon_{\text{sad}}(z^t) \leq O(1) \ln(n) \|\widehat{B}\|_{1,\infty} t^{-1}$ , which is much better than the estimate  $\epsilon_{\text{sad}}(z^t) \leq O(1) \ln(n) \|B\|_{1,\infty} t^{-1}$  for the plain MP (recall that we are dealing with the case of  $\mu := \|\widehat{B}\|_{1,\infty} \ll \|B\|_{1,\infty} = O(1)$ ). To see that **C.2** indeed takes place, note that in our situation, finding a

solution  $\widehat{z}$  to (6.25) reduces to solving the c.-c.s.p. problem (where  $\alpha > 0, \beta > 0, p \in (1, 2)$ )

$$\min_{\|x\|_1 \leq 1} \max_{\|y\|_1 \leq 1} \left[ \alpha \sum_i |x_i|^p - \beta \sum_i |y_i|^p + \sum_i [a_i x_i + b_i y_i + c_i x_i y_i] \right]. \quad (6.37)$$

By duality, this is equivalent to solving the c.-c.s.p. problem

$$\sup_{\mu \geq 0} \inf_{\nu \geq 0} [f(\mu, \nu) := \nu - \mu + \sum_i \min_{x_i} \max_{y_i} [\alpha |x_i|^p + \mu |x_i| - \beta |y_i|^p - \nu |y_i| + a_i x_i + b_i y_i + c_i x_i y_i]].$$

The function  $f(\mu, \nu)$  is convex-concave; computing first-order information on  $f$  reduces to solving  $n$  simple two-dimensional c.-c.s.p. problems  $\min_{x_i} \max_{y_i} [\dots]$  and, for all practical purposes, costs only  $O(n)$  operations. Then we can solve the (two-dimensional) c.-c.s.p. problem  $\max_{\mu \geq 0} \min_{\nu \geq 0} f(\mu, \nu)$  by a polynomial-time first-order algorithm, such as the saddle-point version of the Ellipsoid method (see, e.g., Nemirovski et al., 2010). Thus, solving (6.37) within machine accuracy takes just  $O(n)$  operations.

### 6.4.3 The Strongly Concave Case

#### 6.4.3.1 Situation and Assumptions

Our current goal is to demonstrate that in the situation of the previous section, assuming that  $\phi$  is strongly concave, we can improve the rate of convergence from  $O(1/t)$  to  $O(1/t^2)$ . Let us consider the c.-c.s.p. problem (6.3) and assume that  $\mathcal{X}$  is bounded (while  $\mathcal{Y}$  can be unbounded), and that we are given norms  $\|\cdot\|_x, \|\cdot\|_y$  on the corresponding embedding spaces  $E_x, E_y$ . We assume that we are also given a d.-g.f.  $\omega_x(\cdot)$ , compatible with  $\|\cdot\|_x$ , for  $\mathcal{X}$ , and a d.-g.f.  $\omega_y(\cdot)$  compatible with  $\|\cdot\|_y$ , for the entire  $E_y$  (and not just for  $\mathcal{Y}$ ). W.l.o.g. let  $0 = \operatorname{argmin}_{E_y} \omega_y$ . We keep assumption **C.1** intact and replace assumption **C.2** with its modification:

**C.2'**. It is easy to find a solution  $\widehat{z}$  to the variational inequality (6.25) associated with  $\mathcal{Z}$  and a monotone operator of the form  $\Psi(x, y) = \alpha \mathcal{H}(x, y) + [\alpha_x \omega'_x(x) + e; \alpha_y \omega'_y((y - \bar{y})/R) + f]$  (where  $\alpha, \alpha_x, \alpha_y, R$  are positive and  $\bar{y} \in \mathcal{Y}$ ).

As above, it is easily seen that  $\widehat{z} = (\widehat{x}, \widehat{y})$  is in fact a strong solution to the variational inequality: there exists  $\zeta \in \mathcal{H}(\widehat{z})$  such that

$$\langle \alpha \zeta + [\alpha_x \omega'_x(\widehat{x}) + e; \alpha_y \omega'_y((\widehat{y} - \bar{y})/R) + f], u - \widehat{z} \rangle \geq 0 \quad \forall u \in \mathcal{Z}. \quad (6.38)$$

We assume, as in the case of **C.2**, that when solving (6.25), we get both  $\widehat{z}$  and  $\zeta$ .

Furthermore, there are two new assumptions:

**C.3.** The function  $\phi$  is strongly concave with modulus  $\kappa > 0$  w.r.t.  $\|\cdot\|_y$ :

$$\forall \left( \begin{array}{l} x \in \mathcal{X}, y \in \text{rint } \mathcal{Y}, f \in \partial_y[-\phi(x, y)], \\ y' \in \text{rint } \mathcal{Y}, g \in \partial_y[-\phi(x, y')] \end{array} \right) : \langle f - g, y - y' \rangle \geq \kappa \|y - y'\|_y^2.$$

**C.4.** The  $E_x$ -component of  $G(x, y)$  is independent of  $x$ , that is,  $L_{xx} = 0$  (see (6.27)).

Note that **C.4** is automatically satisfied when  $G(\cdot) = (\nabla_x \tilde{\phi}(\cdot), -\nabla_y \tilde{\phi}(\cdot))$  comes from a bilinear component  $\tilde{\phi}(x, y) = \langle a, x \rangle + \langle b, y \rangle + \langle y, Ax \rangle$  of  $\phi$ .

Observe that since  $\mathcal{X}$  is bounded, the function  $\underline{\phi}(y) = \min_{x \in \mathcal{X}} \phi(x, y)$  is well defined and continuous on  $\mathcal{Y}$ ; by **C.3**, this function is strongly concave and thus has bounded level sets. By remark 5.1,  $\phi$  possesses saddle points, and since  $\underline{\phi}$  is strongly convex, the  $y$ -component of a saddle point is the unique maximizer  $y_*$  of  $\underline{\phi}$  on  $\mathcal{Y}$ . We set

$$\begin{aligned} x_c &= \operatorname{argmin}_{\mathcal{X}} \omega_x(\cdot), \quad \Omega_x = \max_{\mathcal{X}} \omega_x(\cdot) - \min_{\mathcal{X}} \omega_x(\cdot), \\ \Omega_y &= \max_{\|y\|_y \leq 1} \omega_y(y) - \min_y \omega_y(y) = \max_{\|y\|_y \leq 1} \omega_y(y) - \omega_y(0). \end{aligned}$$

#### 6.4.3.2 Algorithm MPb

The idea we intend to implement is the same one we used in Section 5.4 when designing MD for strongly convex optimization: all other things being equal, the efficiency estimate (5.28) is the better, the smaller the domain  $\mathcal{Z}$  (cf. the factor  $\Omega$  in (6.17)). On the other hand, when applying MP to a saddle-point problem with  $\phi(x, y)$  which is strongly concave in  $y$ , we ensure a qualified rate of convergence of  $y^t$  to  $y_*$ , and thus eventually could replace the original domain  $\mathcal{Z}$  with a smaller one by reducing the  $y$ -component. When it happens, we can run MP on this smaller domain, thus accelerating the solution process. This, roughly speaking, is what is going on in the algorithm MPb we are about to present.

**Building Blocks.** Let  $R > 0$ ,  $\bar{y} \in \mathcal{Y}$  and  $\bar{z} = (x_c, \bar{y}) \in \mathcal{Z}$ , so that  $\bar{z} \in \mathcal{Z}$ . Define the following entities:

$$\begin{aligned} \mathcal{Z}_R &= \{(x; y) \in \mathcal{Z} : \|y - \bar{y}\|_y \leq R\}, \\ \mathcal{L}_R &= 2L_{xy} \sqrt{\Omega_x \Omega_y} R + L_{yy} \Omega_y R^2, \\ \alpha &= [L_{xy} \sqrt{\Omega_x \Omega_y} R] / \mathcal{L}_R, \quad \beta = [L_{xy} \sqrt{\Omega_x \Omega_y} R + L_{yy} \Omega_y R^2] / \mathcal{L}_R, \quad (6.39) \\ \omega^{R, \bar{y}}(x, y) &= \frac{\alpha}{\Omega_x} \omega_x(x) + \frac{\beta}{\Omega_y} \omega_y([y - \bar{y}] / R), \\ \|(x, y)\| &= \sqrt{\frac{\alpha}{\Omega_x} \|x\|_x^2 + \frac{\beta}{\Omega_y R^2} \|y\|_y^2} \end{aligned}$$

with  $\|(\xi, \eta)\|_* = \sqrt{\frac{\Omega_x}{\alpha} \|\xi\|_{x,*}^2 + \frac{\Omega_y R^2}{\beta} \|\eta\|_{y,*}^2}$ . It is easily seen that  $\omega^{R, \bar{y}}$  is a d.-g.f. for  $\mathcal{Z}$  compatible with the norm  $\|\cdot\|$ ,  $\bar{z} = \operatorname{argmin}_{\mathcal{Z}} \omega^{R, \bar{y}}(\cdot)$ , and

$$\begin{aligned} (a) \quad & \max_{\mathcal{Z}_R} \omega^{R, \bar{y}}(\cdot) - \min_{\mathcal{Z}_R} \omega^{R, \bar{y}}(\cdot) \leq 1, \\ (b) \quad & \forall (z, z' \in \mathcal{Z}) : \|G(z) - G(z')\|_* \leq \mathcal{L}_R \|z - z'\|. \end{aligned} \quad (6.40)$$

