2 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 1, the methods in question utilize the problem structure in order to convert the original nonsmooth minimization problem into a saddle point problem with smooth convex-concave cost function. This reformulation allows to accelerate significantly the solution process. As in Chapter 1, our emphasis is on methods which, under favorable circumstances, exhibit (nearly) dimension-independent convergence rate. Along with investigating the general "well-structured" situation, we outline possibilities to further accelerate First Order methods by randomization.

2.1 Introduction

The major drawback of the First Order methods (FOMs) considered in chapter 1 is their slow convergence: as the number of steps t grows, the inaccuracy decreases as slowly as $O(1/\sqrt{t})$. As it was explained in Section 1.1, 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 to be completely black-box-oriented. For example, utilizing 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 for 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 problem's structure when looking for medium-accuracy solutions was proposed by Nesterov (2005a); his main observation (whatever 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 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, as compared to the case when problem's structure is ignored, while staying all the time within the scope of computationally cheap FOMs.

In the seminal paper of Nesterov (2005a) the saddle point reformulation of the (convex) problem of interest $\min_{x\in X} f(x)$ is used to construct a computationally cheap smooth convex approximation $\widetilde{f}$ 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 $\widetilde{f}$ deteriorate as $\widetilde{f}$ 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 1.1 (see item (c)), this is the best we can get in the large-scale case when solving as simply-looking problems as $\min_{\|x\|_2\le R} \|Ax-b\|_2$. In what follows, we use as a "working horse" the Mirror Prox (MP) saddle point algorithm from (Nemirovski, 2004) converging at the same rate $O(1/t)$ as Nesterov's smoothing, but 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 simply-looking modification of the saddle point Mirror Descent algorithm from Section 1.6). Another reason is that as compared to smoothing, MP is better suited for accelerating by randomization to be considered in Section 2.5.

The main body of this chapter is organized as follows. In Section 2.2, 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 "working horse" — the Mirror Prox

algorithm with the rate of convergence $O(1/t)$ for solving smooth convex-concave saddle point problems — is presented in Section 2.3. In Section 2.4 we consider two special cases where the MP algorithm can be further accelerated. Another "acceleration option" is considered in Section 2.5, 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 ℓ_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.

Terminology and notation we use in this chapter follow those introduced in Sections 1.2.2, 1.6.1, 1.7 of Chapter 1.

2.2 Saddle Point Reformulations of Convex Minimization Problems

2.2.1 Saddle point representations of convex functions

Let $\mathcal{X}\subset E$ be a nonempty closed convex set in Euclidean space E_x, let $f(x):\mathcal{X}\to\mathbb{R}$ 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

$$\forall x\in\mathcal{X}: f(x)=\overline{\phi}(x):=\sup_{y\in\mathcal{Y}}\phi(x,y). \tag{2.1}$$

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

$$\min_{x\in\mathcal{X}} f(x) \tag{2.2}$$

of minimizing f over $\mathcal{X}$ (cf. (1.2)) to the convex-concave saddle point (c.-c.s.p.) problem

$$\mathrm{SadVal}=\inf_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}}\phi(x,y), \tag{2.3}$$

(cf. (1.31)). Namely, assuming that ϕ has a saddle point on $\mathcal{X}\times\mathcal{Y}$, (2.2) is solvable, and invoking (1.32), we get for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$:

$$\begin{array}{rl} f(x)-\min_{\mathcal{X}} f & =\overline{\phi}(x)-\mathrm{Opt}(P)=\overline{\phi}(x)-\mathrm{SadVal}\\ & \leq \overline{\phi}(x)-\underline{\phi}(y)=\epsilon_{\mathrm{sad}}(x,y).\end{array} \tag{2.4}$$

That is, the x-component of an ϵ-solution to (2.3) (i.e., a point $(x,y)\in\mathcal{X}\times\mathcal{Y}$ with $\epsilon_{\rm sad}(x,y)\le\epsilon$) is an ϵ-solution to (2.2): $f(x)-\min_{\mathcal{X}} f\le\epsilon$.

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 ϕ and simple $\mathcal{Y}$; as a result, the saddle point reformulation (2.3) of the problem (2.2) associated with f can be much better suited for processing by FOMs than the problem (2.2) "as it is." Let us consider some examples (where S_n, S_n^+, Σ_ν, Σ_ν^+ are the standard "flat" and full-dimensional simplexes/spectahedrons, see Section 1.7.1):

1. $f(x):=\max_{1\le\ell\le L} f_\ell(x)=\max_{y\in \mathrm{S}_L}\Big[\phi(x,y):=\sum_{\ell=1}^L y_\ell f_\ell(x)\Big]$; when all f_ℓ are smooth, so is ϕ;
2. $f(x):=\|Ax-b\|_p=\max_{\|y\|_q\le 1}\big[\phi(x,y):=y^T(Ax-b)\big]$, $q=\frac{p}{p-1}$. 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 $f(x):=\|[Ax-b]_+\|_p=\max_{\|y\|_q\le 1,y\ge 0}\phi(x,y)$ and $f(x):=\min_s\|[Ax-b-sc]_+\|_p=\max_{\|y\|_q\le 1,y\ge 0,c^Ty=0}\phi(x,y)$. In particular,
   (a) Let $\mathcal{A}(\cdot)$ be an affine mapping. The problem

$$\mathrm{Opt}=\min_{\xi\in\Xi}\left[f(\xi):=\|\mathcal{A}(\xi)\|_p\right] \tag{2.5}$$

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

$$\min_{x\in\mathrm{S}_{2n}^+}\max_{\|y\|_q\le 1} y^T\mathcal{A}(Jx)\quad[J=[I,-I],\ q=\frac{p}{p-1}] \tag{2.6}$$

   on the product of the standard simplex and the unit $\|\cdot\|_q$-ball. When $\Xi=\{\xi\in\mathbb{R}^{m\times n}:\|\xi\|_{\rm n}\le 1\}$, $\|\cdot\|_{\rm n}$ being the nuclear norm (cf. nuclear norm minimization), representing Ξ as the image of the spectahedron Σ_{m+n}^+ under the linear mapping $x=\left[\begin{array}{cc}u & v\\ v^T & w\end{array}\right]\mapsto\mathcal{J}x:=2v$, (2.5) reduces to the bilinear saddle point problem

$$\min_{x\in\Sigma_{m+n}^+}\max_{\|y\|_q\le 1} y^T\mathcal{A}(\mathcal{J}x); \tag{2.7}$$

   (b) the SVM-type problem

$$\min_{\substack{w\in\mathbb{R}^n,\|w\|\le R,\\ s\in\mathbb{R}}}\Big\|\big[\mathbf{1}-\mathrm{Diag}\{\eta\}(M^Tw+s\mathbf{1})\big]_+\Big\|_p,\quad \mathbf{1}=[1;...;1],$$

reduces to the bilinear saddle point problem

$$\min_{\|x\|\le 1}\max_{\substack{\|y\|_q\le 1,\\ y\ge 0,\ \eta^T y=0}}\left[\phi(x,y):=\sum_j y_j - y^T\,\mathrm{Diag}\{\eta\}RM^Tx\right], \tag{2.8}$$

where $x = w/R$.

3. Let $\mathcal{A}(x) = A_0 + \sum_{i=1}^n x_iA_i$ with $A_0, ..., A_n \in \mathbf{S}^\nu$, and let $S_k(A)$ be the sum of 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, ϕ is "as simple as it could be" — it is just bilinear. The number of examples of this type can be easily increased due to the following observation: the family of convex functions f admitting explicit bilinear saddle point representations (b.s.p.r.'s)

$$f(x) = \max_{y\in\mathcal{Y}}[\langle y, \mathcal{A}x + a\rangle + \langle b, x\rangle + c] \tag{2.9}$$

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, specifically, (a) affine substitution of the argument $x \leftarrow P\xi + p$, (b) multiplication by nonnegative reals, (c) summation, (d) direct summation $\{f_i(x^i)\}_{i=1}^k \mapsto f(x^1, ..., x^k) = \sum_{i=1}^k f_i(x^i)$, and (e) taking maximum. Here (a), (b) are evident, and (c) is nearly so: if

$$f_i(x) = \max_{y^i\in\mathcal{Y}_i}\left[\langle y^i, \mathcal{A}_ix + a^i\rangle + \langle b^i, x\rangle + c_i\right],\ i = 1, ..., k, \tag{2.10}$$

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

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

(d) is an immediate consequence of (a) and (c). To verify (e), let f_i be given by (2.10), let E_i be the embedding space of $\mathcal{Y}_i$, and let $\mathcal{U}_i = \{(u^i, \lambda_i) = (\lambda_iy^i, \lambda_i) : y^i \in \mathcal{Y}_i, \lambda_i \ge 0\} \subset E_i^+ = E_i \times \mathbb{R}$; since $\mathcal{Y}_i$ are convex and compact, the sets $\mathcal{U}_i$ are closed convex cones. Now let

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

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

it is bounded as well. We have

$$\begin{aligned}\max_{1\le i\le k} f_i(x) &= \max_{\lambda\ge 0:\sum_i\lambda_i=1}\sum_{i=1}^k \lambda_i f_i(x) = \max_{\lambda,y^1,...,y^k}\Big\{\textstyle\sum_{i=1}^k\big[\langle\overbrace{\lambda_i y^i}^{u^i},\mathcal{A}_i x+a^i\rangle \\ &\qquad +\langle\lambda_i b^i,x\rangle+\lambda_i c_i\big] : \lambda\ge 0,\textstyle\sum_i\lambda_i=1, y^i\in\mathcal{Y}_i, 1\le i\le k\Big\} \\ &= \max_{u=\{(u^i,\lambda_i):1\le i\le k\}\in\mathcal{U}}\big[\textstyle\sum_{i=1}^k[\langle u^i,\mathcal{A}_i x+a^i\rangle+\langle\lambda_i b^i,x\rangle+\lambda_i c_i]\big],\end{aligned}$$

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

2.3 Mirror Prox Algorithm

We are about to present the basic MP algorithm for the problem (2.3).

