Comput Optim Appl
DOI 10.1007/s10589-014-9723-3

Mirror Prox algorithm for multi-term composite minimization and semi-separable problems

Niao He · Anatoli Juditsky · Arkadi Nemirovski

Received: 21 May 2014

Abstract In the paper, we develop a composite version of Mirror Prox algorithm for solving convex–concave saddle point problems and monotone variational inequalities of special structure, allowing to cover saddle point/variational analogies of what is usually called "composite minimization" (minimizing a sum of an easy-to-handle nonsmooth and a general-type smooth convex functions "as if" there were no nonsmooth component at all). We demonstrate that the composite Mirror Prox inherits the favourable (and unimprovable already in the large-scale bilinear saddle point case) $O(1/\epsilon)$ efficiency estimate of its prototype. We demonstrate that the proposed approach can be successfully applied to Lasso-type problems with several penalizing terms (e.g. acting together ℓ_1 and nuclear norm regularization) and to problems of semi-separable structures considered in the alternating directions methods, implying in both cases methods with the $O(1/\epsilon)$ complexity bounds.

Keywords Numerical algorithms for variational problems · Composite optimization · Minimization problems with multi-term penalty · Proximal methods

Mathematics Subject Classification 65K10 · 65K05 · 90C06 · 90C25 · 90C47

N. He (✉) · A. Nemirovski
Georgia Institute of Technology, Atlanta, GA 30332, USA
e-mail: nhe6@gatech.edu

A. Nemirovski
e-mail: nemirovs@isye.gatech.edu

A. Juditsky
LJK, Université Grenoble Alpes, B.P. 53, 38041 Grenoble Cedex 9, France
e-mail: anatoli.juditsky@imag.fr

Published online: 08 January 2015

1 Introduction

1.1 Motivation

Our work is inspired by the recent trend of seeking efficient ways for solving problems with hybrid regularizations or mixed penalty functions in fields such as machine learning, image restoration, signal processing and many others. We are about to present two instructive examples (for motivations, see, e.g., [2,6,7]).

Example 1 (Matrix completion) Our first motivating example is matrix completion problem, where we want to reconstruct the original matrix $y \in \mathbf{R}^{n\times n}$, known to be both sparse and low-rank, given noisy observations of part of the entries. Specifically, our observation is $b = P_\Omega y + \xi$, where Ω is a given set of cells in an $n \times n$ matrix, $P_\Omega y$ is the restriction of $y \in \mathbf{R}^{n\times n}$ onto Ω, and ξ is a random noise. A natural way to recover y from b is to solve the optimization problem

$$\mathrm{Opt} = \min_{y\in\mathbf{R}^{n\times n}} \left\{ \frac{1}{2}\|P_\Omega y - b\|_2^2 + \lambda\|y\|_1 + \mu\|y\|_{\mathrm{nuc}} \right\} \tag{1}$$

where $\mu, \lambda > 0$ are regularization parameters. Here $\|y\|_2 = \sqrt{\mathrm{Tr}(y^T y)}$ is the Frobenius norm, $\|y\|_1 = \sum_{i,j=1}^n |y_{ij}|$ is the ℓ_1-norm, and $\|y\|_{\mathrm{nuc}} = \sum_{i=1}^n \sigma_i(y)$ ($\sigma_i(y)$ are the singular values of y) is the nuclear norm of a matrix $y \in \mathbf{R}^{n\times n}$.

Example 2 (*Image recovery*) Our second motivating example is image recovery problem, where we want to recover an image $y \in \mathbf{R}^{n\times n}$ from its noisy observations $b = Ay + \xi$, where Ay is a given affine mapping (e.g. the restriction operator P_Ω defined as above, or some blur operator), and ξ is a random noise. Assume that the image can be decomposed as $y = y_{\mathrm{L}} + y_{\mathrm{S}} + y_{\mathrm{sm}}$ where y_{L} is of low rank, y_{sm} is the matrix of contamination by a "smooth background signal", and y_{S} is a sparse matrix of "singular corruption." Under this assumption in order to recover y from b, it is natural to solve the optimization problem

$$\begin{aligned}\mathrm{Opt} = \min_{y_{\mathrm{L}}, y_{\mathrm{S}}, y_{\mathrm{sm}}\in\mathbf{R}^{n\times n}} \{&\|A(y_{\mathrm{L}} + y_{\mathrm{S}} + y_{\mathrm{sm}}) - b\|_2 + \mu_1\|y_{\mathrm{L}}\|_{\mathrm{nuc}} \\ &+ \mu_2\|y_{\mathrm{S}}\|_1 + \mu_3\|y_{\mathrm{sm}}\|_{\mathrm{TV}}\}\end{aligned} \tag{2}$$

where $\mu_1, \mu_2, \mu_3 > 0$ are regularization parameters. Here $\|y\|_{\mathrm{TV}}$ is the total variation of an image y:

$$\begin{aligned}\|y\|_{\mathrm{TV}} &= \|\nabla_i y\|_1 + \|\nabla_j y\|_1, \\ (\nabla_i y)_{ij} &= y_{i+1,j} - y_{i,j},\ [i;j] \in \mathbf{Z}^2 :\ 1 \le i < n-1, 1 \le j < n, \\ (\nabla_j y)_{ij} &= y_{i,j+1} - y_{i,j},\ [i;j] \in \mathbf{Z}^2 :\ 1 \le i < n, 1 \le j < n-1.\end{aligned}$$

These and other examples motivate addressing the following *multi-term composite minimization problem*

$$\min_{y\in Y}\left\{\sum_{k=1}^{K}[\psi_k(A_k y+b_k)+\Psi_k(A_k y+b_k)]\right\}, \tag{3}$$

and, more generally, the *semi-separable problem*

$$\min_{[y^1;...;y^K]\in Y_1\times\cdots\times Y_K}\left\{\sum_{k=1}^{K}\left[\psi_k(y^k)+\Psi_k(y^k)\right]:\ \sum_{k=1}^{K}A_k y^k=b\right\}. \tag{4}$$

Here for $1\le k\le K$ the domains Y_k are closed and convex, $\psi_k(\cdot)$ are convex Lipschitz-continuous functions, and $\Psi_k(\cdot)$ are convex functions which are "simple and fit Y_k".[1]

The problem of multi-term composite minimization (3) has been considered (in a somewhat different setting) in [22] for $K=2$. When $K=1$, problem (3) becomes the usual composite minimization problem:

$$\min_{u\in U}\{\psi(u)+\Psi(u)\} \tag{5}$$

which is well studied in the case where $\psi(\cdot)$ is a *smooth* convex function and $\Psi(\cdot)$ is a simple non-smooth function. For instance, it was shown that the composite versions of Fast Gradient Method originating in Nesterov's seminal work [21] and further developed by many authors (see, e.g., [3,4,8,25,27] and references therein), as applied to (5), work *as if* there were no nonsmooth term at all and exhibit the $O(1/t^2)$ convergence rate, which is the optimal rate attainable by first order algorithms of large-scale smooth convex optimization. Note that these algorithms cannot be directly applied to problems (3) with $K>1$.