For  $u \in \mathcal{Z}$  and  $z \in \mathcal{Z}^o$  we set  $V_z^{R, \bar{y}}(u) = \omega^{R, \bar{y}}(u) - \omega^{R, \bar{y}}(z) - \langle (\omega^{R, \bar{y}}(z))', u - z \rangle$  and define the prox-mapping

$$\operatorname{Prox}_z^{R, \bar{y}}(\xi) = \operatorname{argmin}_{u \in \mathcal{Z}} [\langle \xi, u \rangle + V_z^{R, \bar{y}}(u)].$$

Let  $z_1 = \bar{z}$  and  $\gamma_t > 0$ ,  $t = 1, 2, \dots$ . Consider the following recurrence  $\mathcal{B}$  (cf. Section 6.4.1):

(a) Given  $z_t \in \mathcal{Z}^o$ , we form the monotone operator  $\Psi(z) = \gamma_t \mathcal{H}(z) + (\omega^{R, \bar{y}})'(z) - (\omega^{R, \bar{y}})'(z_t) + \gamma_t G(z_t)$  and solve the variational inequality (6.25) associated with  $\mathcal{Z}$  and this operator; let the solution be denoted by  $w_t$ . Since the operator  $\Psi$  is of the form considered in **C.2'**, as a by-product of our computation we get a vector  $\zeta_t$  such that  $\forall u \in \mathcal{Z}$  :

$$\zeta_t \in \mathcal{H}(w_t) \ \& \ \langle \gamma_t [\zeta_t + G(z_t)] + (\omega^{R, \bar{y}})'(w_t) - (\omega^{R, \bar{y}})'(z_t), u - w_t \rangle \geq 0 \quad (6.41)$$

(cf. (6.38)).

(b) Compute  $z_{t+1} = \operatorname{Prox}_{z_t}^{R, \bar{y}}(\gamma_t (\zeta_t + G(w_t)))$  and

$$z^t(R, \bar{y}) \equiv (x^t(R, \bar{y}), y^t(R, \bar{y})) = \left[ \sum_{\tau=1}^t \gamma_\tau \right]^{-1} \sum_{\tau=1}^t \gamma_\tau w_\tau.$$

Let

$$F_t = \zeta_t + G(w_t), \quad \delta_t = \langle \gamma_t F_t, w_t - z_{t+1} \rangle - V_{z_t}^{R, \bar{y}}(z_{t+1}).$$

**Proposition 6.4.** *Let assumptions **C.1** and **C.2'-C.4** hold. Let the stepsizes satisfy the conditions  $\gamma_\tau \geq \mathcal{L}_R^{-1}$  and  $\delta_\tau \leq 0$  for all  $\tau$  (which certainly is so when  $\gamma_\tau = \mathcal{L}_R^{-1}$  for all  $\tau$ ).*

(i) *Assume that  $\|\bar{y} - y_*\|_y \leq R$ . Then for  $x^t = x^t(R, \bar{y})$ ,  $y^t = y^t(R, \bar{y})$  it holds that*

$$\begin{aligned} (a) \quad & \tilde{\phi}_R(x^t) - \underline{\phi}(y^t) \leq \left[ \sum_{\tau=1}^t \gamma_\tau \right]^{-1} \sum_{\tau=1}^t \gamma_\tau \langle F_\tau, w_\tau - z_* \rangle \\ & \leq \left[ \sum_{\tau=1}^t \gamma_\tau \right]^{-1} \leq \frac{\mathcal{L}_R}{t}, \\ (b) \quad & \|y^t - y_*\|_y^2 \leq \frac{2}{\kappa} [\tilde{\phi}_R(x^t) - \underline{\phi}(y^t)] \leq \frac{2\mathcal{L}_R}{\kappa t}, \end{aligned} \quad (6.42)$$

where  $\tilde{\phi}_R(x) = \max_{y \in \mathcal{Y} : \|y - \bar{y}\|_y \leq R} \phi(x, y)$ .

(ii) *Further, if  $\|\bar{y} - y_*\|_y \leq R/2$  and  $t > \frac{8\mathcal{L}_R}{\kappa R^2}$ , then  $\tilde{\phi}_R(x^t) = \bar{\phi}(x^t) :=$*

$\max_{y \in \mathcal{Y}} \phi(x^t, y)$ , and therefore

$$\epsilon_{\text{sad}}(x^t, y^t) := \bar{\phi}(x^t) - \underline{\phi}(y^t) \leq \frac{\mathcal{L}_R}{t}. \quad (6.43)$$

*Proof.* (i): Exactly the same argument as in the proof of Proposition 6.3, with (6.40.b) in the role of (6.29), shows that

$$\forall u \in \mathcal{Z} : \sum_{\tau=1}^t \gamma_{\tau} \langle F_{\tau}, z_{\tau} - u \rangle \leq V_{z_1}^{R, \bar{y}}(u) + \sum_{\tau=1}^t \delta_{\tau}$$

and that  $\delta_{\tau} \leq 0$ , provided  $\gamma_{\tau} = \mathcal{L}_R^{-1}$ . Thus, under the premise of Proposition 6.4 we have

$$\sum_{\tau=1}^t \gamma_{\tau} \langle F_{\tau}, z_{\tau} - u \rangle \leq V_{z_1}^{R, \bar{y}}(u) \quad \forall u \in \mathcal{Z}.$$

When  $u = (x, y) \in \mathcal{Z}_R$ , the right-hand side of this inequality is  $\leq 1$  by (6.40.a) and due to  $z_1 = \bar{z}$ . Using the same argument as in item 2<sup>o</sup> of the proof of Proposition 6.1, we conclude that the left-hand side in the inequality is  $\geq [\sum_{\tau=1}^t \gamma_{\tau}] [\phi(x^t, y) - \phi(x, y^t)]$ . Thus,

$$\forall u \in \mathcal{Z}_R : \phi(x^t, y) - \phi(x, y^t) \leq \left[ \sum_{\tau=1}^t \gamma_{\tau} \right]^{-1} \sum_{\tau=1}^t \gamma_{\tau} \langle F_{\tau}, z_{\tau} - u \rangle.$$

Taking the supremum of the left hand side of this inequality over  $u \in \mathcal{Z}_R$  and noting that  $\gamma_{\tau} \geq \mathcal{L}_R^{-1}$ , we arrive at (6.42.a). Further,  $\|\bar{y} - y_*\| \leq R$ , whence  $\tilde{\phi}_R(x^t) \geq \phi(x^t, y_*) \geq \underline{\phi}(y_*)$ . Since  $y_*$  is the maximizer of the strongly concave, modulus  $\kappa$  w.r.t.  $\|\cdot\|_y$ , function  $\underline{\phi}(\cdot)$  over  $\mathcal{Y}$ , we have

$$\|y^t - y_*\|_y^2 \leq \frac{2}{\kappa} [\underline{\phi}(y_*) - \underline{\phi}(y^t)] \leq \frac{2}{\kappa} [\tilde{\phi}_R(x^t) - \underline{\phi}(y^t)],$$

which is the first inequality in (6.42.b); the second inequality in (6.42.b) is given by (6.42.a). (i) is proved.