2.3.1 Assumptions and setup

Here we assume that

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

B. The convex-concave function $\phi(x,y):\mathcal{Z}=\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ possesses 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 (2.3), see Section 1.6.1.

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 1.2.2). For $z\in\mathcal{Z}^o$, $w\in\mathcal{Z}$ let (cf. the definition (1.4))

$$V_z(w)=\omega(w)-\omega(z)-\langle\omega'(z),w-z\rangle, \tag{2.11}$$

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

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

and set

$$\Omega=\max_{w\in\mathcal{Z}}V_{z_c}(w)\le\max_{\mathcal{Z}}\omega(\cdot)-\min_{\mathcal{Z}}\omega(\cdot) \tag{2.12}$$

(cf. Section 1.2.2). 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 $\|\cdot\|_*$:

$$\forall(z,z'\in\mathcal{Z}):\|F(z)-F(z')\|_*\le L\|z-z'\|. \tag{2.13}$$

2.3.2 The algorithm

The MP algorithm is given by the recurrence

$$\begin{array}{ll}(a): & z_1=z_c,\\ (b): & w_\tau=\mathrm{Prox}_{z_\tau}(\gamma_\tau F(z_\tau)),\ z_{\tau+1}=\mathrm{Prox}_{z_\tau}(\gamma_\tau F(w_\tau)),\\ (c): & z^\tau=[\sum_{s=1}^\tau\gamma_s]^{-1}\sum_{s=1}^\tau\gamma_s w_s,\end{array} \tag{2.14}$$

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

$$\delta_\tau=\gamma_\tau\langle F(w_\tau),w_\tau-z_{\tau+1}\rangle-V_{z_\tau}(z_{\tau+1}) \tag{2.15}$$

(cf. (1.4)). The convergence properties of the algorithm are given by the following

Proposition 2.1. *Under Assumptions* **A**, **B**

(i) *For every $t\ge1$ it holds (for notation, see* (2.12), (2.15)*)*

$$\epsilon_{\mathrm{sad}}(z^t)\le\left[\sum\nolimits_{\tau=1}^t\gamma_\tau\right]^{-1}\left[\Omega+\sum\nolimits_{\tau=1}^t\delta_\tau\right]. \tag{2.16}$$

(ii) *If the stepsizes satisfy the condition $\gamma_\tau\ge L^{-1}$, $\delta_\tau\le0$ for all τ (which for sure is so when $\gamma_\tau\equiv L^{-1}$), we have*

$$\forall t\ge1:\epsilon_{\mathrm{sad}}(z^t)\le\Omega\left[\sum\nolimits_{\tau=1}^t\gamma_\tau\right]^{-1}\le\Omega L/t. \tag{2.17}$$

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

Lemma 2.2. *Given $z\in\mathcal{Z}^o$, $\xi,\eta\in E$, let $w=\mathrm{Prox}_z(\xi)$ and $z_+=\mathrm{Prox}_z(\eta)$. Then for all $u\in\mathcal{Z}$ it holds*

$$\begin{array}{ll}\langle\eta,w-u\rangle\le V_z(u)-V_{z_+}(u)+\langle\eta,w-z_+\rangle-V_z(z_+) & (a)\\ \le V_z(u)-V_{z_+}(u)+\langle\eta-\xi,w-z_+\rangle-V_z(w)-V_w(z_+) & (b)\\ \le V_z(u)-V_{z_+}(u)+\left[\frac12\|\eta-\xi\|_*\|w-z_+\|-\frac12\|z-w\|^2-\frac12\|z_+-w\|^2\right] & (c)\\ \le V_z(u)-V_{z_+}(u)+\frac12[\|\eta-\xi\|_*^2-\|w-z\|^2] & (d)\end{array} \tag{2.18}$$

Proof. By definition of $z_+=\mathrm{Prox}_z(\eta)$ we have $\langle\eta-\omega'(z)+\omega'(z_+),u-z_+\rangle\ge0$; we obtain (2.18.a) by rearranging terms and taking into account the definition of $V_v(u)$, (cf. the derivation of (1.12)). By definition of $w=\mathrm{Prox}_z(\xi)$ we have $\langle\xi-\omega'(z)+\omega'(w),z_+-w\rangle\ge0$, whence $\langle\eta,w-z_+\rangle\le$

$\langle \eta - \xi, w - z_+\rangle + \langle \omega'(w) - \omega'(z), z_+ - w\rangle$; 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 ω implying that $V_v(u) \geq \frac{1}{2}\|u - v\|^2$, and (d) is an immediate consequence of (c). □

2^0. Applying Lemma 2.2 to $z = z_\tau$, $\xi = \gamma_\tau F(z_\tau)$ (which results in $w = w_\tau$) and $\eta = \gamma_\tau F(w_\tau)$ (which results in $z_+ = z_{\tau+1}$), we obtain due to (2.18.d):

$$\begin{array}{ll} (a) & \gamma_\tau \langle F(w_\tau), w_\tau - u\rangle \leq V_{z_\tau}(u) - V_{z_{\tau+1}}(u) + \delta_\tau \; \forall u \in \mathcal{Z}, \\ (b) & \delta_\tau \leq \frac{1}{2}\left[\gamma_\tau^2 \|F(w_\tau) - F(z_\tau)\|_*^2 - \|w_\tau - z_\tau\|^2\right] \end{array} \tag{2.19}$$

Summing up (2.19.a) over $\tau = 1, ..., t$, taking into account that $V_{z_1}(u) = V_{z_c}(u) \leq \Omega$ by (2.12) 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

$$\forall u \in \mathcal{Z} : \sum_{\tau=1}^t \lambda_\tau \langle F(w_\tau), w_\tau - u\rangle \leq A := \frac{\Omega + \sum_{\tau=1}^t \delta_\tau}{\sum_{\tau=1}^t \gamma_\tau}. \tag{2.20}$$

On the other hand, setting $w_\tau = (x_\tau, y_\tau)$, $z^t = (x^t, y^t)$, $u = (x, y)$ and using (1.37) 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 (2.20) results in $\phi(x^t, y) - \phi(x, y^t) \leq A$ for all $(x, y) \in \mathcal{Z}$. Taking supremum in $(x, y) \in \mathcal{Z}$, we arrive at (2.16); (i) is proved. To prove (ii), note that with $\gamma_t \leq L^{-1}$, (2.19.b) implies that $\delta_\tau \leq 0$, see (2.13). □

2.3.3 Setting up the MP algorithm

Let us restrict ourselves with the *favourable geometry* case defined completely similar to Section 1.7.2, 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 rint $\mathcal{Z}^+$. We assume that the representation $\mathcal{Z}^+ = \mathcal{Z}_1 \times ... \times \mathcal{Z}_K$ 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 $E = E_1 \times ... \times E_K$ of $\mathcal{Z} \subset \mathcal{Z}^+$ with the norm $\|\cdot\|$ and a d.-g.f. $\omega(\cdot)$ according to (1.42), (1.43) (where, for notational consistency, we should replace x^ℓ with z^ℓ 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 α_ℓ in (1.42), (1.43). To this end assume that we have at our disposal upper bounds $L_{\mu\nu} = L_{\nu\mu}$

on the "partial Lipschitz constants $L^*_{\mu\nu}$" of the (Lipschitz continuous by Assumption **B**) vector field $F(z=(x,y)) = (\nabla_x\phi(x,y), -\nabla_y\phi(x,y))$. The latter quantities are defined as follows: the representation $E = E_1 \times ... \times E_K$ induces the representation $F(z=(z^1,...,z^K)) = (F_1(z),...,F_K(z))$, and we set

$$L^*_{\mu\nu} = \sup\left\{\frac{\|F_\mu(z^1,...,z^K) - F_\mu(z^1,...,z^{\nu-1},z^\nu_+,z^{\nu+2},...,z^K)\|_{(\mu),*}}{\|z^\nu - z^\nu_+\|_{(\nu)}} : z^\ell \in \mathcal{Z}_\ell, z^\nu_+ \in \mathcal{Z}_\nu, z^\nu \neq z^\nu_+\right\}.$$

Let Ω_ℓ be defined by (1.46) 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

$$\Omega \le 1 \ \text{ and } \ L \le \mathcal{L} := \sum\nolimits_{\mu,\nu} L_{\mu\nu}\sqrt{\Omega_\mu\Omega_\nu},$$

, so that the bound (2.17) reads:

$$\epsilon_{\rm sad}(z^t) \le \mathcal{L}/t, \ \mathcal{L} = \sum\nolimits_{\mu,\nu} L_{\mu\nu}\sqrt{\Omega_\mu\Omega_\nu}. \tag{2.21}$$

As far as complexity of a step and dependence of the efficiency estimate on problem's dimension are concerned, the present situation is completely similar to that of MD, studied in Section 1.7. In particular, all our considerations in the discussion at the end of Section 1.7.2 remain valid here.