The problem with semi-separable structures (4) for $K=2$, has also been extensively studied using the augmented Lagrangian approach (see, e.g., [5,11,12,16,23,24,26,28] and references therein). In particular, much work was carried out on the alternating directions method of multipliers (ADMM, see [5] for an overview), which optimizes the augmented Lagrangian in an alternating fashion and exhibits an overall $O(1/t)$ convergence rate. Note that the available accuracy bounds for those algorithms involve optimal values of Lagrange multipliers of the equality constraints (cf. [23]). Several variants of this method have been developed recently to adjust to the case for $K>2$ (see, e.g.[10]), however, most of these algorithms require to solve iteratively subproblems of type (5) especially with the presence of non-smooth terms in the objective.

[1] The precise meaning of simplicity and fitting will be specified later. As of now, it suffices to give a couple of examples. When Ψ_k is the ℓ_1 norm, Y_k can be the entire space, or the centered at the origin ℓ_p-ball, $1\le p\le 2$; when Ψ_k is the nuclear norm, Y_k can be the entire space, or the centered at the origin Frobenius/nuclear norm ball.

1.2 Our contribution

In this paper, we do not assume smoothness of functions ψ_k, but instead, we suppose that ψ_k are *saddle point representable*:

$$\psi_k(y^k) = \sup_{z^k \in Z_k} \left[\phi_k(y^k, z^k) - \overline{\Psi}_k(z^k) \right], \quad 1 \le k \le K, \tag{6}$$

where $\phi_k(\cdot, \cdot)$ are smooth functions which are convex–concave (i.e., convex in their first and concave in the second argument), Z_k are convex and compact, and $\overline{\Psi}_k(\cdot)$ are simple convex functions on Z_k. Let us consider, for instance, the multi-term composite minimization problem (3). Under (6), the primal problem (3) allows for the saddle point reformulation:

$$\min_{y \in Y} \max_{[z^1; \ldots; z^k] \in Z_1 \times \cdots \times Z_K} \left\{ \sum_{k=1}^K \left[\phi_k(A_k y + b_k, z^k) - \overline{\Psi}_k(z^k) + \Psi_k(A_k y + b_k) \right] \right\} \tag{7}$$

Note that when there are no Ψ_k, $\overline{\Psi}_k$'s, problem (7) becomes a convex–concave saddle point problem with smooth cost function, studied in [14]. In particular, it was shown in [14] that Mirror Prox (MP) algorithm originating from [17], when applied to the saddle point problem (7), exhibits the "theoretically optimal" convergence rate $O(1/t)$. Our goal in this paper is to develop novel $O(1/t)$-converging first order algorithms for problem (7) (and also the related saddle point reformulation of the problem in (4)), which appears to be the best rate known, under circumstances, from the literature (and established there in essentially less general setting than the one considered below).

Our key observation is that composite problem (3), (6) can be reformulated as a smooth linearly constrained saddle point problem by simply moving the nonsmooth terms into the problem domain. Namely, problem (3) , (6) can be written as

$$\begin{aligned}
&\min_{\substack{y \in Y,\ [y^k; \tau^k] \in Y_k^+ \\ 1 \le k \le K}} \max_{\substack{[z^k; \sigma^k] \in Z_k^+ \\ 1 \le k \le K}} \left\{ \sum_{k=1}^K \left[\phi_k(y^k, z^k) - \sigma^k + \tau^k \right] : y^k \right. \\
&\qquad \left. = A_k y + b_k,\ k = 1, \ldots, K \right\} \\
&Y_k^+ = \left\{ [y^k; \tau^k] : y^k \in Y_k, \tau^k \ge \Psi_k(y^k) \right\}, Z_k^+ \\
&\qquad = \left\{ [z^k; \sigma^k] : z^k \in Z_k, \sigma^k \ge \overline{\Psi}_k(z^k) \right\},\ k = 1, \ldots, K.
\end{aligned}$$

We can further approximate the resulting problem by penalizing the equality constraints, thus passing to

$$\min_{\substack{y \in Y,\ [y^k; \tau^k] \in Y_k^+ \\ 1 \le k \le K}} \max_{\substack{[z^k; \sigma^k] \in Z_k^+ \\ 1 \le k \le K}} \left\{ \sum_{k=1}^K \left[\phi_k(y^k, z^k) - \sigma^k + \tau^k + \rho_k \| y^k - A_k y - b_k \|_2 \right] \right\}$$

$$= \min_{\substack{y\in Y,\ [y^k;\tau^k]\in Y_k^+ \\ 1\le k\le K}} \max_{\substack{w^k\in W_k,\ [z^k;\sigma^k]\in Z_k^+ \\ 1\le k\le K}} \left\{ \sum_{k=1}^K \Big[\phi_k(y^k, z^k) - \sigma^k + \tau^k + \rho_k \langle y^k - A_k y - b_k, w^k \rangle \Big] \right\}, \tag{8}$$

where $\rho_k > 0$ are penalty parameters and $W_k = \{w^k : \|w^k\|_2 \le 1\}, k = 1, \ldots, K$.

We solve the convex–concave saddle point problem (8) with *smooth* cost function by $O(1/t)$-converging Mirror Prox algorithm. It is worth to mention that if the functions ϕ_k, Ψ_k are Lipschitz continuous on the domains $A_k Y + b_k$, and ρ_k are selected properly, the saddle point problem is exactly equivalent to the problem of interest.

The monotone operator F associated with the saddle point problem in (8) has a special structure: the variables can be split into two blocks u (all y-, z- and w-variables) and v (all τ- and σ-variables) in such a way that the induced partition of F is $F = [F_u(u); F_v]$ with the u-component F_u depending solely on u and constant v-component F_v. We demonstrate below that in this case the basic MP algorithm admits a "composite" version which works essentially "as if" there were no v-component at all. This composite version of MP will be the working horse of all subsequent developments.

The main body of this paper is organized as follows. In Sect. 2 we present required background on variational inequalities with monotone operators and convex–concave saddle points. In Sect. 3 we present and justify the composite MP algorithm. In Sects. 4 and 5, we apply our approach to problems (3), (6) and (4), (6). In Sect. 4.4, we illustrate our approach (including numerical results) as applied to the motivating examples. All proofs missing in the main body of the paper are relegated to the Appendix.

2 Preliminaries: variational inequalities and accuracy certificates

Execution protocols and accuracy certificates. Let X be a nonempty closed convex set in a Euclidean space E and $F(x) : X \to E$ be a vector field.