(ii): All we need to derive (ii) from (i) is to prove that under the premise of (ii), the quantities  $\bar{\phi}(x^t) := \max_{y \in \mathcal{Y}} \phi(x^t, y)$  and  $\tilde{\phi}_R(x^t) := \max_{y \in \mathcal{Y}, \|y - \bar{y}\|_y \leq R} \phi(x^t, y)$  are equal to each other. Assume that this is not the case, and let us lead this assumption to a contradiction. Looking at the definitions of  $\bar{\phi}$  and  $\tilde{\phi}_R$ , we see that in the case in question the maximizer  $\tilde{y}$  of  $\phi(x^t, y)$  over  $\mathcal{Y}_R = \{y : \|y - \bar{y}\|_y \leq R\}$  satisfies  $\|\bar{y} - \tilde{y}\|_y = R$ . Since  $\|\bar{y} - y_*\|_y \leq R/2$ , it follows that  $\|y_* - \tilde{y}\|_y \geq R/2$ . Because  $y_* \in \mathcal{Y}_R$ ,  $\tilde{y} = \operatorname{argmax}_{y \in \mathcal{Y}_R} \phi(x^t, y)$  and  $\phi(x^t, y)$  is strongly concave, modulus  $\kappa$  w.r.t.  $\|\cdot\|_y$ , we get  $\phi(x^t, y_*) \leq \phi(x^t, \tilde{y}) - \frac{\kappa}{2} \|y_* - \tilde{y}\|_y^2 \leq \phi(x^t, \tilde{y}) - \frac{\kappa R^2}{8}$ , whence

$\tilde{\phi}_R(x^t) = \phi(x^t, \tilde{y}) \geq \phi(x^t, y_*) + \frac{\kappa R^2}{8}$ . On the other hand,  $\phi(x^t, y_*) \geq \phi(y_*) \geq \phi(y^t)$ , and we arrive at  $\tilde{\phi}_R(x^t) - \phi(y^t) \geq \frac{\kappa R^2}{8}$ . At the same time, (6.42.a) says that  $\tilde{\phi}_R(x^t) - \phi(y^t) \leq \mathcal{L}_R t^{-1} < \frac{\kappa R^2}{8}$ , where the latter inequality is due to  $t > \frac{8\mathcal{L}_R}{\kappa R^2}$ . We arrive at the desired contradiction.  $\square$

**Algorithm MPb.** Let  $R_0 > 0$  and  $y^0 \in \mathcal{Y}$  such that

$$\|y^0 - y_*\| \leq R_0/2 \quad (6.44)$$

are given, and let

$$\begin{aligned} R_k &= 2^{-k/2} R_0, \\ N_k &= \text{Ceil} \left( 16\kappa^{-1} \left[ 2^{\frac{k+1}{2}} L_{xy} \sqrt{\Omega_x \Omega_y} R_0^{-1} + L_{yy} \Omega_y \right] \right), \\ M_k &= \sum_{j=1}^k N_j, \quad k = 1, 2, \dots \end{aligned}$$

Execution of MPb is split into *stages*  $k = 1, 2, \dots$ . At the beginning of stage  $k$ , we have at our disposal  $y^{k-1} \in \mathcal{Y}$  such that

$$\|y^{k-1} - y_*\|_y \leq R_{k-1}/2. \quad (I_{k-1})$$

At stage  $k$ , we compute  $(\hat{x}^k, \hat{y}^k) = z^{N_k}(R_{k-1}, y^{k-1})$ , which takes  $N_k$  steps of the recurrence  $\mathcal{B}$  (where  $R$  is set to  $R_{k-1}$  and  $\bar{y}$  is set to  $y^{k-1}$ ). The stepsize policy can be an arbitrary policy satisfying  $\gamma_\tau \geq \mathcal{L}_{R_{k-1}}^{-1}$  and  $\delta_\tau \leq 0$ , e.g.,  $\gamma_\tau \equiv L_{R_{k-1}}^{-1}$ ; see Proposition 6.4. After  $(\hat{x}^k, \hat{y}^k)$  is built, we set  $y^k = \hat{y}^k$  and pass to stage  $k+1$ .

Note that  $M_k$  is merely the total number of steps of  $\mathcal{B}$  carried out in course of the first  $k$  stages of MPb.

The convergence properties of MPb are given by the following statement (which can be derived from Proposition 6.4 in exactly the same way that Proposition 5.4 was derived from Proposition 5.3):

**Proposition 6.5.** *Let assumptions C.1, C.2'–C.4 hold, and let  $R_0 > 0$  and  $y^0 \in \mathcal{Y}$  satisfy (6.44). Then algorithm MPb maintains relations  $(I_{k-1})$  and*

$$\epsilon_{\text{sad}}(\hat{x}^k, \hat{y}^k) \leq \kappa 2^{-(k+3)} R_0^2, \quad (J_k)$$

$k = 1, 2, \dots$ . Further, let  $k_*$  be the smallest integer  $k$  such that  $k \geq 1$  and  $2^{\frac{k}{2}} \geq k R_0 \frac{L_{xy} \sqrt{\Omega_x \Omega_y}}{L_{yy} \Omega_y + \kappa}$ . Then

- for  $1 \leq k < k_*$ , we have  $M_k \leq O(1)k \frac{L_{yy} \Omega_y + \kappa}{\kappa}$  and  $\epsilon_{\text{sad}}(\hat{x}^k, \hat{y}^k) \leq \kappa 2^{-k} R_0^2$ ;
- for  $k \geq k_*$ , we have  $M_k \leq O(1)N_k$  and  $\epsilon_{\text{sad}}(\hat{x}^k, \hat{y}^k) \leq O(1) \frac{L_{xy}^2 \Omega_x \Omega_y}{\kappa M_k^2}$ .

Note that MPb behaves in the same way as the MD algorithm for strongly convex objectives (cf. Chapter 5, Section 5.4). Specifically, when the approximate solution  $y_k$  is far from the optimal solution  $y_*$ , the method converges linearly and switches to the sublinear rate (now it is  $O(1/t^2)$ ) when approaching  $y_*$ .

### 6.4.3.3 Illustration III

As an instructive application example for algorithm MPb, consider the convex minimization problem

$$\begin{aligned} \text{Opt} &= \min_{\xi \in \Xi} f(\xi), \quad f(\xi) = f_0(\xi) + \sum_{\ell=1}^L \frac{1}{2} \text{dist}^2(A_\ell \xi - b_\ell, U_\ell + V_\ell), \\ \text{dist}^2(w, W) &= \min_{w' \in W} \|w - w'\|_2^2 \end{aligned} \quad (6.45)$$

where

- $\Xi \subset E_\xi = \mathbb{R}^{n_\xi}$  is a convex compact set with a nonempty interior,  $E_\xi$  is equipped with a norm  $\|\cdot\|_\xi$ , and  $\Xi$  is equipped with a d.-g.f.  $\omega_\xi(\xi)$  compatible with  $\|\cdot\|_\xi$ ;
- $f_0(\xi) : \Xi \rightarrow \mathbb{R}$  is a simple continuous convex function, “simple” meaning that it is easy to solve auxiliary problems

$$\min_{\xi \in \Xi} \{ \alpha f_0(\xi) + a^T \xi + \beta \omega_\xi(\xi) \} \quad [\alpha, \beta > 0]$$

- $U_\ell \subset \mathbb{R}^{m_\ell}$  are convex compact sets such that computing metric projection  $\text{Proj}_{U_\ell}(u) = \text{argmin}_{u' \in U_\ell} \|u - u'\|_2$  onto  $U_\ell$  is easy;
- $V_\ell \subset \mathbb{R}^{m_\ell}$  are polytopes given as  $V_\ell = \text{Conv}\{v_{\ell,1}, \dots, v_{\ell,n_\ell}\}$ .

On a close inspection, problem (6.45) admits a saddle-point reformulation. Specifically, recalling that  $S_k = \{x \in \mathbb{R}_+^k : \sum_i x_i = 1\}$  and setting

$$\begin{aligned} \mathcal{X} &= \{x = [\xi; x^1; \dots; x^L] \in \Xi \times S_{n_1} \times \dots \times S_{n_L}\} \subset E_x = \mathbb{R}^{n_\xi + n_1 + \dots + n_L}, \\ \mathcal{Y} &= E_y := \mathbb{R}_{y^1}^{m_1} \times \dots \times \mathbb{R}_{y^L}^{m_L}, \\ g(y = (y^1, \dots, y^L)) &= \sum_{\ell} g_\ell(y^\ell), \quad g_\ell(y^\ell) = \frac{1}{2} [y^\ell]^T y^\ell + \max_{u_\ell \in U_\ell} u_\ell^T y^\ell, \\ B_\ell &= [v_{\ell,1}, \dots, v_{\ell,n_\ell}], \\ A[\xi; x^1; \dots; x^L] - b &= [A_1 \xi - B_1 x^1; \dots; A_L \xi - B_L x^L] - [b_1; \dots; b_L], \\ \phi(x, y) &= f_0(\xi) + y^T [Ax - b] - g(y), \end{aligned}$$

we get a continuous convex-concave function  $\phi$  on  $\mathcal{X} \times \mathcal{Y}$  such that

$$f(\xi) = \min_{\eta=(x^1, \dots, x^L): (\xi, \eta) \in \mathcal{X}} \max_{y \in \mathcal{Y}} \phi((\xi, \eta), y),$$

so that if a point  $(x = [\xi; x^1; \dots; x^L], y = [y^1; \dots; y^L]) \in \mathcal{X} \times \mathcal{Y}$  is an  $\epsilon$ -solution

to the c.-c.s.p. problem  $\inf_{x \in \mathcal{X}} \sup_y \phi(x, y)$ ,  $\xi$  is an  $\epsilon$ -solution to the problem of interest (6.45):

$$\epsilon_{\text{sad}}(x, y) \leq \epsilon \Rightarrow f(\xi) - \text{Opt} \leq \epsilon.$$

Now we apply algorithm MPb to the saddle-point problem  $\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$ . The required setup is as follows:

1. Given positive  $\alpha, \alpha_1, \dots, \alpha_L$  (parameters of the construction), we equip the embedding space  $E_x$  of  $\mathcal{X}$  with the norm

$$\|[\xi; x^1; \dots; x^L]\|_x = \sqrt{\alpha \|\xi\|^2 + \sum_{\ell=1}^L \alpha_\ell \|x^\ell\|_1^2},$$

and  $\mathcal{X}$  itself with the d.-g.f.

$$\omega_x([\xi; x^1; \dots; x^L]) = \alpha \omega_\xi(\xi) + \sum_{\ell=1}^L \alpha_\ell \text{Ent}(x^\ell), \quad \text{Ent}(u) = \sum_{i=1}^{\dim u} u_i \ln u_i,$$

which, it can immediately be seen, is compatible with  $\|\cdot\|_x$ .