2.3.3.1 Illustration I

As simple and instructive illustrations, consider problems (2.8), (2.5).
1. Consider problem (2.8), 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 η is a ± 1 vector which has both positive and negative entries. When $p = 2$, (2.8) is a bilinear saddle point problem on the product of the unit $\|\cdot\|_r$-ball and a simple part of $\|\cdot\|_2$-ball. Combining (2.21) with what was said in Section 1.7.2, we arrive at the efficiency estimate

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

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: (2.8) is now a bilinear saddle point problem on the product of the unit $\|\cdot\|_r$-ball and a simple subset of the unit box $\{y : \|y\|_\infty \le 1\}$, or, which is the same, a simple subset of the centered at the origin Euclidean ball of the radius $\rho = \sqrt{\dim\eta}$. 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 $\rho R\,\mathrm{Diag}\{\eta\}M^T$. As a result, we arrive at the dimension-dependent efficiency estimate

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

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

2. Now consider problem (2.5), and let $p\in\{2,\infty\}$.

2.1. Let us start with the case of $\Xi=\{\xi\in\mathbb{R}^n:\|\xi\|_1\le 1\}$, so that $\mathcal{A}(Jx)=A_0+Ax$, where A is an $m\times 2n$ matrix. Here (2.6) 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 (2.21) with derivations in Section 1.7.2, the efficiency estimate of MP reads

$$\epsilon_{\rm sad}(x^t,y^t)\le O(1)\sqrt{\ln(n)}(\ln(m))^{\frac12-\frac1p}\left[\max_{1\le j\le\dim x}\|A_j\|_p\right]t^{-1},\quad(2.22)$$

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 $\Xi=\{(\xi^1,...,\xi^k)\in\mathbb{R}^{d_1}\times...\times\mathbb{R}^{d_k}:\sum_j\|\xi^j\|_2\le 1\}$, so that problem (2.5) is of the form arising in "block Lasso" ($p=2$) or "block Dantzig selector" ($p=\infty$). Given $d_i=\dim\xi^i$, consider the block-diagonal structure $\nu=(d_1+1,...,d_k+1)$, and let $\mathcal{X}$ be the part of the corresponding spectahedron Σ^+_ν comprised of matrices from this set which have "arrow" diagonal blocks $\mathrm{Arrow}(\tau_i,\xi^i):=\left[\begin{array}{c|c}\tau_i & [\xi^i]^T\\ \hline \xi^i & \tau_i I_{d_i}\end{array}\right]$. Note that Ξ is nothing but the image of $\mathcal{X}$ under the linear mapping $x\mapsto\mathcal{J}x$ which maps a matrix $x=\mathrm{Diag}\left\{\left[\begin{array}{c|c}\tau_i & [\xi^i]^T\\ \hline \xi^i & T_i\end{array}\right],1\le i\le k\right\}\in\mathbf{S}^\nu$ ($\tau_i\in\mathbb{R}$) into the collection $(\xi^1,...,\xi^k)$. Thus, denoting by A the matrix of the homogeneous part of the affine mapping $\mathcal{A}(\cdot)$, problem (2.5) is equivalent to $\mathrm{Opt}=\min_{\mathcal{X}}\|A\mathcal{J}x-b\|_p$, and thus is equivalent to the bilinear saddle point problem

$$\mathrm{Opt}=\min_{x\in\mathcal{X}}\max_{\|y\|_q\le 1}y^T[\mathcal{B}x-b],\ \mathcal{B}=A\mathcal{J}.$$

Equipping the embedding space $E_x=\mathbf{S}^\nu$ of $\mathcal{X}$ with the trace norm $|\cdot|_1$ and $\mathcal{X}$ — with the inherited from Σ^+_ν d.-g.f. $\omega_x(x)=2\mathrm{Ent}(\lambda(x))$ (see item 3c in Section 1.7.1) and applying the results of Section 2.3.3, the efficiency estimate of MP reads

$$\epsilon_{\rm sad}(z^t)\le O(1)\left(\ln\left(\sum\nolimits_{i=1}^k(d_i+1)\right)\right)^{\frac12}(\ln(m))^{\frac12-\frac1p}\pi(\mathcal{B})t^{-1},\quad(2.23)$$

where $\pi(\mathcal{B})$ is the norm of the linear mapping $x \mapsto \mathcal{B}x$ induced by the trace norm in the argument space and the norm $\|\cdot\|_p$ in the image space.[1]

The best part of the story is that the prox-mapping is easy to compute in this setup. The only non-evident 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, of the form $g(x) = \frac{1}{2}\omega_x(x) + \langle b, x\rangle$. Here is the verification: the eigenvalues of the matrix $\mathrm{Arrow}(\tau, \xi)$, $\dim \xi = d$, are $\tau + \|\xi\|_2$, $\tau - \|\xi\|_2$ and τ with multiplicity $d-1$. Thus, for $x = \mathrm{Diag}\{\mathrm{Arrow}(\tau_1, \xi_1), ..., \mathrm{Arrow}(\tau_k, \xi^k)\} \in \mathcal{X}$, we have

$$g(x) = \sum_{i=1}^{k} \big[(\tau_i + \|\xi^i\|_2)\ln(\tau_i + \|\xi^i\|_2) + (\tau_i - \|\xi^i\|_2)\ln(\tau_i - \|\xi^i\|_2) \\ +(d_i - 1)\tau_i \ln(\tau_i) + \alpha_i\tau_i - \beta_i^T\xi^i\big].$$

Note that at the minimizer of this function over $\mathcal{X}$, the vectors ξ^i are nonnegative multiples of β_i, and finding the minimizer reduces to specifying τ_i and $\sigma_i = \|\xi^i\|_2$. The latter quantities form the optimal solution to the simple "nearly separable" convex program

$$\min_{\tau_i,\sigma_i} \textstyle\sum_{i=1}^k \big[(\tau_i + \sigma_i)\ln(\tau_i + \sigma_i) + (\tau_i - \sigma_i)\ln(\tau_i - \sigma_i) + (d_i - 1)\tau_i \ln \tau_i \\ +\alpha_i\tau_i - \|\beta_i\|_2\sigma_i : 0 \le \sigma_i \le \tau_i, \textstyle\sum_i \tau_i \le 1\big]$$

The above problem clearly can be solved within machine accuracy in $O(\sum_i d_i)$ a.o. As a result, the arithmetic cost of a step of MP in our situation is, for all practical purposes, dominated by $O(1)$ computations of matrix-vector products involving A and A^T.

2.3. Finally, consider the case when Ξ is the unit nuclear norm ball, so that $\mathcal{A}(\mathcal{J}x) = a_0 + [\mathrm{Tr}(A_1x); ...; \mathrm{Tr}(A_kx)]$ with $A_i \in \mathbf{S}^{m+n}$, and (2.7) is a bilinear saddle point problem on the direct product of the spectahedron Σ^+_{m+n} and the unit $\|\cdot\|_q$ ball in $\mathbb{R}^k$. Applying the results of Section 2.3.3, the efficiency estimate of MP reads

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

The complexity of a step is dominated by $O(1)$ computations of the values of $\mathcal{A}$ and matrices of the form $\sum_{i=1}^k y_iA_i$, 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 reformu-

1. It is immediately seen that the norm of the mapping $x \mapsto \mathcal{J}x$ induced by the trace norm in the argument space and the norm $\sum_i \|\xi^i\|_2$ in the image space is equal to 1, so that $\pi(\mathcal{B})$ is at most the norm of the mapping $\xi \mapsto A\xi$ induced by the norm $\sum_i \|\xi^i\|_2$ in the argument space and the norm $\|\cdot\|_p$ in the image space.

lation of (2.5) induces straightforwardly a feasible solution ξ^t to the problem of interest (2.5) such that $f(\xi^t) - \mathrm{Opt} \le \epsilon_{\mathrm{sad}}(x^t, y^t)$.

2.4 Accelerating Mirror Prox algorithm

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

2.4.1 Splitting

2.4.1.1 Situation and assumptions

Consider the c.-c.s.p. problem (2.3) 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 compatible with the respective norms d.-g.f.'s $\omega_x(\cdot)$ for $\mathcal{X}$ and $\omega_y(\cdot)$ for $\mathcal{Y}$.