Suppose that we process (X, F) by an algorithm which generates a sequence of search points $x_t \in X$, $t = 1, 2, \ldots$, and computes the vectors $F(x_t)$, so that after t steps we have at our disposal *t-step execution protocol* $\mathcal{I}_t = \{x_\tau, F(x_\tau)\}_{\tau=1}^t$. By definition, an *accuracy certificate* for this protocol is simply a collection $\lambda^t = \{\lambda_\tau^t\}_{\tau=1}^t$ of nonnegative reals summing up to 1. We associate with the protocol $\mathcal{I}_t$ and accuracy certificate λ^t two quantities as follows:

- *Approximate solution* $x^t(\mathcal{I}_t, \lambda^t) := \sum_{\tau=1}^t \lambda_\tau^t x_\tau$, which is a point of X;
- *Resolution* $\mathrm{Res}(X'|\mathcal{I}_t, \lambda^t)$ *on a subset* $X' \ne \emptyset$ of X given by

$$\mathrm{Res}(X'|\mathcal{I}_t, \lambda^t) = \sup_{x\in X'} \sum_{\tau=1}^t \lambda_\tau^t \langle F(x_\tau), x_\tau - x \rangle. \tag{9}$$

The role of those notions in the optimization context is explained next.[2]

Variational inequalities. Assume that F is *monotone*, i.e.,

$$\langle F(x) - F(y), x - y\rangle \geq 0, \quad \forall x, y \in X \tag{10}$$

and let our goal be to approximate a weak solution to the variational inequality (v.i.) vi(X, F) associated with (X, F); weak solution is defined as a point $x_* \in X$ such that

$$\langle F(y), y - x_*\rangle \geq 0 \ \forall y \in X. \tag{11}$$

A natural (in)accuracy measure of a candidate weak solution $x \in X$ to vi(X, F) is the *dual gap function*

$$\epsilon_{\mathrm{VI}}(x\big|X, F) = \sup_{y\in X}\langle F(y), x - y\rangle \tag{12}$$

This inaccuracy is a convex nonnegative function which vanishes exactly at the set of weak solutions to the vi(X, F) .

Proposition 1 *For every t, every execution protocol $\mathcal{I}_t = \{x_\tau \in X, F(x_\tau)\}_{\tau=1}^t$ and every accuracy certificate λ^t one has $x^t := x^t(\mathcal{I}_t, \lambda^t) \in X$. Besides this, assuming F monotone, for every closed convex set $X' \subset X$ such that $x^t \in X'$ one has*

$$\epsilon_{\mathrm{VI}}\Big(x^t\big|X', F\Big) \leq Res\Big(X'\big|\mathcal{I}_t, \lambda^t\Big). \tag{13}$$

Proof Indeed, x^t is a convex combination of the points $x_\tau \in X$ with coefficients λ_τ^t, whence $x^t \in X$. With X' as in the premise of Proposition, we have

$$\begin{aligned}\forall y \in X' : \langle F(y), x^t - y\rangle &= \sum_{\tau=1}^t \lambda_\tau^t\langle F(y), x_\tau - y\rangle \leq \sum_{\tau=1}^t \lambda_\tau^t\langle F(x_\tau), x_\tau - y\rangle\\ &\leq \mathrm{Res}(X'\big|\mathcal{I}_t, \lambda^t),\end{aligned}$$

where the first $\leq$ is due to monotonicity of F. □

Convex–concave saddle point problems Now let $X = X_1 \times X_2$, where X_i is a closed convex subset in Euclidean space E_i, $i = 1, 2$, and $E = E_1 \times E_2$, and let $\Phi(x^1, x^2) : X_1 \times X_2 \to \mathbf{R}$ be a locally Lipschitz continuous function which is convex in $x^1 \in X_1$ and concave in $x^2 \in X_2$. X_1, X_2, Φ give rise to the saddle point problem

$$\mathrm{SadVal} = \min_{x^1\in X_1}\max_{x^2\in X_2}\Phi(x^1, x^2), \tag{14}$$

[2] Our exposition follows.

two induced convex optimization problems

$$\begin{aligned}\mathrm{Opt}(P) &= \min_{x^1\in X_1}\left[\overline{\Phi}(x^1) = \sup_{x^2\in X_2}\Phi(x^1,x^2)\right] (P)\\ \mathrm{Opt}(D) &= \max_{x^2\in X_2}\left[\underline{\Phi}(x^2) = \inf_{x^1\in X_1}\Phi(x^1,x^2)\right] (D)\end{aligned} \tag{15}$$

and a vector field $F(x^1,x^2) = [F_1(x^1,x^2); F_2(x^1,x^2)]$ specified (in general, non-uniquely) by the relations

$$\forall (x^1,x^2)\in X_1\times X_2 : F_1(x^1,x^2)\in \partial_{x^1}\Phi(x^1,x^2),\ F_2(x^1,x^2)\in \partial_{x^2}[-\Phi(x^1,x^2)].$$

It is well known that F is monotone on X, and that weak solutions to the $\mathrm{vi}(X,F)$ are exactly the saddle points of Φ on $X_1\times X_2$. These saddle points exist if and only if (P) and (D) are solvable with equal optimal values, in which case the saddle points are exactly the pairs (x^1_*, x^2_*) comprised by optimal solutions to (P) and (D). In general, $\mathrm{Opt}(P)\geq \mathrm{Opt}(D)$, with equality definitely taking place when at least one of the sets X_1, X_2 is bounded; if both are bounded, saddle points do exist. To avoid unnecessary complications, from now on, when speaking about a convex–concave saddle point problem, we assume that the problem is *proper*, meaning that $\mathrm{Opt}(P)$ and $\mathrm{Opt}(D)$ are reals; this definitely is the case when X is bounded.

A natural (in)accuracy measure for a candidate $x=[x^1;x^2]\in X_1\times X_2$ to the role of a saddle point of Φ is the quantity

$$\begin{aligned}\epsilon_{\mathrm{sad}}(x\big|X_1,X_2,\Phi) &= \overline{\Phi}(x^1)-\underline{\Phi}(x^2)\\ &= [\overline{\Phi}(x^1)-\mathrm{Opt}(P)]+[\mathrm{Opt}(D)-\underline{\Phi}(x^2)]\\ &\quad+\underbrace{[\mathrm{Opt}(P)-\mathrm{Opt}(D)]}_{\geq 0}\end{aligned} \tag{16}$$

This inaccuracy is nonnegative and is the sum of the duality gap $\mathrm{Opt}(P)-\mathrm{Opt}(D)$ (always nonnegative and vanishing when one of the sets X_1, X_2 is bounded) and the inaccuracies, in terms of respective objectives, of x^1 as a candidate solution to (P) and x^2 as a candidate solution to (D).