2. We equip  $\mathcal{Y} = E_y = \mathbb{R}_y^{m_1 + \dots + m_L}$  with the standard Euclidean norm  $\|y\|_2$  and the d.-g.f.  $\omega_y(y) = \frac{1}{2} y^T y$ .

3. The monotone operator  $\Phi$  associated with  $(\phi, z)$  is

$$\Phi(x, y) = \{\partial_x[\phi(x, y) + \chi x(x)]\} \times \{\partial_y[-\phi(x, y)]\}, \quad \chi_Q(u) = \begin{cases} 0, & u \in Q \\ +\infty, & u \notin Q \end{cases}.$$

We define its splitting, required by **C.1**, as

$$\begin{aligned} \mathcal{H}(x, y) &= \{\{\partial_\xi[f_0(\xi) + \chi \Xi(\xi)] \times \{0\} \dots \times \{0\}\} \times \{\partial_y[\sum_{\ell=1}^L g_\ell(y^\ell)]\}\}, \\ G(x, y) &= (\nabla_x[y^T[Ax - b]]) = A^T y, \quad -\nabla_y[y^T[Ax - b]] = b - Ax. \end{aligned}$$

With this setup, we satisfy **C.1** and **C.3-C.4** (**C.3** is satisfied with  $\kappa = 1$ ). Let us verify that **C.2'** is satisfied as well. Indeed, in our current situation, finding a solution  $\hat{z}$  to (6.25) means solving the pair of convex optimization problems

$$\begin{aligned} (a) \quad & \min_{[\xi; x^1; \dots; x^L] \in \mathcal{X}} [p\alpha\omega_\xi(\xi) + qf_0(\xi) + e^T \xi \\ & \quad \quad \quad + \sum_{\ell=1}^L [p\alpha_\ell \text{Ent}(x^\ell) + e_\ell^T x^\ell]] \\ (b) \quad & \min_{y=[y^1; \dots; y^L]} \sum_{\ell=1}^L [\frac{r}{2}[y^\ell]^T y^\ell + sg_\ell(y^\ell) + f_\ell^T y^\ell] \end{aligned} \tag{6.46}$$

where  $p, q, r$ , and  $s$  are positive. Due to the direct product structure of  $\mathcal{X}$ , (6.46.a) decomposes into the uncoupled problems  $\min_{\xi \in \Xi} [p\alpha\omega_\xi(\xi) + f_0(\xi) + e^T \xi]$  and  $\min_{x^\ell \in \mathcal{S}_\ell} [p\alpha_\ell \text{Ent}(x^\ell) + e_\ell^T x^\ell]$ . We have explicitly assumed that the first of these problems is easy; the remaining ones admit closed form

solutions (cf. (5.39)). (6.46.b) also is easy: a simple computation yields  $y^\ell = -\frac{1}{r+s}[s\text{Proj}_{U_\ell}(-s^{-1}f_\ell) + f_\ell]$ , and it was assumed that it is easy to project onto  $U_\ell$ .

The bottom line is that we can solve (6.45) by algorithm MPb, the resulting efficiency estimate being

$$f(\widehat{\xi}^k) - \text{Opt} \leq O(1) \frac{L_{xy}^2 \Omega_x}{M_k^2}, k \geq k_* = O(1) \ln(R_0 L_{xy} \sqrt{\Omega_x} + 2)$$

(see Proposition 6.5 and take into account that we are in the situation of  $\kappa = 1, \Omega_y = \frac{1}{2}, L_{yy} = 0$ ). We can further use the parameters  $\alpha, \alpha_1, \dots, \alpha_L$  to optimize the quantity  $L_{xy}^2 \Omega_x$ . A rough optimization leads to the following: let  $\mu_\ell$  be the norm of the linear mapping  $\xi \rightarrow A_\ell \xi$  induced by the norms  $\|\cdot\|_\xi, \|\cdot\|_2$  in the argument and the image spaces, respectively, and let  $\nu_\ell = \max_{1 \leq j \leq n_\ell} \|v_{\ell,j}\|_2$ . Choosing

$$\alpha = \sum_{\ell=1}^L \mu_\ell^2, \alpha_\ell = \nu_\ell^2, 1 \leq \ell \leq L$$

results in  $L_{xy}^2 \Omega_x \leq O(1) [\Omega_\xi \sum_{\ell} \mu_\ell^2 + \sum_{\ell} \nu_\ell^2 \ln(n_\ell + 1)]$ ,  $\Omega_\xi = \max_{\Xi} \omega_\xi(\cdot) - \min_{\Xi} \omega_\xi(\cdot)$ .

## 6.5 Accelerating First-Order Methods by Randomization

We have seen in Section 6.2.1 that many important well-structured convex minimization programs reduce to just *bilinear* saddle-point problems

$$\text{SadVal} = \min_{x \in \mathcal{X} \subset E_x} \max_{y \in \mathcal{Y} \subset E_y} [\phi(x, y) := \langle a, x \rangle + \langle y, Ax - b \rangle], \quad (6.47)$$

the corresponding monotone operator admitting an *affine* selection

$$\begin{aligned} F(z = (x, y)) &= (a + A^T y, b - Ax) = (a, b) + \mathcal{F}z, \\ \mathcal{F}(x, y) &= (A^T y, -Ax). \end{aligned} \quad (6.48)$$

Computing the value of  $F$  requires two matrix-vector multiplications involving  $A$  and  $A^T$ . When  $\mathcal{X}, \mathcal{Y}$  are simple and the problem is large-scale with dense  $A$  (which is the case in many machine learning and signal processing applications), these matrix-vector multiplications dominate the computational cost of an iteration of an FOM; as the sizes of  $A$  grow, these multiplications can become prohibitively time consuming. The idea of what follows is that *matrix-vector multiplications is easy to randomize, and this randomization, under favorable circumstances, allows for dramatic acceleration of FOMs in the extremely large-scale case.*

### 6.5.1 Randomizing Matrix-Vector Multiplications

Let  $u \in \mathbf{R}^n$ . Computing the image of  $u$  under a linear mapping  $u \mapsto Bu = \sum_{j=1}^n u_j b_j : \mathbf{R}^n \rightarrow E$  are easy to randomize: treat the vector  $[|u_1|; \dots; |u_n|]/\|u\|_1$  as a probability distribution on the set  $\{b_1, \dots, b_n\}$ , draw from this distribution a sample  $b_j$  and set  $\xi_u = \|u\|_1 \text{sign}(u_j) b_j$ , thus getting an unbiased ( $\mathbf{E}\{\xi_u\} = Bu$ ) random estimate of  $Bu$ . When  $b_j$  are represented by readily available arrays, the arithmetic cost of sampling from the distribution  $P_u$  of  $\xi_u$ , modulo the setup cost  $O(n)$  a.o. of computing the cumulative distribution  $\{\|u\|_1^{-1} \sum_{i=1}^j |u_i|\}_{j=1}^n$  is just  $O(\ln(n))$  a.o. to generate  $j$  plus  $O(\dim E)$  a.o. to compute  $\|u\|_1 \text{sign}(u_j) b_j$ . Thus, the total cost of getting a single realization of  $\xi_u$  is  $O(n) + \dim E$ . For large  $n$  and  $\dim E$  this is much less than the cost  $O(n \dim E)$ , assuming  $b_j$  are dense, of a straightforward precise computation of  $Bu$ .

We can generate a number  $k$  of independent samples  $\xi^\ell \sim P_u$ ,  $\ell = 1, \dots, k$ , and take, as an unbiased estimate of  $Bu$ , the average  $\xi = \frac{1}{k} \sum_{\ell=1}^k \xi^\ell$ , thus reducing the estimate's variability; with this approach, the setup cost is paid only once.