We already know that if the convex-concave cost function ϕ is smooth (i.e., possesses 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, ϕ can be represented as a sum of a "simple" and smooth parts. Specifically, let us assume that

C.1. The monotone operator Φ associated with (2.3) (see Section 1.6.1) admits "splitting:" we can point out a *Lipschitz continuous on* $\mathcal{Z}$ vector field $G(z) = (G_x(z), G_y(z)) : \mathcal{Z} \to E = E_x \times E_y$ and a point-to-set monotone operator $\mathcal{H}$ with the same domain as Φ such that the sets $\mathcal{H}(z)$, $z \in Dom\mathcal{H}$, are convex and nonempty, the graph of $\mathcal{H}$ (the set $\{(z,h) : z \in \mathrm{Dom}\ \mathcal{H}, h \in \mathcal{H}(z)\}$) is closed, and

$$\forall z \in \mathrm{Dom}\ \mathcal{H} : \mathcal{H}(z) + G(z) \subset \Phi(z). \tag{2.24}$$

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 $\widehat{z} \in \mathcal{Z}$ satisfying

$$\forall (z \in \mathrm{rint}\ \mathcal{Z},\ F \in \Psi(z)) : \langle F, z - \widehat{z}\rangle \ge 0. \tag{2.25}$$

It is easily seen that in the case of **C.1** (2.25) has a unique solution $\widehat{z} = (\widehat{x}, \widehat{y})$ which belongs to $\mathrm{Dom}\ \Phi \cap \mathcal{Z}^o$ and in fact is a strong solution: there exists $\zeta \in \mathcal{H}(\widehat{z})$ such that

$$\forall z \in \mathcal{Z} : \langle \alpha\zeta + [\alpha_z\omega_x'(\widehat{x}) + e; \alpha_y\omega_y'(\widehat{y}) + f], z - \widehat{z}\rangle \ge 0. \tag{2.26}$$

We assume that when solving (2.25), we get both $\widehat{z}$ and ζ.

We intend to demonstrate that under Assumptions **C.1** and **C.2** we can solve (2.3) "as if" there were no $\mathcal{H}$-component at all.

2.4.2 Algorithm MPa

2.4.2.1 Preliminaries

Recall that the mapping $G(x,y) = (G_x(x,y), G_y(x,y)) : \mathcal{Z} \to E$ 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

$$\forall (z=(x,y)\in\mathcal{Z}, z'=(x',y')\in\mathcal{Z}):\\ \begin{cases} \|G_x(x',y)-G_x(x,y)\|_{x,*} \le L_{xx}\|x'-x\|_x, \\ \|G_y(x,y')-G_y(x,y)\|_{y,*} \le L_{yy}\|y'-y\|_y \\ \|G_x(x,y')-G_x(x,y)\|_{x,*} \le L_{xy}\|y'-y\|_y, \\ \|G_y(x',y)-G_y(x,y)\|_{y,*} \le L_{xy}\|x'-x\|_x \end{cases} \tag{2.27}$$

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

$$\begin{array}{l} \Omega_x = \max_{\mathcal{X}}\omega_x(\cdot) - \min_{\mathcal{X}}\omega_x(\cdot),\ \Omega_y = \max_{\mathcal{Y}}\omega_y(\cdot) - \min_{\mathcal{Y}}\omega_y(\cdot) \\ \mathcal{L} = L_{xx}\Omega_x + L_{xy}\Omega_y + 2L_{xy}\sqrt{\Omega_x\Omega_y}, \\ \alpha = [L_{xx}\Omega_x + L_{xy}\sqrt{\Omega_x\Omega_y}]/\mathcal{L},\ \beta = [L_{yy}\Omega_y + L_{xy}\sqrt{\Omega_x\Omega_y}]/\mathcal{L}, \\ \omega(x,y) = \frac{\alpha}{\Omega_x}\omega_x(x) + \frac{\beta}{\Omega_y}\omega_y(y) : \mathcal{Z}\to\mathbb{R}, \\ \|(x,y)\| = \sqrt{\frac{\alpha}{\Omega_x}\|x\|_x^2 + \frac{\beta}{\Omega_y}\|y\|_y^2} \end{array} \tag{2.28}$$

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 2.3.3). Observe that $\omega(\cdot)$ is a d.-g.f. on $\mathcal{Z}$ compatible with the norm $\|\cdot\|$. It is easily seen that $\Omega := 1 \ge \max_{z\in\mathcal{Z}}\omega(z) - \min_{z\in\mathcal{Z}}\omega(z)$ and

$$\forall (z,z'\in\mathcal{Z}):\ \|G(z)-G(z')\|_* \le \mathcal{L}\|z-z'\|. \tag{2.29}$$

2.4.2.2 Algorithm MPa

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

1. *Initialization:* Set $z_1 = \mathrm{argmin}_{\mathcal{Z}}\,\omega(\cdot)$.

2. *Step* $\tau = 1,2,...$: Given $z_\tau \in \mathcal{Z}^o$ and a stepsize $\gamma_\tau > 0$, we find w_τ satisfying

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

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

$$\forall(u \in \mathcal{Z}) : \langle \gamma_\tau(\zeta_\tau + G(z_\tau)) + \omega'(w_\tau) - \omega'(z_\tau), u - w_\tau \rangle \geq 0; \tag{2.30}$$

by assumption **C.2**, computation of ω_τ and ζ_τ is easy. Next, we compute

$$\begin{array}{rcl} z_{\tau+1} & = & \mathrm{Prox}_{z_\tau}(\gamma_\tau(\zeta_\tau + G(w_\tau))) \\ & := & \mathrm{argmin}_{z\in\mathcal{Z}} \left[\langle \gamma_\tau(\zeta_\tau + G(w_\tau)) + V_{z_\tau}(z)\right], \end{array} \tag{2.31}$$

where $V_\cdot(\cdot)$ is defined in (2.11). We set

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

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

Proposition 2.3. *Under Assumptions* **C.1** *and* **C.2**, *algorithm MPa ensures that*

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

$$\epsilon_{\mathrm{sad}}(z^t) \leq \left[\sum\nolimits_{\tau=1}^{t} \gamma_\tau\right]^{-1} \left[1 + \sum\nolimits_{\tau=1}^{t} \delta_\tau\right]. \tag{2.32}$$

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

$$\forall t \geq 1 : \epsilon_{\mathrm{sad}}(z^t) \leq \left[\sum\nolimits_{\tau=1}^{t} \gamma_\tau\right]^{-1} \leq \mathcal{L}/t. \tag{2.33}$$

Proof. Relation (2.30) expresses exactly the fact that $w_\tau = \mathrm{Prox}_{z_\tau}(\gamma_\tau(\zeta_\tau + G(z_\tau)))$. With this in mind, Lemma 2.2 implies that

$$\begin{array}{ll} (a) & \gamma_\tau \langle \zeta_\tau + G(w_\tau), w_\tau - u \rangle \leq V_{z_\tau}(u) - V_{z_{\tau+1}}(u) + \delta_\tau \ \forall u \in \mathcal{Z}, \\ (b) & \delta_\tau \leq \frac{1}{2}\left[\gamma_\tau^2 \|G(w_\tau) - G(z_\tau)\|_*^2 - \|w_\tau - z_\tau\|^2\right], \end{array} \tag{2.34}$$

cf. (2.19). It remains to repeat word by word the reasoning in items 2^0-3^0 of the proof of Proposition 2.1, keeping in mind (2.29) and the fact that, by the origin of ζ_τ and in view of (2.24), we have $\zeta_\tau + G(w_\tau) \in \Phi(w_\tau)$. □

2.4.2.3 Illustration II

Consider a Dantzig selector type problem

$$\text{Opt} = \min_{\|x\|_1 \le 1} \|A^T(Ax - b)\|_\infty \quad [A : m \times n, m \le n] \tag{2.35}$$

(cf. (2.5)) along with its saddle point reformulation

$$\text{Opt} = \min_{\|x\|_1 \le 1} \max_{\|y\|_1 \le 1} y^T[Bx - c], \; B = A^T A, \; c = A^T b. \tag{2.36}$$

As we have already mentioned, the efficiency estimate for the basic MP as applied to this problem is $\epsilon_{\text{sad}}(z^t) \le O(1)\sqrt{\ln(n)}\|B\|_{1,\infty} t^{-1}$, 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* $\mu(A) = \max_{i \ne j} |A_i^T A_j| / A_i^T A_i$ is $\ll 1$. In other words, the diagonal entries in B are of order of 1, and the magnitudes of off-diagonal entries do not exceed $\mu \ll 1$. For example, for a typical randomly selected A, μ is as small as $O(\sqrt{\ln(n)/m})$. Now, the monotone operator associated with (2.36) 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 (2.36) are unit ℓ_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 Section 1.7.1, 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 $\epsilon_{\text{sad}}(z^t) \le O(1)\ln(n)\|\widehat{B}\|_{1,\infty} t^{-1}$, which is much better than the estimate $\epsilon_{\text{sad}}(z^t) \le O(1)\ln(n)\|B\|_{1,\infty} t^{-1}$ for the plain MP (recall that we are in the case of $\mu := \|\widehat{B}\|_{1,\infty} \ll \|B\|_{1,\infty} = O(1)$).
To see that we indeed are in the case of **C.2**, note that in our situation finding a solution $\widehat{z}$ to (2.25) reduces to solving the c.-c.s.p. problem (where $\alpha > 0, \beta > 0, p \in (1, 2)$)

$$\min_{\|x\|_1 \le 1} \max_{\|y\|_1 \le 1} \left[\alpha \sum_i |x_i|^p - \beta \sum_i |y_i|^p + \sum_i [a_i x_i + b_i y_i + c_i x_i y_i] \right]. \tag{2.37}$$

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

$$\sup_{\mu\geq 0}\inf_{\nu\geq 0}\big[f(\mu,\nu):=\nu-\mu \\ +\sum_i \min_{x_i}\max_{y_i}\left[\alpha|x_i|^p+\mu|x_i|-\beta|y_i|^p-\nu|y_i|+a_ix_i+b_iy_i+c_ix_iy_i\right]\big].$$

The function $f(\mu,\nu)$ is concave in μ and convex in ν; computing first order information on f reduces to solving n simple two-dimensional c.-c.s.p. problems $\min_{x_i}\max_{y_i}[...]$ and, for all practical purposes, costs just $O(n)$ 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, e.g., the saddle point version of the Ellipsoid method (see, e.g., Nemirovski et al., 2009b). Thus, solving (2.37) within machine accuracy takes just $O(n)$ operations.