The role of accuracy certificates in convex–concave saddle point problems stems from the following observation: □

Proposition 2 *Let X_1, X_2 be nonempty closed convex sets, $\Phi: X:=X_1\times X_2\to \mathbf{R}$ be a locally Lipschitz continuous convex–concave function, and F be the associated monotone vector field on X.*

Let $\mathcal{I}_t=\{x_\tau=[x^1_\tau;x^2_\tau]\in X, F(x_\tau)\}_{\tau=1}^t$ be a t-step execution protocol associated with (X,F) and $\lambda^t=\{\lambda^t_\tau\}_{\tau=1}^t$ be an associated accuracy certificate. Then $x^t := x^t(\mathcal{I}_t,\lambda^t)=[x^{1,t};x^{2,t}]\in X$.

Assume, further, that $X'_1\subset X_1$ and $X'_2\subset X_2$ are closed convex sets such that

$$x^t\in X' := X'_1\times X'_2. \tag{17}$$

Then

$$\epsilon_{\text{Sad}}\left(x^t\middle|X_1', X_2', \Phi\right) = \sup_{x^2\in X_2'} \Phi\left(x^{1,t}, x^2\right) - \inf_{x^1\in X_1'} \Phi\left(x^1, x^{2,t}\right) \le Res\left(X'\middle|\mathcal{I}_t, \lambda^t\right). \tag{18}$$

In addition, setting $\widetilde{\Phi}(x^1) = \sup_{x^2\in X_2'} \Phi\left(x^1, x^2\right)$, *for every* $\widehat{x}^1 \in X_1'$ *we have*

$$\widetilde{\Phi}\left(x^{1,t}\right) - \widetilde{\Phi}\left(\widehat{x}^1\right) \le \widetilde{\Phi}\left(x^{1,t}\right) - \Phi(\widehat{x}^1, x^{2,t}) \le Res(\{\widehat{x}^1\} \times X_2'|\mathcal{I}_t, \lambda^t). \tag{19}$$

In particular, when the problem $Opt = \min_{x^1\in X_1'} \widetilde{\Phi}(x^1)$ *is solvable with an optimal solution* x_*^1, *we have*

$$\widetilde{\Phi}(x^{1,t}) - Opt \le Res\left(\{x_*^1\} \times X_2'|\mathcal{I}_t, \lambda^t\right). \tag{20}$$

Proof The inclusion $x^t \in X$ is evident. For every set $Y \subset X$ we have

$$\begin{aligned}
&\forall [p;q] \in Y:\\
&\text{Res}(Y|\mathcal{I}_t, \lambda^t) \ge \sum_{\tau=1}^{t} \lambda_\tau^t \left[\langle F_1(x_\tau^1), x_\tau^1 - p\rangle + \langle F_2(x_\tau^2), x_\tau^2 - q\rangle\right]\\
&\ge \sum_{\tau=1}^{t} \lambda_\tau^t \left[[\Phi\left(x_\tau^1, x_\tau^2\right) - \Phi\left(p, x_\tau^2\right)] + \left[\Phi\left(x_\tau^1, q\right) - \Phi\left(x_\tau^1, x_\tau^2\right)\right]\right]\\
&\text{[by the origin of } F \text{and since } \Phi \text{is convex–concave]}\\
&= \sum_{\tau=1}^{t} \lambda_\tau^t \left[\Phi\left(x_\tau^1, q\right) - \Phi\left(p, x_\tau^2\right)\right] \ge \Phi\left(x^{1,t}, q\right) - \Phi\left(p, x^{2,t}\right)\\
&\text{[by origin of } x^t \text{and since } \Phi \text{is convex–concave]}
\end{aligned}$$

Thus, for every $Y \subset X$ we have

$$\sup_{[p;q]\in Y} \left[\Phi\left(x^{1,t}, q\right) - \Phi\left(p, x^{2,t}\right)\right] \le \text{Res}(Y|\mathcal{I}_t, \lambda^t). \tag{21}$$

Now assume that (17) takes place. Setting $Y = X' := X_1' \times X_2'$ and recalling what ϵ_{Sad} is, (21) yields (18). With $Y = \{\widehat{x}^1\} \times X_2'$, (21) yields the second inequality in (19); the first inequality in (19) is evident due to $x^{2,t} \in X_2'$. □

3 Composite Mirror Prox algorithm

3.1 The situation

Let U be a nonempty closed convex domain in a Euclidean space E_u, E_v be a Euclidean space, and X be a nonempty closed convex domain in $E = E_u \times E_v$. We denote vectors from E by $x = [u; v]$ with blocks u, v belonging to E_u and E_v, respectively.

We assume that

A1: E_u is equipped with a norm $\|\cdot\|$, the conjugate norm being $\|\cdot\|_*$, and U is equipped with a *distance-generating function* (d.g.f.) $\omega(\cdot)$ (that is, with a continuously differentiable convex function $\omega(\cdot) : U \to \mathbf{R}$) which is *compatible* with $\|\cdot\|$, meaning that ω is strongly convex, modulus 1, w.r.t. $\|\cdot\|$.
Note that d.g.f. ω defines the *Bregman distance*

$$V_u(w) := \omega(w) - \omega(u) - \langle \omega'(u), w - u\rangle \geq \frac{1}{2}\|w - u\|^2,\ u, w \in U, \quad (22)$$

where the concluding inequality follows from strong convexity, modulus 1, of the d.g.f. w.r.t. $\|\cdot\|$.
In the sequel, we refer to the pair $\|\cdot\|$, $\omega(\cdot)$ as to *proximal setup* for U.

A2: the image PX of X under the projection $x = [u; v] \mapsto Px := u$ is contained in U.

A3: we are given a vector field $F(u, v) : X \to E$ on X of the special structure as follows:

$$F(u, v) = [F_u(u); F_v],$$

with $F_u(u) \in E_u$ and $F_v \in E_v$. Note that F is independent of v. We assume also that

$$\forall u, u' \in U : \|F_u(u) - F_u(u')\|_* \leq L\|u - u'\| + M \quad (23)$$

with some $L < \infty$, $M < \infty$.

A4: the linear form $\langle F_v, v\rangle$ of $[u; v] \in E$ is bounded from below on X and is coercive on X w.r.t. v: whenever $[u_t; v_t] \in X, t = 1, 2, \ldots$ is a sequence such that $\{u_t\}_{t=1}^{\infty}$ is bounded and $\|v_t\|_2 \to \infty$ as $t \to \infty$, we have $\langle F_v, v_t\rangle \to \infty$, $t \to \infty$.