### 6.5.2 Randomized Algorithm for Solving Bilinear Saddle-Point Problem

We are about to present a randomized version MPr of the mirror-prox algorithm for solving the bilinear saddle-point problem (6.47).

#### 6.5.2.1 Assumptions and Setup

1. As usual, we assume that  $\mathcal{X}$  and  $\mathcal{Y}$  are nonempty compact convex subsets of Euclidean spaces  $E_x, E_y$ ; these spaces are equipped with the respective norms  $\|\cdot\|_x, \|\cdot\|_y$ , while  $\mathcal{X}, \mathcal{Y}$  are equipped with d.-g.f.'s  $\omega_x(\cdot), \omega_y(\cdot)$  compatible with  $\|\cdot\|_x$ , resp.,  $\|\cdot\|_y$ , and define  $\Omega_x, \Omega_y$  according to (6.28). Further, we define  $\|A\|_{x,y}$  as the norm of the linear mapping  $x \mapsto Ax : E_x \rightarrow E_y$  induced by the norms  $\|\cdot\|_x, \|\cdot\|_y$  on the argument and the image spaces.

2. We assume that every point  $u \in \mathcal{X}$  is associated with a probability distribution  $\Pi_u$  supported on  $\mathcal{X}$  such that  $\mathbf{E}_{\xi \sim \Pi_u}\{\xi\} = u$ , for all  $u \in \mathcal{X}$ . Similarly, we assume that every point  $v \in \mathcal{Y}$  is associated with a probability distribution  $P_v$  on  $E_y$  with a bounded support and such that  $\mathbf{E}_{\eta \sim P_v}\{\eta\} = v$  for all  $v \in \mathcal{Y}$ . We refer to the case when  $P_v$ , for every  $v \in \mathcal{Y}$ , is supported on  $\mathcal{Y}$ , as the *inside* case, as opposed to the *general* case, where support of  $P_v$ ,  $v \in \mathcal{Y}$ , does not necessarily belong to  $\mathcal{Y}$ . We will use  $\Pi_x, P_y$  to randomize matrix-vector multiplications. Specifically, given two positive

integers  $k_x, k_y$  (parameters of our construction), and given  $u \in \mathcal{X}$ , we build a randomized estimate of  $Au$  as  $A\xi_u$ , where  $\xi_u = \frac{1}{k_x} \sum_{i=1}^{k_x} \xi_i$  and  $\xi_i$  are sampled, independently of each other, from  $\Pi_u$ . Similarly, given  $v \in \mathcal{Y}$ , we estimate  $A^T v$  by  $A^T \eta_v$ , where  $\eta_v = \frac{1}{k_y} \sum_{i=1}^{k_y} \eta_i$ , with  $\eta_i$  sampled independently of each other from  $P_v$ . Note that  $\xi_v \in \mathcal{X}$ , and in the inside case  $\eta_u \in \mathcal{Y}$ . Of course, a randomized estimation of  $Au, A^T v$  makes sense only when computing  $A\xi, \xi \in \text{supp}(\Pi_u), A^T \eta, \eta \in \text{supp}(P_v)$  is much easier than computing  $Au, A^T v$  for a general type  $u$  and  $v$ .

We introduce the quantities

$$\begin{aligned} \sigma_x^2 &= \sup_{u \in \mathcal{X}} \mathbf{E}\{\|A[\xi_u - u]\|_{y,*}^2\}, \quad \sigma_y^2 = \sup_{v \in \mathcal{Y}} \mathbf{E}\{A^T[\eta_v - v]\|_{x,*}^2\}, \\ \Theta &= 2 [\Omega_x \sigma_y^2 + \Omega_y \sigma_x^2]. \end{aligned} \quad (6.49)$$

where  $\xi_u, \eta_v$  are the random vectors just defined, and, as always,  $\|\cdot\|_{x,*}, \|\cdot\|_{y,*}$  are the norms conjugate to  $\|\cdot\|_x$  and  $\|\cdot\|_y$ .

**3.** The setup for the algorithm MPr is given by the norm  $\|\cdot\|$  on  $E = E_x \times E_y \supset \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ , and the compatible with this norm d.-g.f.  $\omega(\cdot)$  for  $\mathcal{Z}$  which are given by

$$\|(x, y)\| = \sqrt{\frac{1}{2\Omega_x} \|x\|_x^2 + \frac{1}{2\Omega_y} \|y\|_y^2}, \quad \omega(x, y) = \frac{1}{2\Omega_x} \omega_x(x) + \frac{1}{2\Omega_y} \omega_y(y),$$

so that

$$\|(\xi, \eta)\|_* = \sqrt{2\Omega_x \|\xi\|_{x,*}^2 + 2\Omega_y \|\eta\|_{y,*}^2}. \quad (6.50)$$

For  $z \in \mathcal{Z}^o, w \in \mathcal{Z}$  let (cf. the definition (5.4))

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

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

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

It can immediately be seen that

$$\Omega[\mathcal{Z}] = \max_{\mathcal{Z}} \omega(\cdot) - \min_{\mathcal{Z}} \omega(\cdot) = 1. \quad (6.51)$$

and the affine monotone operator  $F(z)$  given by (6.48) satisfies the relation

$$\forall z, z' : \|F(z) - F(z')\|_* \leq \mathcal{L} \|z - z'\|, \quad \mathcal{L} = 2\|A\|_{x,y} \sqrt{\Omega_x \Omega_y}. \quad (6.52)$$

### 6.5.2.2 Algorithm

For simplicity, we present here the version of MPr where the number of steps,  $N$ , is fixed in advance. Given  $N$ , we set

$$\gamma = \min \left[ \frac{1}{\sqrt{3\mathcal{L}}}, \frac{1}{\sqrt{3\Theta N}} \right] \quad (6.53)$$

and run  $N$  steps of the following randomized recurrence:

1. *Initialization:* We set  $z_1 = \operatorname{argmin}_z \omega(\cdot)$ .
2. *Step  $t = 1, 2, \dots, N$ :* Given  $z_t = (x_t, y_t) \in \mathcal{Z}^o$ , we generate  $\xi_{x_t}, \eta_{y_t}$  as explained above, set  $\zeta_t = (\xi_{x_t}, \eta_{y_t})$ , and compute  $F(\zeta_t) = (a + A^T \eta_{y_t}, b - A \xi_{x_t})$  and

$$w_t = (\hat{x}_t, \hat{y}_t) = \operatorname{Prox}_{z_t}(\gamma F(\zeta_t)).$$

We next generate  $\xi_{\hat{x}_t}, \eta_{\hat{y}_t}$  as explained above, set  $\hat{\zeta}_t = (\xi_{\hat{x}_t}, \eta_{\hat{y}_t})$ , and compute  $F(\hat{\zeta}_t) = (a + A^T \eta_{\hat{y}_t}, b - A \xi_{\hat{x}_t})$  and  $z_{t+1} = \operatorname{Prox}_{z_t}(\gamma F(\hat{\zeta}_t))$ .

3. *Termination  $t = N$ :* we output

$$z^N = (x^N, y^N) = \frac{1}{N} \sum_{t=1}^N (\xi_{\hat{x}_t}, \eta_{\hat{y}_t}), \quad \text{and} \quad F(z^N) = \frac{1}{N} \sum_{t=1}^N F(\hat{\zeta}_t)$$

(recall that  $F(\cdot)$  is affine).

The efficiency estimate of algorithm MPr is given by the following

**Proposition 6.6.** *For every  $N$ , the random approximate solution  $z^N = (x^N, y^N)$  generated by algorithm MPr possesses the following properties:*

- (i) *In the inside case,  $z^N \in \mathcal{Z}$  and*

$$\mathbf{E}\{\epsilon_{\text{sad}}(z^N)\} \leq \epsilon_N := \max \left[ \frac{2\sqrt{3\Theta}}{\sqrt{N}}, \frac{4\sqrt{3}\|A\|_{x,y}\sqrt{\Omega_x\Omega_y}}{N} \right]; \quad (6.54)$$

- (ii) *In the general case,  $x^N \in \mathcal{X}$  and  $\mathbf{E}\{\bar{\phi}(x^N)\} - \min_{\mathcal{X}} \bar{\phi} \leq \epsilon_N$ .*

Observe that in the general case we do not control the error  $\epsilon_{\text{sad}}(z^N)$  of the saddle-point solution. Yet the bound (ii) of Proposition 6.6 allows us to control the accuracy  $f(x^N) - \min_{\mathcal{X}} f$  of the solution  $x^N$  when the saddle-point problem is used to minimize the convex function  $f = \bar{\phi}$  (cf. (6.4)).