2.4.3 The strongly concave case

2.4.3.1 Situation and assumptions

Our current goal is to demonstrate that in the situation of the previous section, assuming ϕ 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 (2.3) 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 as follows:

C.2′. It is easy to find a solution $\widehat{z}$ to the variational inequality (2.25) 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 $\bar{y}\in\mathcal{Y}$).

Same 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

$$\langle\alpha\zeta+[\alpha_x\omega_x'(\widehat{x})+e;\alpha_y\omega_y'((\widehat{y}-\bar{y})/R)+f],u-\widehat{z}\rangle\geq 0\,\forall u\in\mathcal{Z}. \qquad (2.38)$$

We assume, as in the case of **C.2**, that when solving (2.25), we get both $\widehat{z}$ and ζ.

Furthermore, there are two new assumptions:

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

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

C.4. The E_x-component of $G(x,y)$ is independent of x, that is, $L_{xx}=0$ (see (2.27)).

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

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 1.6, ϕ possesses saddle points, and since $\underline{\phi}$ is strongly convex, the y-component of a saddle point is the unique maximizer, let it be denoted y_*, of $\underline{\phi}$ on $\mathcal{Y}$. We set

$$x_c=\mathrm{argmin}_{\mathcal{X}}\omega_x(\cdot),\ \Omega_x=\max_{\mathcal{X}}\omega_x(\cdot)-\min_{\mathcal{X}}\omega_x(\cdot),$$
$$\Omega_y=\max_{\|y\|_y\le1}\omega_y(y)-\min_{\mathcal{Y}}\omega_y(y)=\max_{\|y\|_y\le1}\omega_y(y)-\omega_y(0).$$

2.4.3.2 Algorithm MPb

The idea we intend to implement is the same we used in Section 1.4 when designing MD for strongly convex optimization: all other things being equal, the efficiency estimate (1.28) is the better the smaller is the domain $\mathcal{Z}$ (cf. the factor Ω in (2.17)). 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 $\bar{z}\in\mathcal{Z}$. Let us define the following entities:

$$\begin{array}{l}
\mathcal{Z}_R=\{(x;y)\in\mathcal{Z}:\|y-\bar{y}\|_y\le R\},\\
\mathcal{L}_R=2L_{xy}\sqrt{\Omega_x\Omega_y}R+L_{yy}\Omega_yR^2,\\
\alpha=[L_{xy}\sqrt{\Omega_x\Omega_y}R]/\mathcal{L}_R,\ \beta=[L_{xy}\sqrt{\Omega_x\Omega_y}R+L_{yy}\Omega_yR^2]/\mathcal{L}_R,\\
\omega^{R,\bar{y}}(x,y)=\frac{\alpha}{\Omega_x}\omega_x(x)+\frac{\beta}{\Omega_y}\omega_y([y-\bar{y}]/R)\\
\|(x,y)\|=\sqrt{\frac{\alpha}{\Omega_x}\|x\|_x^2+\frac{\beta}{\Omega_yR^2}\|y\|_y^2}
\end{array}\qquad(2.39)$$

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

$$\begin{array}{ll}
(a) & \max_{\mathcal{Z}_R}\omega^{R,\bar{y}}(\cdot)-\min_{\mathcal{Z}_R}\omega^{R,\bar{y}}(\cdot)\le1,\\
(b) & \forall(z,z'\in\mathcal{Z}):\|G(z)-G(z')\|_*\le\mathcal{L}_R\|z-z'\|.
\end{array}\qquad(2.40)$$

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

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

Let $z_1 = \bar{z}$ and consider for $t = 1, 2, ...$, $\gamma_t > 0$, the following recurrence $\mathcal{B}$ (cf. Section 2.4.1):

(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 (2.25) associated with $\mathcal{Z}$ and this operator; let the solution be denoted by w_t. Since the operator Ψ is of the form considered in **C.2′**, as a byproduct of our computation, we get a vector ζ_t such that $\forall u \in \mathcal{Z}$,

$$\zeta_t \in \mathcal{H}(w_t) \;\&\; \langle \gamma_t[\zeta_t + G(z_t)] + (\omega^{R,\bar{y}})'(w_t) - (\omega^{R,\bar{y}})'(z_t), u - w_t\rangle \geq 0, \quad (2.41)$$

cf. (2.38).

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

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

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 2.4. *Let Assumptions* **C.1**, **C.2′-C.4** *hold. Let the stepsizes satisfy the condition* $\gamma_\tau \geq \mathcal{L}_R^{-1}$ *and* $\delta_\tau \leq 0$ *for all* τ *(which for sure is so when* $\gamma_\tau = \mathcal{L}_R^{-1}$ *for all* τ*).*

(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:*

$$\begin{array}{lrcl} (a) & \widetilde{\phi}_R(x^t) - \underline{\phi}(y^t) \leq & \big[\sum_{\tau=1}^t \gamma_\tau\big]^{-1} \sum_{\tau=1}^t \gamma_\tau \langle F_\tau, w_\tau - z_*\rangle & \\ & \leq & \big[\sum_{\tau=1}^t \gamma_\tau\big]^{-1} \leq \frac{\mathcal{L}_R}{t}, & (2.42) \\ (b) & \|y^t - y_*\|_y^2 \leq & \frac{2}{\kappa}[\widetilde{\phi}_R(x^t) - \underline{\phi}(y^t)] \leq \frac{2\mathcal{L}_R}{\kappa t}, & \end{array}$$

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

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

$$\epsilon_{\mathrm{sad}}(x^t, y^t) := \overline{\phi}(x^t) - \underline{\phi}(y^t) \leq \frac{\mathcal{L}_R}{t}. \quad (2.43)$$

Proof. (i): Exactly the same argument as in the proof of Proposition 2.3,

with (2.40.b) in the role of (2.29), shows that

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

and that $\delta_\tau \le 0$ provided $\gamma_\tau = \mathcal{L}_R^{-1}$. Thus, under the premise of Proposition 2.4 we have

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

When $u = (x, y) \in \mathcal{Z}_R$, the right hand side of this inequality is ≤ 1 by (2.40.a) and due to $z_1 = \bar{z}$. Using the argument as in 2^0 of the proof of Proposition 2.1 we conclude that the left hand side in the inequality is $\ge \left[\sum_{\tau=1}^{t} \gamma_\tau\right] \left[\phi(x^t, y) - \phi(x, y^t)\right]$. Thus,

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

Taking the supremum of the left hand side of this inequality over $u \in \mathcal{Z}_R$ and noting that $\gamma_\tau \ge \mathcal{L}_R^{-1}$, we arrive at (2.42.$a$). Further, $\|\bar{y} - y_*\| \le R$, whence $\widetilde{\phi}_R(x^t) \ge \phi(x^t, y_*) \ge \underline{\phi}(y_*)$. Since y_* is the maximizer of the strongly concave, modulus κ w.r.t. $\|\cdot\|_y$, function $\underline{\phi}(\cdot)$ over $\mathcal{Y}$, we have

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

which is the first inequality in (2.42.b); the second inequality in (2.42.b) is given by (2.42.a). (i) is proved.

(ii): All we need to derive (ii) from (i), is to prove that under the premise of (ii) the quantities $\overline{\phi}(x^t) := \max_{y \in \mathcal{Y}} \phi(x^t, y)$ and $\widetilde{\phi}_R(x^t) := \max_{y \in \mathcal{Y}, \|y - \bar{y}\|_y \le R} \phi(x^t, y)$ 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 $\overline{\phi}$ and $\widetilde{\phi}_R$, we see that in the case in question the maximizer $\widetilde{y}$ of $\phi(x^t, y)$ over $\mathcal{Y}_R = \{y :\in \mathcal{Y} : \|y - \bar{y}\|_y \le R\}$ satisfies $\|\bar{y} - \widetilde{y}\|_y = R$. Since $\|\bar{y} - y_*\|_y \le R/2$, it follows that $\|y_* - \widetilde{y}\|_y \ge R/2$. Because $y_* \in \mathcal{Y}_R$, $\widetilde{y} = \operatorname{argmax}_{y \in \mathcal{Y}_R} \phi(x^t, y)$ and $\phi(x^t, y)$ is strongly concave, modulus κ w.r.t. $\|\cdot\|_y$, we get $\phi(x^t, y_*) \le \phi(x^t, \widetilde{y}) - \frac{\kappa}{2}\|y_* - \widetilde{y}\|_y^2 \le \phi(x^t, \widetilde{y}) - \frac{\kappa R^2}{8}$, whence $\widetilde{\phi}_R(x^t) = \phi(x^t, \widetilde{y}) \ge \phi(x^t, y_*) + \frac{\kappa R^2}{8}$. On the other hand, $\phi(x^t, y_*) \ge \underline{\phi}(y_*) \ge \underline{\phi}(y^t)$, and we arrive at $\widetilde{\phi}_R(x^t) - \underline{\phi}(y^t) \ge \frac{\kappa R^2}{8}$. At the same time, (2.42.a) says that $\widetilde{\phi}_R(x^t) - \underline{\phi}(y^t) \le \mathcal{L}_R t^{-1} < \frac{\kappa R^2}{8}$, where the latter inequality is due to $t > \frac{8\mathcal{L}_R}{\kappa R^2}$. We arrive at a desired contradiction. □