Our goal in this section is to show that *in the situation in question, proximal type processing F* (say, F is monotone on X, and we want to solve the variational inequality given by F and X) *can be implemented "as if" there were no v-components in the domain and in F.*

A generic application we are aiming at is as follows. We want to solve a "composite" saddle point problem

$$\mathrm{SadVal} = \min_{u_1 \in U_1} \max_{u_2 \in U_2} \left[\phi(u_1, u_2) + \Psi_1(u_1) - \Psi_2(u_2)\right], \quad (24)$$

where

- $U_1 \subset E_1$ and $U_2 \subset E_2$ are nonempty closed convex sets in Euclidean spaces E_1, E_2
- ϕ is a smooth (with Lipschitz continuous gradient) convex–concave function on $U_1 \times U_2$
- $\Psi_1 : U_1 \to \mathbf{R}$ and $\Psi_2 : U_2 \to \mathbf{R}$ are convex functions, perhaps nonsmooth, but "fitting" the domains U_1, U_2 in the following sense: for $i = 1, 2$, we can equip E_i with a norm $\|\cdot\|_{(i)}$, and U_i - with a compatible with this norm d.g.f. $\omega_i(\cdot)$ in such a way that optimization problems of the form

$$\min_{u_i \in U_i} [\alpha\omega_i(u_i) + \beta\Psi_i(u_i) + \langle \xi, u_i \rangle] \qquad [\alpha > 0, \beta > 0] \tag{25}$$

are easy to solve.

Our ultimate goal is to solve (24) "as if" there were no (perhaps) nonsmooth terms Ψ_i. With our approach, we intend to "get rid" of the nonsmooth terms by "moving" them into the description of problem's domains. To this end, we act as follows:

- For $i = 1, 2$, we set $X_i = \{x_i = [u_i; v_i] \in E_i \times \mathbf{R} : u_i \in U_i, v_i \geq \Psi_i(u_i)\}$ and set

$$\begin{aligned} U &:= U_1 \times U_2 \subset E_u := E_1 \times E_2, E_v = \mathbf{R}^2, \\ X &= \Big\{x = [u = [u_1; u_2]; v = [v_1; v_2]] : u_i \in U_i, v_i \geq \Psi_i(u_i), i = 1, 2\Big\} \\ &\subset E_u \times E_v, \end{aligned}$$

  thus ensuring that $PX \subset U$, where $P[u; v] = u$;
- We rewrite the problem of interest equivalently as

$$\mathrm{SadVal} = \min_{x^1 = [u_1; v_1] \in X_1} \max_{x^2 = [u_2; v_2] \in X_2} [\Phi(u_1, v_1; u_2, v_2) = \phi(u_1, u_2) + v_1 - v_2] \tag{26}$$

  Note that Φ is convex–concave and smooth. The associated monotone operator is

$$\begin{aligned} &F(u = [u_1; u_2], v = [v_1; v_2]) \\ &\quad = \big[F_u(u) = [\nabla_{u_1}\phi(u_1, u_2); -\nabla_{u_2}\phi(u_1, u_2)]; F_v = [1; 1]\big] \end{aligned}$$

  and is of the structure required in **A3**. Note that F is Lipschitz continuous, so that (23) is satisfied with properly selected L and with $M = 0$.

We intend to process the reformulated saddle point problem (26) with a properly modified state-of-the-art MP saddle point algorithm [17]. In its basic version and as applied to a variational inequality with Lipschitz continuous monotone operator (in particular, to a convex–concave saddle point problem with smooth cost function), this algorithm exhibits $O(1/t)$ rate of convergence, which is the best rate achievable with First Order saddle point algorithms as applied to large-scale saddle point problems (even those with bilinear cost function). The basic MP would require to equip the

domain $X = X_1 \times X_2$ of (26) with a d.g.f. $\omega(x_1, x_2)$ resulting in an easy-to-solve auxiliary problems of the form

$$\min_{x=[u_1;u_2;v_1;v_2]\in X} [\omega(x) + \langle \xi, x\rangle], \tag{27}$$

which would require to account in ω, in a nonlinear fashion, for the v-variables (since ω should be a strongly convex in both u- and v-variables). While it is easy to construct ω from our postulated "building blocks" ω_1, ω_2 leading to easy-to-solve problems (25), this construction results in auxiliary problems (27) somehow more complicated than problems (25). To overcome this difficulty, below we develop a "composite" MP algorithm taking advantage of the special structure of F, as expressed in **A3**, and preserving the favorable efficiency estimates of the prototype. The modified MP operates with the auxiliary problems of the form

$$\min_{x=[u_1;u_2;v_1;v_2]\in X_1\times X_2} \sum_{i=1}^{2} [\alpha_i \omega_i(u_i) + \beta_i v_i + \langle \xi_i, u_i\rangle], \quad [\alpha_i > 0, \beta_i > 0]$$

that is, with pairs of uncoupled problems

$$\min_{[u_i;v_i]\in X_i} [\alpha_i \omega_i(u_i) + \beta_i v_i + \langle \xi_i, u_i\rangle], \ i = 1, 2;$$

recalling that $X_i = \{[u_i; v_i] : u_i \in U_i, v_i \geq \Psi_i(u_i)\}$, these problems are nothing but the easy-to-solve problems (25).

3.2 Composite Mirror Prox algorithm

Given the situation described in Sect. 3.1, we define the associated *prox-mapping*: for $\xi = [\eta; \zeta] \in E$ and $x = [u; v] \in X$,

$$\begin{aligned} P_x(\xi) &\in \operatorname*{Argmin}_{[s;w]\in X} \left\{\langle \eta - \omega'(u), s\rangle + \langle \zeta, w\rangle + \omega(s)\right\} \\ &\equiv \operatorname*{Argmin}_{[s;w]\in X} \{\langle \eta, s\rangle + \langle \zeta, w\rangle + V_u(s)\} \end{aligned} \tag{28}$$

Observe that $P_x([\eta; \gamma F_v])$ is well defined whenever $\gamma > 0$—the required Argmin is nonempty due to the strong convexity of ω on U and assumption **A4** (for verification, see item 0° in Appendix 1). Now consider the process as follows:

$$\begin{aligned} x_1 &:= [u_1; v_1] \in X; \\ y_\tau &:= [u'_\tau; v'_\tau] = P_{x_\tau}(\gamma_\tau F(x_\tau)) = P_{x_\tau}(\gamma_\tau [F_u(u_\tau); F_v]) \\ x_{\tau+1} &:= [u_{\tau+1}; v_{\tau+1}] = P_{x_\tau}(\gamma_\tau F(y_\tau)) = P_{x_\tau}(\gamma_\tau [F_u(u'_\tau); F_v]), \end{aligned} \tag{29}$$

where $\gamma_\tau > 0$; the latter relation, due to the above, implies that the recurrence (29) is well defined.

Theorem 1 *In the setting of Sect. 3.1, assuming that* **A1**–**A4** *hold, consider the Composite Mirror Prox recurrence 29 (CoMP) with stepsizes* $\gamma_\tau > 0$, $\tau = 1, 2, \ldots$ *satisfying the relation:*

$$\delta_\tau := \gamma_\tau \langle F_u(u'_\tau) - F_u(u_\tau), u'_\tau - u_{\tau+1} \rangle - V_{u'_\tau}(u_{\tau+1}) - V_{u_\tau}(u'_\tau) \le \gamma_\tau^2 M^2. \quad (30)$$

Then the corresponding execution protocol $\mathcal{I}_t = \{y_\tau, F(y_\tau)\}_{\tau=1}^t$ *admits accuracy certificate* $\lambda^t = \{\lambda_\tau^t = \gamma_\tau / \sum_{i=1}^t \gamma_i\}$ *such that for every* $X' \subset X$ *it holds*

$$Res(X' | \mathcal{I}_t, \lambda^t) \le \frac{\Theta[X'] + M^2 \sum_{\tau=1}^t \gamma_\tau^2}{\sum_{\tau=1}^t \gamma_\tau}, \quad \Theta[X'] = \sup_{[u;v] \in X'} V_{u_1}(u). \quad (31)$$

Relation (30) is definitely satisfied when $0 < \gamma_\tau \le (\sqrt{2}L)^{-1}$, *or, in the case of* $M = 0$, *when* $\gamma_\tau \le L^{-1}$.