*Proof.* Setting  $F_t = F(\zeta_t), \hat{F}_t = F(\hat{\zeta}_t), F_t^* = F(z_t), \hat{F}_t^* = F(w_t), V_z(u) =$

$\omega(u) - \omega(z) - \langle \omega'(z), u - z \rangle$  and invoking Lemma 6.2, we get

$$\begin{aligned} \forall u \in \mathcal{Z} : \gamma \langle \widehat{F}_t, w_t - u \rangle &\leq V_{z_t}(u) - V_{z_{t+1}}(u) + \Delta_t, \\ \Delta_t &= \frac{1}{2} \left[ \gamma^2 \|F_t - \widehat{F}_t\|_*^2 - \|z_t - w_t\|^2 \right], \end{aligned}$$

whence, taking into account that  $V_{z_1}(u) \leq \Omega[\mathcal{Z}] = 1$  (see (6.51)) and that  $V_{z_{N+1}}(u) \geq 0$ ,

$$\forall u = (x, y) \in \mathcal{Z} : \gamma \sum_{t=1}^N \langle \widehat{F}_t, \widehat{\zeta}_t - u \rangle \leq 1 + \overbrace{\sum_{t=1}^N \Delta_t}^{\alpha_N} + \overbrace{\gamma \sum_{t=1}^N \langle \widehat{F}_t, \widehat{\zeta}_t - w_t \rangle}^{\beta_N}.$$

Substituting the values of  $\widehat{F}_t$  and taking expectations, the latter inequality (where the right-hand side is independent of  $u$ ) implies that

$$\begin{aligned} \mathbf{E} \left\{ \max_{(x,y) \in \mathcal{Z}} \gamma N [\phi(x^N, y) - \phi(x, y^N)] \right\} &\leq 1 + \mathbf{E}\{\alpha_N\} + \mathbf{E}\{\beta_N\}, \quad (6.55) \\ \beta_N &= \gamma \sum_{t=1}^N [\langle a, \xi_{\widehat{x}_t} - \widehat{x}_t \rangle + \langle b, \eta_{\widehat{y}_t} - \widehat{y}_t \rangle + \langle A \xi_{\widehat{x}_t}, \widehat{y}_t \rangle - \langle A^T \eta_{\widehat{x}_t}, \widehat{x}_t \rangle]. \end{aligned}$$

Now let  $\mathbf{E}_{w_t}\{\cdot\}$  stand for the expectation conditional to the history of the solution process up to the moment when  $w_t$  is generated. We have  $\mathbf{E}_{w_t}\{\xi_{\widehat{x}_t}\} = \widehat{x}_t$  and  $\mathbf{E}_{w_t}\{\eta_{\widehat{y}_t}\} = \widehat{y}_t$ , so that  $\mathbf{E}\{\beta_N\} = 0$ . Further, we have

$$\Delta_t \leq \frac{1}{2} \left[ 3\gamma^2 \left[ \|\widehat{F}_t^* - F_t^*\|_*^2 + \|\widehat{F}_t^* - \widehat{F}_t\|_*^2 + \|F_t^* - F_t\|_*^2 \right] - \|z_t - w_t\|^2 \right]$$

and, recalling the origin of  $F$ s,  $\|F_t^* - \widehat{F}_t^*\|_* \leq \mathcal{L}\|z_t - w_t\|$  by (6.52). Since  $3\gamma^2 \leq \mathcal{L}^2$  by (6.53), we get

$$\mathbf{E}\{\Delta_t\} \leq \frac{3\gamma^2}{2} \mathbf{E}\{\|\widehat{F}_t^* - \widehat{F}_t\|_*^2 + \|F_t^* - F_t\|_*^2\} \leq 3\gamma^2 \Theta,$$

where the concluding inequality is due to the definitions of  $\Theta$  and of the norm  $\|\cdot\|_*$  (see (6.49) and (6.50), respectively). Thus, (6.55) implies that

$$\mathbf{E} \left\{ \max_{(x,y) \in \mathcal{Z}} [\phi(x^N, y) - \phi(x, y^N)] \right\} \leq 1/(N\gamma) + 3\Theta\gamma \leq \epsilon_N, \quad (6.56)$$

due to the definition of  $\epsilon_N$ . Now, in the inside case we clearly have  $(x^N, y^N) \in \mathcal{Z}$ , and therefore (6.56) implies (6.54). In the general case we have  $x^N \in \mathcal{X}$ . In addition, let  $x_*$  be the  $x$ -component of a saddle point of  $\phi$  on  $\mathcal{Z}$ . Replacing in the left-hand side of (6.56) maximization over all pairs  $(x, y)$  from  $\mathcal{Z}$  with maximization only over the pair  $(x_*, y)$  with  $y \in \mathcal{Y}$  (which can only decrease

the left-hand side), we get from (6.56) that

$$\mathbf{E}\{\bar{\phi}(x^N)\} \leq \epsilon_N + \mathbf{E}\{\phi(x_*, y^N)\} = \epsilon_N + \phi(x_*, \mathbf{E}\{y^N\}). \quad (6.57)$$

Observe that  $\mathbf{E}_{w_t}\{\eta_{\hat{y}_t}\} = \hat{y}_t \in \mathcal{Y}$ . We conclude that

$$\mathbf{E}\{y^N\} = \mathbf{E}\left\{\frac{1}{N} \sum_{t=1}^N \eta_{\hat{y}_t}\right\} = \mathbf{E}\left\{\frac{1}{N} \sum_{t=1}^N \hat{y}_t\right\} \in \mathcal{Y}.$$

Thus, the right-hand side in (6.57) is  $\leq \epsilon_N + \text{SadVal}$ , and (ii) follows.  $\square$

**Remark 6.1.** We stress here that *MPr*, along with the approximate solution  $(x^N, y^N)$ , returns the value  $F(x^N, y^N)$ . This allows for easy computation, not requiring matrix-vector multiplications, of  $\bar{\phi}(x^N)$  and  $\underline{\phi}(y^N)$ .

### 6.5.2.3 Illustration IV: $\ell_1$ -Minimization

Consider problem (6.5) with  $\Xi = \{\xi \in \mathbb{R}^n : \|\xi\|_1 \leq 1\}$ . Representing  $\Xi$  as the image of the standard simplex  $\mathbb{S}_{2n} = \{x \in \mathbb{R}_+^{2n} : \sum_i x_i = 1\}$  under the mapping  $x \mapsto J_n x$ ,  $J_n = [I_n, -I_n]$ , the problem reads

$$\text{Opt} = \min_{x \in \mathbb{S}_{2n}} \|Ax - b\|_p \quad [A \in \mathbb{R}^{m \times 2n}]. \quad (6.58)$$

We consider two cases:  $p = \infty$  (*uniform fit*, as in the Dantzig selector) and  $p = 2$  ( *$\ell_2$ -fit*, as in Lasso).

**Uniform Fit.** Here (6.58) can be converted into the bilinear saddle-point problem

$$\text{Opt} = \min_{x \in \mathbb{S}_{2n}} \max_{y \in \mathbb{S}_{2m}} [\phi(x, y) := y^T J_m^T [Ax - b]]. \quad (6.59)$$

Setting  $\|\cdot\|_x = \|\cdot\|_1$ ,  $\omega_x(x) = \text{Ent}(x)$ ,  $\|\cdot\|_y = \|\cdot\|_1$ ,  $\omega_y(y) = \text{Ent}(y)$ , let us specify  $\Pi_u$ ,  $u \in \mathbb{S}_{2n}$ , and  $P_v$ ,  $v \in \mathbb{S}_{2m}$ , according to the recipe from Section 6.5.1, that is, the random vector  $\xi_u \sim \Pi_u$  with probability  $u_i$  is the  $i$ th basic orth,  $i = 1, \dots, m$ , and similarly for  $\eta_v \sim P_v$ . This is the inside case, and when we set  $\|A\|_{1,\infty} = \max_{i,j} |A_{ij}|$ , we get  $\sigma_x^2 = O(1) \frac{\|A\|_{1,\infty}^2 \ln(2m)}{k_x + \ln(2m)}$ ,  $\sigma_y^2 =$

$$O(1) \frac{\|A\|_{1,\infty}^2 \ln(2n)}{k_y + \ln(2n)}, 2 \text{ and}$$

$$\begin{aligned} \Omega_x &= \ln(2n), \quad \Omega_y = \ln(2m), \quad \mathcal{L} = 2\|A\|_{1,\infty} \sqrt{\ln(2n) \ln(2m)}, \\ \Theta &\leq O(1) \|A\|_{1,\infty}^2 \left[ \frac{\ln^2(2m)}{k_x + \ln(2m)} + \frac{\ln^2(2n)}{k_y + \ln(2n)} \right] \end{aligned}$$

In this setting Proposition 6.6 reads:

**Corollary 6.7.** *For all positive integers  $k_x, k_y, N$  one can find a random feasible solution  $(x^N, y^N)$  to (6.59) along with the quantities  $\bar{\phi}(x^N) = \|Ax^N - b\|_\infty \geq \text{Opt}$  and a lower bound  $\underline{\phi}(y^N)$  on  $\text{Opt}$  such that*

$$\text{Prob} \left\{ \bar{\phi}(x^N) - \underline{\phi}(y^N) \leq O(1) \frac{\|A\|_{1,\infty} \ln(2mn)}{\sqrt{N} \sqrt{\min[N, k_x + \ln(2m), k_y + \ln(2n)]}} \right\} \geq \frac{1}{2} \quad (6.60)$$

in  $N$  steps, the computational effort per step dominated by the necessity to extract  $2k_x$  columns and  $2k_y$  rows from  $A$ , given their indexes.