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

$$\|y^0 - y_*\| \le R_0/2, \tag{2.44}$$

be given, and let

$$R_k = 2^{-k/2}R_0,\ N_k = \mathrm{Ceil}\left(16\kappa^{-1}\left[2^{\frac{k+1}{2}}L_{xy}\sqrt{\Omega_x\Omega_y}R_0^{-1} + L_{yy}\Omega_y\right]\right),$$
$$M_k = \textstyle\sum_{j=1}^k N_j,\ k = 1, 2, ... \tag{2.45}$$

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

$$\|y^{k-1} - y_*\|_y \le R_{k-1}/2. \tag{I_{k-1}}$$

At stage k, we compute $(\widehat{x}^k, \widehat{y}^k) = z^{N_k}(R_{k-1}, y^{k-1})$, 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}). As about the stepsize policy, it can be an arbitrary policy satisfying $\gamma_\tau \ge \mathcal{L}^{-1}_{R_{k-1}}$ and $\delta_\tau \le 0$, e.g., $\gamma_\tau \equiv L^{-1}_{R_{k-1}}$, see Proposition 2.4. After $(\widehat{x}^k, \widehat{y}^k)$ is built, we set $y^k = \widehat{y}^k$ and pass to stage $k+1$.

Note that M_k is nothing but 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 2.4 in exactly the same way as Proposition 1.4 was derived from Proposition 1.3):

Proposition 2.5. *Let Assumptions* **C.1**,**C.2′-C.4** *hold, and let* $R_0 > 0$ *and* $y^0 \in Y$ *satisfy* (2.44). *Then algorithm MPb maintains relations* (I_{k-1}) *and*

$$\epsilon_{\mathrm{sad}}(\widehat{x}^k, \widehat{y}^k) \le \kappa 2^{-(k+3)}R_0^2, \tag{J_k}$$

$k = 1, 2, ...$ *Let, further,* k_* *be the smallest integer* k *such that* $k \ge 1$ *and* $2^{\frac{k}{2}} \ge kR_0\frac{L_{xy}\sqrt{\Omega_x\Omega_y}}{L_{yy}\Omega_y+\kappa}$. *Then*
— *for* $1 \le k < k_*$, *we have* $M_k \le O(1)k\frac{L_{yy}\Omega_y+\kappa}{\kappa}$ *and* $\epsilon_{\mathrm{sad}}(\widehat{x}^k, \widehat{y}^k) \le \kappa 2^{-k}R_0^2$;
— *for* $k \ge k_*$, *we have* $M_k \le O(1)N_k$ *and* $\epsilon_{\mathrm{sad}}(\widehat{x}^k, \widehat{y}^k) \le O(1)\frac{L_{xy}^2\Omega_x\Omega_y}{M_k^2}$.

Note that MPb behaves in the same way as the MD algorithm for strongly convex objectives of Section 1.4: 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_*.

2.4.3.3 Illustration III

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

$$\begin{gathered}\mathrm{Opt} = \min_{\xi\in\Xi} f(\xi), \quad f(\xi) = f_0(\xi) + \sum_{\ell=1}^{L} \tfrac{1}{2}\mathrm{dist}^2(A_\ell\xi - b_\ell, U_\ell + V_\ell), \\ \mathrm{dist}^2(w, W) = \min_{w'\in W}\|w - w'\|_2^2\end{gathered} \tag{2.46}$$

where

• $\Xi \subset E_\xi = \mathbb{R}^{n_\xi}$ is a convex compact set with a nonempty interior, E_ξ is equipped with a norm $\|\cdot\|_\xi$, and Ξ is equipped with a d.-g.f. $\omega_\xi(\xi)$ compatible with $\|\cdot\|_\xi$;

• $f_0(\xi) : \Xi \to \mathbb{R}$ is a simple continuous convex function, simplicity meaning that it is easy to solve auxiliary problems

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

• $U_\ell \subset \mathbb{R}^{m_\ell}$ are convex compact sets such that computing metric projection $\mathrm{Proj}_{U_\ell}(u) = \mathrm{argmin}_{u'\in U_\ell}\|u - u'\|_2$ onto U_ℓ is easy;

• $V_\ell \subset \mathbb{R}^{m_\ell}$ are polytopes given as $V_\ell = \mathrm{Conv}\{v_{\ell,1}, ..., v_{\ell,n_\ell}\}$.

On a close inspection, problem (2.46) admits a saddle point reformulation. Specifically, recalling that $\mathrm{S}_k = \{x \in \mathbb{R}^k_+ : \sum_i x_i = 1\}$ and setting

$$\begin{aligned}&\mathcal{X} = \{x = [\xi; x^1; ...; x^L] \in \Xi \times \mathrm{S}_{n_1} \times ... \times \mathrm{S}_{n_L}\} \subset E_x = \mathbb{R}^{n_\xi + n_1 + ... + n_L}, \\ &\mathcal{Y} = E_y := \mathbb{R}^{m_1}_{y^1} \times ... \times \mathbb{R}^{m_L}_{y^L}, \\ &g(y = (y^1, ..., y^L)) = \textstyle\sum_\ell g_\ell(y^\ell),\ g_\ell(y^\ell) = \frac{1}{2}[y^\ell]^T y^\ell + \max_{u_\ell\in U_\ell} u_\ell^T y^\ell,\end{aligned}$$

$$\begin{aligned}&B_\ell = [v_{\ell,1}, ..., v_{\ell,n_\ell}], \\ &A[\xi; x^1; ...; x^L] - b = [A_1\xi - B_1x^1; ...; A_L\xi - B_\ell x^L] - [b_1; ...; b_L], \\ &\phi(x, y) = f_0(\xi) + y^T[Ax - b] - g(y),\end{aligned}$$

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

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

so that a point $(x = [\xi; x^1; ...; x^L], y = [y^1; ...; y^L]) \in \mathcal{X} \times \mathcal{Y}$ which is an ϵ-solution to the c.-c.s.p. problem $\inf_{x\in\mathcal{X}} \sup_{\mathcal{Y}} \phi(x, y)$ induces an ϵ-solution, specifically, ξ, to the problem of interest (2.46):

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

Now let us apply to the saddle problem $\inf_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} \phi(x, y)$ algorithm

MPb. The required setup is as follows:

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

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

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

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

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

2. We equip $\mathcal{Y} = E_y = \mathbb{R}_y^{m_1+...+m_L}$ 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 Φ associated with (ϕ, z) is

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

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

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

Observe that with this setup, we satisfy **C.1** and **C.3-4** (**C.3** is satisfied with $\kappa = 1$). Let us verify that **C.2′** is satisfied as well. Indeed, in our current situation to find a solution $\widehat{z}$ to (2.25) means to solve the pair of convex optimization problems

$$\begin{array}{ll} (a) & \min\limits_{[\xi;x^1;...;x^L]\in\mathcal{X}} \big[p\alpha\omega_\xi(\xi) + qf_0(\xi) + e^T\xi \\ & \qquad\qquad + \sum_{\ell=1}^L \big[p\alpha_\ell \mathrm{Ent}(x^\ell) + e_\ell^T x^\ell\big]\big] \\ (b) & \min\limits_{y=[y^1;...;y^L]} \sum_{\ell=1}^L \big[\frac{r}{2}[y^\ell]^T y^\ell + sg_\ell(y^\ell) + f_\ell^T y^\ell\big] \end{array} \qquad (2.47)$$

where p, q, r, s are positive. Due to the direct product structure of $\mathcal{X}$, (2.47.a) decomposes into uncoupled problems $\min_{\xi\in\Xi}[p\alpha\omega_\xi(\xi) + f_0(\xi) + e^T\xi]$ and $\min_{x^\ell\in \mathrm{S}_\ell} \big[p\alpha_\ell \mathrm{Ent}(x^\ell) + e_\ell^T x^\ell\big]$. We have explicitly assumed that the first of these problems is easy; the remaining ones admit closed form solutions, cf. (1.39). (2.47.b) also is easy: a simple computation yields $y^\ell = -\frac{1}{r+s}[s\mathrm{Proj}_{U_\ell}(-s^{-1}f_\ell) + f_\ell]$, and it was assumed that it is easy to project onto U_ℓ.