Invoking Propositions 1, 2, we arrive at the following

Corollary 1 *Under the premise of Theorem 1, for every* $t = 1, 2, \ldots$, *setting*

$$x^t = [u^t; v^t] = \frac{1}{\sum_{\tau=1}^t \gamma_\tau} \sum_{\tau=1}^t \gamma_\tau y_\tau.$$

we ensure that $x^t \in X$ *and that*
(i) *In the case when* F *is monotone on* X, *we have*

$$\epsilon_{\mathrm{VI}}(x^t | X, F) \le \left[\sum_{\tau=1}^t \gamma_\tau\right]^{-1} \left[\Theta[X] + M^2 \sum_{\tau=1}^t \gamma_\tau^2\right]. \quad (32)$$

(ii) *Let* $X = X_1 \times X_2$, *and let* F *be the monotone vector field associated with the saddle point problem (14) with convex–concave locally Lipschitz continuous cost function* Φ. *Then*

$$\epsilon_{\mathrm{Sad}}(x^t | X_1, X_2, \Phi) \le \left[\sum_{\tau=1}^t \gamma_\tau\right]^{-1} \left[\Theta[X] + M^2 \sum_{\tau=1}^t \gamma_\tau^2\right]. \quad (33)$$

In addition, assuming that problem (P) *in (15) is solvable with optimal solution* x_*^1 *and denoting by* $x^{1,t}$ *the projection of* $x^t \in X = X_1 \times X_2$ *onto* X_1, *we have*

$$\overline{\Phi}(x^{1,t}) - Opt(P) \le \left[\sum_{\tau=1}^t \gamma_\tau\right]^{-1} \left[\Theta[\{x_*^1\} \times X_2] + M^2 \sum_{\tau=1}^t \gamma_\tau^2\right]. \quad (34)$$

Remark When F is Lipschitz continuous (that is, (23) holds true with some $L > 0$ and $M = 0$), the requirements on the stepsizes imposed in the premise of Theorem 1 reduce to $\delta_\tau \le 0$ for all τ and are definitely satisfied with the constant stepsizes $\gamma_\tau = 1/L$. Thus, in the case under consideration we can assume w.l.o.g. that $\gamma_\tau \ge 1/L$, thus ensuring that the upper bound on $\mathrm{Res}(X' | \mathcal{I}_t, \lambda^t)$ in (31) is $\le \Theta[X'] L t^{-1}$. As a result, (34) becomes

$$\overline{\Phi}(x^{1,t}) - \mathrm{Opt}(P) \le \Theta[\{x_*^1\} \times X_2] L t^{-1}. \quad (35)$$

3.3 Modifications

In this section, we demonstrate that in fact our algorithm admits some freedom in building approximate solutions, freedom which can be used to improve to some extent solutions' quality. Modifications to be presented originate from [19]. We assume that we are in the situation described in Sect. 3.1, and assumptions **A1**–**A4** are in force. In addition, we assume that

A5: The vector field F described in **A3** is monotone, and the variational inequality given by (X, F) has a weak solution:

$$\exists x_* = [u_*; v_*] \in X : \langle F(y), y - x_* \rangle \geq 0 \ \forall y \in X \tag{36}$$

Lemma 1 *In the situation from Sect. 3.1 and under assumptions* **A1**–**A5**, *for* $R \geq 0$ *let us set*

$$\widehat{\Theta}(R) = \max_{u,u' \in U} \left\{ V_u(u') : \|u - u_1\| \leq R, \|u' - u_1\| \leq R \right\} \tag{37}$$

(this quantity is finite since ω *is continuously differentiable on* U*), and let*

$$\{x_\tau = [u_\tau; v_\tau] : \tau \leq N + 1, y_\tau : \tau \leq N\}$$

be the trajectory of the N*-step MP algorithm (29) with stepsizes* $\gamma_\tau > 0$ *which ensure (30) for* $\tau \leq N$. *Then for all* $u \in U$ *and* $t \leq N + 1$,

$$0 \leq V_{u_t}(u) \leq \widehat{\Theta}(\max[R_N, \|u - u_1\|]), \quad R_N := 2\left(2V_{u_1}(u_*) + M^2 \sum_{\tau=1}^{N-1} \gamma_\tau^2\right)^{1/2}, \tag{38}$$

with u_* *defined in (36).*

Proposition 3 *In the situation of Sect. 3.1 and under assumptions* **A1**–**A5**, *let* N *be a positive integer, and let* $\mathcal{I}_N = \{y_\tau, F(y_\tau)\}_{\tau=1}^N$ *be the execution protocol generated by* N*-step CoMP (29) with stepsizes* γ_τ *ensuring (30). Let also* $\lambda^N = \{\lambda_1, \ldots, \lambda_N\}$ *be a collection of positive reals summing up to 1 and such that*

$$\lambda_1/\gamma_1 \leq \lambda_2/\gamma_2 \leq \cdots \leq \lambda_N/\gamma_N. \tag{39}$$

Then for every $R \geq 0$, *with* $X_R = \{x = [u; v] \in X : \|u - u_1\| \leq R\}$ *one has*

$$Res(X_R|\mathcal{I}_N, \lambda^N) \leq \frac{\lambda_N}{\gamma_N} \widehat{\Theta}(\max[R_N, R]) + M^2 \sum_{\tau=1}^{N} \lambda_\tau \gamma_\tau, \tag{40}$$

with $\widehat{\Theta}(\cdot)$ *and* R_N *defined by (37) and (38).*

Invoking Propositions 1, 2, we arrive at the following modification of Corollary 1.