Note that our computation yields, along with  $(x^N, y^N)$ , the quantities  $\bar{\phi}(x^N)$  and  $\underline{\phi}(y^N)$ . Thus, when repeating the computation  $\ell$  times and choosing the best among the resulting  $x$ - and  $y$ -components of the solutions we make the probability of the left-hand side event in (6.60) as large as  $1 - 2^{-\ell}$ . For example, with  $k_x = k_y = 1$ , assuming  $\delta = \epsilon/\|A\|_{1,\infty} \leq 1$ , finding an  $\epsilon$ -solution to (6.59) with reliability  $\geq 1 - \beta$  costs  $O(1) \ln^2(2mn) \ln(1/\beta) \delta^{-2}$  steps of the outlined type, that is,  $O(1)(m+n) \ln^2(2mn) \ln(1/\beta) \delta^{-2}$  a.o. For comparison, when  $\delta$  stays fixed and  $m, n$  are large, the lowest known (so far) cost of finding an  $\epsilon$ -solution to problem (6.58) with uniform fit is  $O(1) \sqrt{\ln(m) \ln(n)} \delta^{-1}$  steps, with the effort per step dominated by the necessity to compute  $O(1)$  matrix-vector multiplications involving  $A$  and  $A^T$  (this cost is achieved by Nesterov's smoothing or with MP; see (6.22)). When  $A$  is a general-type dense  $m \times n$  matrix, the cost of the deterministic computation is  $O(1)mn \sqrt{\ln(m) \ln(n)} \delta^{-1}$ . We see that for fixed relative accuracy  $\delta$  and large  $m, n$ , randomization does accelerate the solution process, the gain growing with  $m, n$ .

---

2. The bound for  $\sigma_x^2$  and  $\sigma_y^2$  is readily given by the following fact (see, e.g., Juditsky and Nemirovski, 2008): when  $\xi_1, \dots, \xi_k \in \mathbf{R}^n$  are independent zero mean random vectors with  $\mathbf{E}\{\|\xi_i\|_\infty^2\} \leq 1$  for all  $i$ , one has  $\mathbf{E}\{\|\frac{1}{k} \sum_{i=1}^k \xi_i\|_\infty^2\} \leq O(1) \min[1, \ln(n)/k]$ ; this inequality remains true when  $\mathbf{R}^n$  is replaced with  $\mathbf{S}^n$ , and  $\|\cdot\|_\infty$  is replaced with the standard matrix norm (largest singular value).

$\ell_2$ -fit. Here (6.58) can be converted into the bilinear saddle-point problem

$$\text{Opt} = \min_{x \in \mathbb{S}_{2n}} \max_{\|y\|_2 \leq 1} [\phi(x, y) := y^T [Ax - b]]. \quad (6.61)$$

In this case we keep  $\|\cdot\|_x = \|\cdot\|_1$ ,  $\omega_x(x) = \text{Ent}(x)$ , and set  $\|\cdot\|_y = \|\cdot\|_2$ ,  $\omega_y(y) = \frac{1}{2}y^T y$ . We specify  $\Pi_u$ ,  $u \in \mathbb{S}_{2n}$ , exactly as in the case of uniform fit, and define  $P_v$ ,  $v \in \mathcal{Y} = \{y \in \mathbb{R}^m : \|y\|_2 \leq 1\}$  as follows:  $\eta_v \sim P_v$  takes values  $\text{sign}(u_i)\|u\|_1 e_i$ ,  $e_i$  being basic orths, with probabilities  $|u_i|/\|u\|_1$ . Note that we are not in the inside case anymore. Setting  $\|A\|_{1,2} = \max_{1 \leq j \leq 2n} \|A_j\|_2$ ,  $A_j$  being the columns of  $A$ , we get

$$\begin{aligned} \Omega_x &= \ln(2n), \Omega_y = \frac{1}{2}, \mathcal{L} = \|A\|_{1,2} \sqrt{2 \ln(2n)}, \\ \Theta &\leq O(1) \left[ \frac{1}{k_x} \|A\|_{1,2}^2 + \frac{\ln^2(2n)}{k_y + \ln(2n)} [\|A\|_{1,2} + \sqrt{m} \|A\|_{1,\infty}]^2 \right]. \end{aligned}$$

Now Proposition 6.6 reads:

**Corollary 6.8.** *For all positive integers  $k_x$ ,  $k_y$ ,  $N$ , one can find a random feasible solution  $x^N$  to (6.58) (where  $p = 2$ ), along with the vector  $Ax^N$ , such that*

$$\begin{aligned} \text{Prob} \left\{ \|Ax^N - b\|_2 \leq \text{Opt} \right. \\ \left. + O(1) \frac{\|A\|_{1,2} \sqrt{\ln(2n)}}{\sqrt{N}} \sqrt{\frac{1}{N} + \frac{1}{k_x \ln(2n)} + \frac{\ln(2n) \Gamma^2(A)}{k_y + \ln(2n)}} \right\} \geq \frac{1}{2}, \quad (6.62) \end{aligned}$$

$$\Gamma(A) = \sqrt{m} \|A\|_{1,\infty} / \|A\|_{1,2}$$

in  $N$  steps, the computational effort per step dominated by the necessity to extract  $2k_x$  columns and  $2k_y$  rows from  $A$ , given their indexes.

Here again, repeating the computation  $\ell$  times and choosing the best among the resulting solutions to (6.58), we make the probability of the left-hand side event in (6.62) as large as  $1 - 2^{-\ell}$ . For instance, with  $k_x = k_y = 1$ , assuming  $\delta := \epsilon / \|A\|_{1,2} \leq 1$ , finding an  $\epsilon$ -solution to (6.58) with reliability  $\geq 1 - \beta$  costs  $O(1) \ln(2n) \ln(1/\beta) \Gamma^2(A) \delta^{-2}$  steps of the outlined type, that is,  $O(1)(m+n) \ln(2n) \ln(1/\beta) \Gamma^2(A) \delta^{-2}$  a.o. Assuming that a precise multiplication of a vector by  $A$  takes  $O(mn)$  a.o., the best known (so far) deterministic counterpart of the above complexity bound is  $O(1)mn \sqrt{\ln(2n)} \delta^{-1}$  a.o. (cf. (6.22)). Now the advantages of randomization when  $\delta$  is fixed and  $m, n$  are large are not as evident as in the case of uniform fit, since the complexity bound for the randomized computation contains an extra factor  $\Gamma^2(A)$  which may be as large as  $O(m)$ . Fortunately, we may “nearly kill”  $\Gamma(A)$  by randomized preprocessing of the form  $[A, b] \mapsto [\bar{A}, \bar{b}] = [UDA, UDb]$ , where  $U$  is a deterministic orthogonal matrix with entries of order  $O(1/\sqrt{m})$ , and  $D$  is a random diagonal matrix with i.i.d. diagonal entries taking values  $\pm 1$  with

equal probabilities. This preprocessing converts (6.58) into an equivalent problem, and it is easily seen that for every  $\beta \ll 1$ , for the transformed matrix  $\bar{A}$  with probability  $\geq 1 - \beta$  it holds that  $\Gamma(\bar{A}) \leq O(1)\sqrt{\ln(mn\beta^{-1})}$ . This implies that, modulo preprocessing's cost, the complexity estimate of the randomized computation reduces to  $O(1)(m+n)\ln(n)\ln(mn/\beta)\ln(1/\beta)\delta^{-2}$ . Choosing  $U$  as a cosine transform or Hadamard matrix, so that the cost of computing  $Uu$  is  $O(m\ln(m))$  a.o., the cost of preprocessing does not exceed  $O(mn\ln(m))$ , which, for small  $\delta$ , is a small fraction of the cost of deterministic computation. Thus, there is a meaningful range of values of  $\delta, m, n$  where randomization is highly profitable. It should be added that in some applications (e.g., in compressed sensing) typical values of  $\Gamma(A)$  are quite moderate, and thus no preprocessing is needed.

## 6.6 Notes and Remarks

1. The research of the second author was partly supported by ONR grant N000140811104, BSF grant 2008302, and NSF grants DMI-0619977 and DMS-0914785.