The bottom line is that *we can solve* (2.46) *by algorithm MPb, the*

resulting efficiency estimate being

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

(see Proposition 2.5 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 $\alpha, \alpha_1, ..., \alpha_L$ to optimize the quantity $L_{xy}^2\Omega_x$. A rough optimization leads to the following: let μ_ℓ be the norm of the linear mapping $\xi \to A_\ell\xi$ induced by the norms $\|\cdot\|_\xi$, $\|\cdot\|_2$ in the argument and the image spaces, respectively, and let $\nu_\ell = \max_{1\le j\le n_\ell}\|v_{\ell,j}\|_2$. Choosing

$$\alpha = \sum\nolimits_{\ell=1}^{L}\mu_\ell^2,\ \alpha_\ell = \nu_\ell^2,\ 1 \le \ell \le L,$$

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

2.5 Accelerating First Order Methods by Randomization

We have seen in Section 2.2.1 that many important "well-structured" convex minimization programs reduce to just *bilinear* saddle point problems

$$\mathrm{SadVal} = \min_{x\in\mathcal{X}\subset E_x}\max_{y\in\mathcal{Y}\subset E_y}[\phi(x,y) := \langle a, x\rangle + \langle y, Ax - b\rangle] \tag{2.48}$$

the corresponding monotone operator admitting an *affine* selection

$$F(z=(x,y)) = (a + A^Ty, b - Ax) = (a,b) + \mathcal{F}z, \quad \mathcal{F}(x,y) = (A^Ty, -Ax). \tag{2.49}$$

Computing 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 a 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 favourable circumstances, allows for dramatic acceleration of FOMs in the extremely large-scale case.*

2.5.1 Randomizing matrix-vector multiplications

Let $u \in \mathbf{R}^n$. Computing the image of u under a linear mapping $u \mapsto Bu = \sum_{j=1}^n u_j b_j : \mathbf{R}^n \to E$ is easy to randomize: treat the vector $[|u_1|; ...; |u_n|]/\|u\|_1$ as a probability distribution on the set $\{b_1, ..., b_n\}$, draw from this distribution a sample $b_\jmath$ and set $\xi_u = \|u\|_1 \mathrm{sign}(u_\jmath) b_\jmath$, 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 ξ_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 $\jmath$ plus $O(\dim E)$ a.o. to compute $\|u\|_1 \mathrm{sign}(u_\jmath) b_\jmath$. Thus, the total cost of getting a single realization of ξ_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.

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

2.5.2 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 (2.48).

2.5.2.1 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 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 Ω_x, Ω_y according to (2.28). Further, we define $\|A\|_{x,y}$ as the norm of the linear mapping $x \mapsto Ax : E_x \to E_y$ 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 Π_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 to the *inside* case, as opposed to the *general* case, where support of P_v, $v \in \mathcal{Y}$, not necessary belongs to $\mathcal{Y}$. We will use Π_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 ξ_i are sampled, independently of each other, from Π_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 η_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 $A\xi$, $\xi \in \mathrm{supp}(\Pi_u)$, $A^T\eta$, $\eta \in \mathrm{supp}(P_v)$ is much easier than computing Au, $A^T v$ for a "general position" u and v.

We introduce the quantities

$$\sigma_x^2 = \sup_{u\in\mathcal{X}} \mathbf{E}\{\|A[\xi_u - u]\|_{y,*}^2\},\ \sigma_y^2 = \sup_{v\in\mathcal{Y}} \mathbf{E}\{\|A^T[\eta_v - v]\|_{x,*}^2\},$$
$$\Theta = 2\left[\Omega_x\sigma_y^2 + \Omega_y\sigma_x^2\right]. \tag{2.50}$$

where ξ_u, η_v are the just defined random vectors, 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 $\mathcal{Z}$ given by

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

what implies

$$\|(\xi,\eta)\|_* = \sqrt{2\Omega_x\|\xi\|_{x,*}^2 + 2\Omega_y\|\eta\|_{y,*}^2}. \tag{2.51}$$

For $z \in \mathcal{Z}^o$, $w \in \mathcal{Z}$ let (cf. the definition (1.4))

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

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

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

It is immediately seen that

$$\Omega[\mathcal{Z}] = \max_{\mathcal{Z}}\omega(\cdot) - \min_{\mathcal{Z}}\omega(\cdot) = 1. \tag{2.52}$$

and the affine monotone operator $F(z)$ given by (2.49) satisfies the relation

$$\forall z, z' : \|F(z) - F(z')\|_* \le \mathcal{L}\|z - z'\|,\ \mathcal{L} = 2\|A\|_{x,y}\sqrt{\Omega_x\Omega_y}. \tag{2.53}$$

2.5.2.2 Algorithm

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

$$\gamma = \min\left[\frac{1}{\sqrt{3}\mathcal{L}}, \frac{1}{\sqrt{3\Theta N}}\right] \tag{2.54}$$

and run N steps of the following randomized recurrence:

1. *Initialization:* We set $z_1 = \mathrm{argmin}_{\mathcal{Z}}\, \omega(\cdot)$;

2. *Step* $t = 1, 2, ..., N$*:* Given $z_t = (x_t, y_t) \in \mathcal{Z}^o$, we generate ξ_{x_t}, η_{y_t} as explained above, set $\zeta_t = (\xi_{x_t}, \eta_{y_t})$, compute $F(\zeta_t) = (a + A^T\eta_{y_t}, b - A\xi_{x_t})$ and

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

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

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

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

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

The efficiency estimate of algorithm MPr is given by the following

Proposition 2.6. *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*

$$\mathbf{E}\{\epsilon_{\mathrm{sad}}(z^N)\} \le \epsilon_N := \max\left[\frac{2\sqrt{3\Theta}}{\sqrt{N}}, \frac{4\sqrt{3}\|A\|_{x,y}\sqrt{\Omega_x\Omega_y}}{N}\right]; \tag{2.55}$$

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

Observe that in the general case we do not control the error $\epsilon_{\mathrm{sad}}(z^N)$ of the saddle point solution. Yet the bound (ii) of Proposition 2.55 allows 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 a "well structured" convex function $f = \overline{\phi}$ (cf. (2.4)).

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

$\omega(u) - \omega(z) - \langle \omega'(z), u - z\rangle$ and invoking Lemma 2.2, we get

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

whence, taking into account that $V_{z_1}(u) \le \Omega[\mathcal{Z}] = 1$ (see (2.52)) and $V_{z_{N+1}}(u) \ge 0$,

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

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

$$\mathbf{E}\left\{\max_{(x,y)\in\mathcal{Z}} \gamma N\left[\phi(x^N, y) - \phi(x, y^N)\right]\right\} \le 1 + \mathbf{E}\{\alpha_N\} + \mathbf{E}\{\beta_N\}, \quad (2.56)$$

$$\beta_N = \gamma\sum_{t=1}^N \left[\langle a, \xi_{\widehat{x}_t} - \widehat{x}_t\rangle + \langle b, \eta_{\widehat{y}_t} - \widehat{y}_t\rangle + \langle A\xi_{\widehat{x}_t}, \widehat{y}_t\rangle - \langle A^T\eta_{\widehat{x}_t}, \widehat{x}_t\rangle\right].$$

Let now $\mathbf{E}_{w_t}\{\cdot\}$ 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

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

and, recalling the origin of F's, $\|F_t^* - \widehat{F}_t^*\|_* \le \mathcal{L}\|z_t - w_t\|$ by (2.53). Since $3\gamma^2 \le \mathcal{L}^2$ by (2.54), we get

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

where the concluding inequality is due to the definition (2.50) of Θ and (2.51) of the norm $\|\cdot\|_*$. Thus, (2.56) implies that

$$\mathbf{E}\left\{\max_{(x,y)\in\mathcal{Z}}\left[\phi(x^N, y) - \phi(x, y^N)\right]\right\} \le 1/(N\gamma) + 3\Theta\gamma \le \epsilon_N, \quad (2.57)$$

due to the definition of ϵ_N. Now, in the inside case we clearly have $(x^N, y^N) \in \mathcal{Z}$, and therefore (2.57) implies (2.55). In the general case, we have $x^N \in \mathcal{X}$. Besides this, let x_* be the x-component of a saddle point of ϕ on $\mathcal{Z}$. Replacing in the left hand side of (2.57) 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 (2.57) that

$$\mathbf{E}\{\overline{\phi}(x^N)\} \le \epsilon_N + \mathbf{E}\{\phi(x_*, y^N)\} = \epsilon_N + \phi\big(x_*, \mathbf{E}\{y^N\}\big). \tag{2.58}$$

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

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

Thus, the right hand side in (2.58) is $\le \epsilon_N + \mathrm{SadVal}$, and (ii) follows. □

Remark 2.7. *We would like to stress here that MPr, along with 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 $\overline{\phi}(x^N)$ and $\underline{\phi}(y^N)$.*

2.5.2.3 Illustration IV: ℓ_1 minimization

Consider problem (2.5) with $\Xi = \{\xi \in \mathbb{R}^n : \|\xi\|_1 \le 1\}$. Representing Ξ as the image of the standard simplex $\mathrm{S}_{2n} = \{x \in \mathbb{R}^{2n}_+ : \sum_i x_i = 1\}$ under the mapping $x \mapsto J_n x$, $J_n = [I_n, -I_n]$, the problem reads

$$\mathrm{Opt} = \min_{x \in \mathrm{S}_{2n}} \|Ax - b\|_p \quad [A \in \mathbb{R}^{m \times 2n}] \tag{2.59}$$

We consider two cases: $p = \infty$ ("uniform fit," as in Dantzig selector) and $p = 2$ ("ℓ_2 fit," as in Lasso).