Corollary 2 *Under the premise and in the notation of Proposition 3, setting*

$$x^N = [u^N; v^N] = \sum_{\tau=1}^{N} \lambda_\tau y_\tau.$$

we ensure that $x^N \in X$. *Besides this,*
(i) *Let* X' *be a closed convex subset of* X *such that* $x^N \in X'$ *and the projection of* X' *on the u-space is contained in* $\|\cdot\|$*-ball of radius* R *centered at* u_1. *Then*

$$\epsilon_{\rm VI}(x^N | X', F) \le \frac{\lambda_N}{\gamma_N} \widehat{\Theta}(\max[R_N, R]) + M^2 \sum_{\tau=1}^{N} \lambda_\tau \gamma_\tau. \tag{41}$$

(ii) *Let* $X = X_1 \times X_2$ *and* F *be the monotone vector field associated with saddle point problem (14) with convex–concave locally Lipschitz continuous cost function* Φ. *Let, further,* X_i' *be closed convex subsets of* X_i, $i = 1, 2$, *such that* $x^N \in X_1' \times X_2'$ *and the projection of* $X_1' \times X_2'$ *onto the u-space is contained in* $\|\cdot\|$*-ball of radius* R *centered at* u_1. *Then*

$$\epsilon_{\rm Sad}(x^N | X_1', X_2', \Phi) \le \frac{\lambda_N}{\gamma_N} \widehat{\Theta}(\max[R_N, R]) + M^2 \sum_{\tau=1}^{N} \lambda_\tau \gamma_\tau. \tag{42}$$

4 Multi-term composite minimization

In this section, we focus on the problem (3), (6) of multi-term composite minimization.

4.1 Problem setting

We intend to consider problem (3), (6) in the situation as follows. For a nonnegative integer K and $0 \le k \le K$ we are given

1. Euclidean spaces E_k and $\overline{E}_k$ along with their nonempty closed convex subsets Y_k and Z_k, respectively;
2. Proximal setups for (E_k, Y_k) and $(\overline{E}_k, Z_k)$, that is, norms $p_k(\cdot)$ on E_k, norms $q_k(\cdot)$ on $\overline{E}_k$, and d.g.f.'s $\omega_k(\cdot) : Y_k \to \mathbf{R}$, $\overline{\omega}_k(\cdot) : Z_k \to \mathbf{R}$ compatible with $p_k(\cdot)$ and $q_k(\cdot)$, respectively;
3. Affine mappings $y^0 \mapsto A_k y^0 + b_k : E_0 \to E_k$, where $y^0 \mapsto A_0 y^0 + b_0$ is the identity mapping on E_0;
4. Lipschitz continuous convex functions $\psi_k(y^k) : Y_k \to \mathbf{R}$ along with their *saddle point representations*

$$\psi_k(y^k) = \sup_{z^k \in Z_k} [\phi_k(y^k, z^k) - \overline{\Psi}_k(z^k)], \quad 0 \le k \le K, \tag{43}$$

where $\phi_k(y^k, z^k) : Y_k \times Z_k \to \mathbf{R}$ are smooth (with Lipschitz continuous gradients) functions convex in $y^k \in Y_k$ and concave in $z^k \in Z_k$, and $\overline{\Psi}_k(z^k) : Z_k \to \mathbf{R}$ are Lipschitz continuous convex functions such that the problems of the form

$$\min_{z^k \in Z_k} \left[\overline{\omega}_k(z^k) + \langle \xi^k, z^k \rangle + \alpha \overline{\Psi}_k(z^k) \right] \quad [\alpha > 0] \tag{44}$$

are easy to solve;

5. Lipschitz continuous convex functions $\Psi_k(y^k) : Y_k \to \mathbf{R}$ such that the problems of the form

$$\min_{y^k \in Y_k} \left[\omega_k(y^k) + \langle \xi^k, y^k \rangle + \alpha \Psi_k(y^k) \right] \quad [\alpha > 0] \tag{45}$$

are easy to solve;

6. For $1 \le k \le K$, the norms $\pi_k^*(\cdot)$ on E_k are given, with conjugate norms $\pi_k(\cdot)$, along with d.g.f.'s $\widehat{\omega}_k(\cdot) : W_k := \{w^k \in E_k : \pi_k(w^k) \le 1\} \to \mathbf{R}$ which are strongly convex, modulus 1, w.r.t. $\pi_k(\cdot)$ such that the problems

$$\min_{w^k \in W_k} \left[\widehat{\omega}_k(w^k) + \langle \xi^k, w^k \rangle \right] \tag{46}$$

are easy to solve.

The outlined data define the sets

$$Y_k^+ = \left\{ [y^k; \tau^k] : \ y^k \in Y_k, \tau^k \ge \Psi_k(y^k) \right\} \subset E_k^+ := E_k \times \mathbf{R}, \ 0 \le k \le K,$$
$$Z_k^+ = \left\{ [z^k; \sigma^k] : \ z^k \in Z_k, \sigma^k \ge \overline{\Psi}_k(z^k) \right\} \subset \overline{E}_k^+ := \overline{E}_k \times \mathbf{R}, \ 0 \le k \le K.$$

The problem of interest (3), (6) along with its saddle point reformulation in the just defined situation read

$$\mathrm{Opt} = \min_{y^0 \in Y_0} \left\{ f(y^0) := \sum_{k=0}^K \left[\psi_k(A_k y^0 + b_k) + \Psi_k(A_k y^0 + b_k) \right] \right\} \tag{47a}$$

$$= \min_{y^0 \in Y_0} \left\{ f(y^0) = \max_{\{z^k \in Z_k\}_{k=0}^K} \sum_{k=0}^K \left[\phi_k(A_k y^0 + b_k, z^k) \right. \right. \tag{47b}$$
$$\left. \left. + \Psi_k(A_k y^0 + b_k) - \overline{\Psi}_k(z^k) \right] \right\}$$

which we rewrite equivalently as

$$\mathrm{Opt} = \min_{\substack{\{[y^k;\tau^k]\}_{k=0}^K \\ \in Y_0^+ \times \cdots \times Y_K^+}} \max_{\substack{\{[z^k;\sigma^k]\}_{k=0}^K \\ \in Z_0^+ \times \cdots \times Z_K^+}} \left\{ \sum_{k=0}^K \left[\phi_k(y^k, z^k) + \tau^k - \sigma^k \right] : \right. \tag{47c}$$
$$\left. y^k = A_k y^0 + b_k, \ 1 \le k \le K \right\}.$$

From now on we make the following assumptions

B1: *We have* $A_k Y_0 + b_k \subset Y_k$, $1 \le k \le K$;
B2: *For* $0 \le k \le K$, *the sets* Z_k *are bounded. Further, the functions* Ψ_k *are below bounded on* Y_k, *and the functions* $f_k = \psi_k + \Psi_k$ *are coercive on* Y_k: *whenever* $y_t^k \in Y_k$, $t = 1, 2, \ldots$, *are such that* $p_k(y_t^k) \to \infty$ *as* $t \to \infty$, *we have* $f_k(y_t^k) \to \infty$.

Note that **B1** and **B2** imply that the saddle point problem (47c) is solvable; let $\left\{[y_*^k; \tau_*^k]\right\}_{0 \le k \le K}$; $\left\{[z_*^k; \sigma_*^k]\right\}_{0 \le k \le K}$ be the corresponding saddle point.