2. The mirror-prox algorithm was proposed by Nemirovski (2004); its modification able to handle the stochastic case, where the precise values of the monotone operator associated with (6.3) are replaced by unbiased random estimates of these values (cf. Chapter 5, Section 5.5) is developed by Juditsky et al. (2008). The MP combines two basic ideas: (a) averaging of the search trajectory to get approximate solutions (this idea goes back to Bruck (1977) and Nemirovskii and Yudin (1978)) and (b) exploiting *extragradient* steps: instead of the usual gradient-type update  $z \mapsto z_+ = \text{Prox}_z(\gamma F(z))$  used in the saddle-point MP (Section 5.6), the update  $z \mapsto w = \text{Prox}_z(\gamma F(z)) \mapsto z_+ = \text{Prox}_z(\gamma F(w))$  is used. This construction goes back to Korpelevich (1983, 1976), see also Noor (2003) and references therein. Note that a different implementation of the same ideas is provided by Nesterov (2007b) in his dual extrapolation algorithm.

3. The material in Sections 6.4.1 and 6.4.3 is new; this being said, problem settings and complexity results considered in these sections (but not the associated algorithms) are pretty close, although not fully identical, to those covered by the excessive gap technique of Nesterov (2005b). For example, the situation considered in illustration III can be treated equally well via Nesterov's technique, which perhaps is not the case for illustration II. It should be added that splitting like the one in Section 6.4.1, in a slightly more general context of variational inequalities with monotone

operators, was considered by Tseng (2000), although without averaging and thus without any efficiency estimate. These missing elements were added in the recent papers of Monteiro and Svaiter (2010b,a) which in this respect can be viewed as independently developed Euclidean case version of Section 6.4.1. For other schemes of accelerating FOMs via exploiting a problem's structure, see Nesterov (2007a), Beck and Teboulle (2009), Tseng (2008), Goldfarb and Scheinberg (2010), and references therein.

4. The material of Section 6.5.1 originated with Juditsky et al. (2010), where one can find various versions of MPr and (rather encouraging) results of preliminary numerical experiments. Note that the “cheap randomized matrix-vector multiplication” outlined in Section 6.5.1 admits extensions which can be useful when solving semidefinite programs (see Juditsky et al., 2010, Section 2.1.4).

Obviously, the idea of improving the numerical complexity of optimization algorithms by utilizing random subsampling of problem data is not new. For instance, such techniques have been applied to support vector machine classification in Kumar et al. (2008), and to solving certain semidefinite programs in Arora and Kale (2007) and d'Aspremont (2009). Furthermore, as we have already mentioned, both MD and MP admit modifications (see Nemirovski et al., 2009; Juditsky et al., 2008) capable to handle c.-c.s.p. problems (not necessarily bilinear) in the situation where instead of the precise values of the associated monotone operator, unbiased random estimates of these values are used. A common drawback of these modifications is that while we have at our disposal explicit nonasymptotical upper bounds on the *expected* inaccuracy of random approximate solutions  $z^N$  (which, as in the basic MP, are averages of the search points  $w_t$ ) generated by the algorithm, we do *not* know what the actual quality of  $z^N$  is. In the case of a bilinear problem (6.47) and with the randomized estimates of  $F(w_t)$  defined as  $F(\hat{\zeta}_t)$ , we get a new option: to define  $z^N$  as the average of the points  $\hat{\zeta}_t$ . As a result, we do know  $F(z^N)$  and thus can easily assess the quality of  $z^N$  (n.b. remark 6.1). To the best of our knowledge, this option has been realized (implicitly) only once, namely, in the randomized sublinear-time matrix game algorithm of Grigoriadis and Khachiyan (1995) (that ad hoc algorithm is close, although not identical, to MPr as applied to problem (6.59), which is equivalent to a matrix game).

On the other hand, the possibility to assess, in a computationally cheap fashion, the quality of an approximate solution to (6.47) is crucial when solving *parametric* bilinear saddle-point problems. Specifically, many important

applications reduce to problems of the form

$$\max \left\{ \rho : \text{SadVal}(\rho) := \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \phi_\rho(x, y) \leq 0 \right\}, \quad (6.63)$$

where  $\phi_\rho(x, y)$  is a bilinear function affinely depending on  $\rho$ . For example, the  $\ell_1$ -minimization problem as it arises in sparsity-oriented signal processing is  $\text{Opt} = \min_{\xi} \{ \|\xi\|_1 : \|A\xi - b\|_p \leq \delta \}$ , which is nothing but

$$\frac{1}{\text{Opt}} = \max \left\{ \rho : \text{SadVal}(\rho) := \min_{\|x\|_1 \leq 1} \max_{\|y\|_{p/(p-1)} \leq 1} [y^T [Ax - \rho b] - \rho \delta] \leq 0 \right\}.$$

From the complexity viewpoint, the best known (to us) way to process (6.63) is to solve the master problem  $\max\{\rho : \text{SadVal}(\rho) \leq 0\}$  by an appropriate first-order root-finding routine, the (approximate) first-order information on  $\text{SadVal}(\cdot)$  being provided by a first-order saddle-point algorithm. The ability of the MPr algorithm to provide accurate bounds of the value  $\text{SadVal}(\cdot)$  of the inner saddle-point problems makes it the method of choice when solving extremely large parametric saddle-point problems (6.63). For more details on this subject, see Juditsky et al. (2010).

## 6.7 References

- S. Arora and S. Kale. A combinatorial primal-dual approach to semidefinite programs. In: *Proceedings of the 39th Annual ACM Symposium on the Theory of Computations*, pages 227–236, 2007.
- A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- R. Bruck. On weak convergence of an ergodic iteration for the solution of variational inequalities with monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 61(1):15–164, 1977.
- A. d’Aspremont. Subsampling algorithms for semidefinite programming. Technical report, arXiv:0803.1990v5, <http://arxiv.org/abs/0803.1990>, November 2009.
- D. Goldfarb and K. Scheinberg. Fast first order method for separable convex optimization with line search. Technical report, Department of Industrial Engineering and Operations Research, Columbia University, 2010.
- M. D. Grigoriadis and L. G. Khachiyan. A sublinear-time randomized approximation algorithm for matrix games. *Operations Research Letters*, 18(2):53–58, 1995.
- A. Juditsky and A. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. Technical report, HAL: hal-00318071, <http://hal.archives-ouvertes.fr/hal-00318071/>, 2008.
- A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror prox algorithm. Technical report, HAL: hal-00318043,

- <http://hal.archives-ouvertes.fr/hal-00318043/>, 2008.
- A. Juditsky, F. K. Karzan, and A. Nemirovski.  $\ell_1$ -minimization via randomized first order algorithms. Technical report, Optimization Online, [http://www.optimization-online.org/DB\\_FILE/2010/05/2618.pdf](http://www.optimization-online.org/DB_FILE/2010/05/2618.pdf), 2010.
- G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12:747–756, 1976. (in Russian).
- G. M. Korpelevich. Extrapolation gradient methods and relation to modified lagrangeans. *Ekonomika i Matematicheskie Metody*, 19:694–703, 1983. (in Russian).
- K. Kumar, C. Bhattacharya, and R. Hariharan. A randomized algorithm for large scale support vector learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. MIT Press, 2008.
- R. D. C. Monteiro and B. F. Svaiter. Complexity of variants of Tseng’s modified F-B splitting and Korpelevich’s methods for generalized variational inequalities with applications to saddle point and convex optimization problems. Technical report, Optimization Online, [http://www.optimization-online.org/DB\\_HTML/2010/07/2675.html](http://www.optimization-online.org/DB_HTML/2010/07/2675.html), 2010a.
- R. D. C. Monteiro and B. F. Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20:2755–2787, 2010b.
- 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 Journal on Optimization*, 15:229–251, 2004.
- A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
- A. Nemirovski, S. Onn, and U. Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35:52–78, 2010.
- 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.
- Y. Nesterov. A method for solving a convex programming problem with rate of convergence  $o(1/k^2)$ . *Soviet Math. Doklady*, 27(2):372–376, 1983.
- Y. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming, Series A*, 103:127–152, 2005a.
- Y. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, 16(1):235–239, 2005b.
- Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 2007/76, Center for Operations Research and Econometrics, Catholic University of Louvain, [http://www.uclouvain.be/cps/ucl/doc/core/documents/coredp2007\\_76.pdf](http://www.uclouvain.be/cps/ucl/doc/core/documents/coredp2007_76.pdf), 2007a.
- Y. Nesterov. Dual extrapolation and its application for solving variational inequalities and related problems. *Mathematical Programming, Series A*, 109(2–3): 319–344, 2007b.
- M. A. Noor. New extragradient-type methods for general variational inequalities.

- Journal of Mathematical Analysis and Applications*, 277:379–394, 2003.
- P. Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.
- P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, <http://www.math.washington.edu/~tseng/papers/apgm.pdf>, 2008.