Uniform fit. Here (2.59) can be converted into the bilinear saddle point problem

$$\mathrm{Opt} = \min_{x \in \mathrm{S}_{2n}} \max_{y \in \mathrm{S}_{2m}} \left[\phi(x, y) := y^T J_m^T [Ax - b]\right]. \tag{2.60}$$

Setting $\|\cdot\|_x = \|\cdot\|_1$, $\omega_x(x) = \mathrm{Ent}(x)$, $\|\cdot\|_y = \|\cdot\|_1$, $\omega_y(y) = \mathrm{Ent}(y)$, let us specify Π_u, $u \in \mathrm{S}_{2n}$, and P_v, $v \in \mathrm{S}_{2m}$, according to the recipe from Section 2.5.1, that is, the random vector $\xi_u \sim \Pi_u$ with probability u_i is i-th basic orth, $i = 1, ..., m$, and similarly for $\eta_v \sim P_v$. We are in the inside case, and when setting $\|A\|_{1,\infty} = \max\limits_{i,j} |A_{ij}|$, 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] and

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

In this setting Proposition 2.6 reads:

Corollary 2.8. *For all positive integers k_x, k_y, N, one can find a random feasible solution (x^N, y^N) to (2.60) along with the quantities $\overline{\phi}(x^N) = \|Ax^N - b\|_\infty \ge$ Opt and a lower bound $\underline{\phi}(y^N)$ on* Opt *such that*

$$\text{Prob}\left\{\overline{\phi}(x^N) - \underline{\phi}(y^N) \le O(1)\frac{\|A\|_{1,\infty}\ln(2mn)}{\sqrt{N}\sqrt{\min[N, k_x+\ln(2m), k_y+\ln(2n)]}}\right\} \ge \frac{1}{2} \tag{2.61}$$

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

Note that our computation yields, along with (x^N, y^N), the quantities $\overline{\phi}(x^N)$ and $\underline{\phi}(y^N)$. Thus, when repeating the computation ℓ 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 (2.61) as large as $1-2^{-\ell}$. For example, with $k_x = k_y = 1$, assuming $\delta = \epsilon/\|A\|_{1,\infty} \le 1$, finding, with reliability $\ge 1-\beta$, an ϵ-solution to (2.60) 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 δ stays fixed and m, n are large, the lowest known so far cost of finding an ϵ-solution to problem (2.59) with unform 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 (2.22)). 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 δ and large m, n randomization does accelerate the solution process, the gain growing with m, n.

---

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

ℓ_2 fit. Here (2.59) can be converted into the bilinear saddle point problem

$$\mathrm{Opt} = \min_{x\in \mathrm{S}_{2n}} \max_{\|y\|_2\le 1} \left[\phi(x,y) := y^T[Ax-b]\right]. \tag{2.62}$$

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

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

Now Proposition 2.6 reads:

Corollary 2.9. *For all positive integers k_x, k_y, N, one can find a random feasible solution x^N to (2.59) (where $p = 2$), along with the vector Ax^N, such that*

$$\begin{aligned}&\mathrm{Prob}\Big\{\|Ax^N - b\|_2 \le \mathrm{Opt}\\ &\qquad + O(1)\tfrac{\|A\|_{2,\infty}\sqrt{\ln(2n)}}{\sqrt{N}}\sqrt{\tfrac{1}{N} + \tfrac{1}{k_x\ln(2n)} + \tfrac{\ln(2n)\Gamma^2(A)}{k_y+\ln(2n)}}\Big\} \ge \tfrac{1}{2},\end{aligned} \tag{2.63}$$

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

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

Here again, repeating the computation ℓ times and choosing the best among the resulting solutions to (2.59), we make the probability of the left hand side event in (2.63) as large as $1-2^{-\ell}$. For instance, with $k_x = k_y = 1$, assuming $\delta := \epsilon/\|A\|_{2,\infty} \le 1$, finding, with reliability $\ge 1-\beta$, an ϵ-solution to (2.59) 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. (2.22)). Now the advantages of randomization when δ is fixed and m, n are large are not as evident as in the case of uniform fit – 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 of $O(1/\sqrt{m})$, and D is a random diagonal matrix with i.i.d. diagonal entries taking values ± 1

with equal probabilities. This preprocessing converts (2.59) 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 $\Gamma(\bar{A}) \leq O(1)\sqrt{\ln(mn\beta^{-1})}$. 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 δ, is a small fraction of the cost of deterministic computation. Thus, there is a meaningful range of values of δ, 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.

2.6 Notes and Remarks

1. The Mirror Prox algorithm was proposed in (Nemirovski, 2004); its modification capable to handle the stochastic case, where the precise values of the monotone operator associated with (2.3) are replaced by unbiased random estimates of these values (cf. section 1.5) is developed in (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; Nemirovskii and Yudin, 1978)) and (b) exploiting *extragradient* steps: instead of the usual gradient-type update $z \mapsto z_+ = \mathrm{Prox}_z(\gamma F(z))$ used in the saddle point MP (section 1.6), the update $z \mapsto w = \mathrm{Prox}_z(\gamma F(z)) \mapsto z_+ = \mathrm{Prox}_z(\gamma F(w))$ 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 Yu. Nesterov in his dual-extrapolation algorithm (Nesterov, 2007b).

2. The material in sections 2.4.1, 2.4.3 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 equally well treated via Nesterov's technique, which perhaps is not the case for Illustration II. For other schemes of accelerating FOMs via exploiting problem's structure see (Nesterov, 2007a; Beck and Teboulle, 2008) and references therein.

3. The material of section 2.5.1 originates from (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 2.5.1 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.), and to solving certain semidefinite programs in (Arora and Kale, 2007; d'Aspremont, 2009). Furthermore, as we have already mentioned, both MD and MP admit modifications (see Nemirovski et al., 2009a; Juditsky et al., 2008) capable to handle c.-c.s.p. problems (not necessary 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 non-asymptotical upper bounds on the *expected* inaccuracy of random approximate solutions z^N (which, same 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 bilinear problem (2.48) and with the randomized estimates of $F(w_t)$ defined as $F(\widehat{\zeta}_t)$, we get a new option — to define z^N as the average of the points $\widehat{\zeta}_t$. As a result, we do know $F(z^N)$ and thus can easily assess the quality of z^N, n.b. Remark 2.7. To the best of our knowledge, this option has being 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 (2.60) 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 (2.48) is crucial when solving *parametric* bilinear saddle point problems. Specifically, many important applications reduce to problems of the form

$$\max\left\{\rho : \mathrm{SadVal}(\rho) := \min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\phi_\rho(x,y) \le 0\right\}, \tag{2.64}$$

where $\phi_\rho(x,y)$ is a bilinear function affinely depending on ρ. For example, the "true" ℓ_1 minimization problem as arising in sparsity-oriented Signal Processing is $\mathrm{Opt} = \min_\xi\{\|\xi\|_1 : \|A\xi - b\|_p \le \delta\}$, which is nothing but

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

From the complexity viewpoint, the best known to us way to process (2.64) is to solve the "master" problem $\max\{\rho : \mathrm{SadVal}(\rho) \le 0\}$ by an appropriate

first order root-finding routine, the (approximate) first order information on SadVal(·) being provided by a first order saddle point algorithm. The ability of MPr algorithm to provide accurate bounds of the value SadVal(·) of the "inner" saddle point problems, makes it the method of choice when solving extremely large parametric saddle point problems (2.64) (for more details on this subject, see Juditsky et al., 2010).

(Teboulle, 1997)

(Goldfarb and Scheinberg, 2010)

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

References

A. Arora and S. Kale. A combinatorial primal-dual approach to semidefinite programs. In: *Proc. of the 39-th annual ACM Symp. on Theory of Comp.*, pages 227–236, 2007.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2008.

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

A. d'Aspremont. Subsampling algorithms for semidefinite programming. Technical Report arXiv:0803.1990v5, November 2009. URL http://arxiv.org/abs/0803.1990, 2009.

D. Goldfarb and K. Scheinberg. Fast first order method for separable convex optimization with line search. *Technical report, Department IEOR, Columbia University*, 2010.

M. D. Grigoriadis and L. G. Khachiyan. A sublinear-time randomized approximation algorithm for matrix games. *Oper. Res. Lett.*, 18:53–58, 1995.

A. Juditsky and A. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. . http://www2.isye.gatech.edu/ nemirovs/LargeDevSubmitted.pdf, 2008.

A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror prox algorithm. . http://www2.isye.gatech.edu/∼nemirovs/SMP_240408.pdf, 2008.

A. Juditsky, F. Kilinc Karzan, and A. Nemirovski. ℓ_1 minimization via randomized first order algorithms. http://www.optimization-online.org/DB_HTML/2010/05/2618.html, 2010.

G. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody (in Russian)*, 12: 747–756, 1976.

G. Korpelevich. Extrapolation gradient methods and relation to modified lagrangeans. *Ekonomika i Matematicheskie Metody (in Russian)*, 19:694–703, 1983.

K. Kumar, C. Bhattacharya, and R. Harihan. A randomized algorithm for large scale support vector learning. In: J.C. Platt, D. Koller, Y. Singer, and S Roweis, editors, *Advances in Neural Information Processing Systems* **20** (2008), MIT Press.

A. Nemirovski. Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.*, 15:229–251, 2004.

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

A. Nemirovski, S. Onn, and U. Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35:52–78, 2009b.

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

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

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

Yu. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM J. Optim.*, 16(1):235–239, 2005b.

Yu. Nesterov. Gradient methods for minimizing composite objective function. *CORE Discussion Paper 2007/76*, 2007a.

Yu. Nesterov. Dual extrapolation and its application for solving variational inequalities and related problems. *Math. Prog.*, 109(2-3):319–344, 2007b.

M. A. Noor. New extragradient-type methods for general variational inequalities. *J. Math. Anal. Appl.*, 277:379–394, 2003.

M. Teboulle. Convergence of proximal-like algorithms. *SIAM J. Optim.*, 7:

1069–1083, 1997.