4.2 Course of actions

Given $\rho_k > 0$, $1 \le k \le K$, we approximate (47c) by the problem

$$\widehat{\mathrm{Opt}} = \min_{\substack{\{[y^k;\tau^k]\}_{k=0}^K \\ \in Y_0^+ \times \cdots \times Y_K^+ \in Z_0^+ \times \cdots \times Z_K^+}} \max_{\{[z^k;\sigma^k]\}_{k=0}^K} \left\{ \sum_{k=0}^K \left[\phi_k\left(y^k, z^k\right) + \tau^k - \sigma^k \right] + \sum_{k=1}^K \rho_k \pi_k^*\left(y^k - A_k y^0\right) \right\} \tag{48a}$$

$$= \min_{\substack{x^1 \in X_1 \\ := Y_0^+ \times \cdots \times Y_K^+ := Z_0^+ \times \cdots \times Z_K^+ \times W_1 \times \cdots W_K}} \max_{x^2 \in X_2} \Phi\left(\underbrace{\left\{[y^k; \tau^k]\right\}_{k=0}^K}_{x^1}, \underbrace{\left[\left\{[z^k; \sigma^k]\right\}_{k=0}^K ; \{w^k\}_{k=1}^K \right]}_{x^2} \right) \tag{48b}$$

where

$$\Phi(x^1, x^2) = \sum_{k=0}^K \left[\phi_k(y^k, z^k) + \tau^k - \sigma^k \right] + \sum_{k=1}^K \rho_k \langle w^k, y^k - A_k y^0 - b_k \rangle.$$

Observe that the monotone operator $F(x^1, x^2) = \left[F_1(x^1, x^2); F_2(x^1, x^2)\right]$ associated with the saddle point problem in (48b) is given by

$$\begin{aligned} F_1(x^1, x^2) &= \left[\nabla_{y^0} \phi_0\left(y^0, z^0\right) - \sum_{k=1}^K \rho_k A_k^T w^k; 1; \left\{ \nabla_{y^k} \phi_k\left(y^k, z^k\right) + \rho_k w^k; 1 \right\}_{k=1}^K \right], \\ F_2(x^1, x^2) &= \left[\left\{ -\nabla_{z^k} \phi_k\left(y^k, z^k\right); 1 \right\}_{k=0}^K ; \left\{ -\rho_k [y^k - A_k y^0 - b_k] \right\}_{k=1}^K \right]. \end{aligned} \tag{49}$$

Now let us set

- $U = \left\{ u = \left[y^0; \ldots; y^K; z^0; \ldots; z^K; w^1; \ldots; w^K\right] : y^k \in Y_k,\ z^k \in Z_k,\ 0 \le k \le K, \pi_k(w^k) \le 1,\ 1 \le k \le K \right\}$,
- $X = \left\{ \begin{array}{l} x = \left[u = \left[y^0; \ldots; y^K; z^1; \ldots; z^K; w^1; \ldots; w^K\right]; \right. \\ \qquad v = \left[\tau^0; \ldots; \tau^K; \sigma^0; \ldots; \sigma^K\right] : \\ u \in U,\ \tau^k \ge \Psi_k(y^k),\ \sigma^k \ge \overline{\Psi}_k(z^k),\ 0 \le \\ \quad k \le K \end{array} \right\}$,

so that $PX \subset U$, cf. assumption **A2** in Sect. 3.1.

The variational inequality associated with the saddle point problem in (48b) can be treated as the variational inequality on the domain X with the monotone operator

$$F(x = [u; v]) = [F_u(u); F_v],$$

where

$$F_u(\underbrace{[y^0; \ldots; y^K;\ z^0; \ldots; z^K;\ w^1; \ldots; w^K]}_{u}) = \begin{bmatrix} \nabla_y \phi_0(y^0, z^0) - \sum_{k=1}^K \rho_k A_k^T w^k \\ \{\nabla_y \phi_k(y^k, z^k) + \rho_k w^k\}_{k=1}^K \\ \{-\nabla_z \phi_k(y^k, z^k\}_{k=0}^K \\ \{-\rho_k[y^k - A_k y^0 - b_k]\}_{k=1}^K \end{bmatrix}$$
$$F_v(\underbrace{[\tau^0; \ldots; \tau^K; \sigma^0; \ldots; \sigma^K]}_{v}) = [1; \ldots; 1]. \tag{50}$$

This operator meets the structural assumptions **A3** and **A4** from Sect. 3.1 (**A4** is guaranteed by **B2**). We can equip U and its embedding space E_u with the proximal setup $\|\cdot\|,\ \omega(\cdot)$ given by

$$\begin{aligned} \|u\| &= \sqrt{\sum_{k=0}^K \left[\alpha_k p_k^2(y^k) + \beta_k q_k^2(z^k)\right] + \sum_{k=1}^K \gamma_k \pi_k^2(w^k)}, \\ \omega(u) &= \sum_{k=0}^K \left[\alpha_k \omega_k(y^k) + \beta_k \overline{\omega}_k(z^k)\right] + \sum_{k=1}^K \gamma_k \widehat{\omega}_k(w^k), \end{aligned} \tag{51}$$

where $\alpha_k, \beta_k, 0 \le k \le K$, and $\gamma_k, 1 \le k \le K$, are positive aggregation parameters[3]. Observe that carrying out a step of the CoMP algorithm presented in Sect. 3.2 requires computing F at $O(1)$ points of X and solving $O(1)$ auxiliary problems of the form

$$\begin{aligned} \min_{\substack{[y^0;\ldots;y^K;z^0;\ldots;z^K],\\ [;w^1;\ldots;w^K;\tau^0;\ldots;\tau^K;\sigma^0;\ldots;\sigma^K]}} & \left\{\sum_{k=0}^K \left[a_k \omega_k(y^k) + \langle \xi_k, y^k\rangle + b_k \tau^k\right]\right. \\ & \left. + \sum_{k=0}^K \left[c_k \overline{\omega}_k(z^k) + \langle \eta_k, z^k\rangle + d_k \sigma^k\right] + \sum_{k=1}^K \left[e_k \widehat{\omega}_k(w^k) + \langle \zeta_k, w^k\rangle\right]\right\} : \\ & y^k \in Y_k, \tau^k \ge \Psi_k(y^k),\ z^k \in Z_k, \sigma^k \ge \overline{\Psi}_k(y^k),\ 0 \le k \le K, \\ & \pi_k(w^k) \le 1,\ 1 \le k \le K, \end{aligned}$$

with positive $a_k, \ldots, e_k$, and we have assumed that these problems are easy to solve.

[3] In principle, these parameters should be chosen to optimize the resulting efficiency estimates; this indeed is doable, provided that we have at our disposal upper bounds on the Lipschitz constants of the components of F_u and that U is bounded, see [17, Section 5] or [14, Section 6.3.3].

4.3 "Exact penalty"

Let us make one more assumption:

C: *For* $1 \le k \le K$,
- ψ_k *are Lipschitz continuous on* Y_k *with constants* G_k *w.r.t.* $\pi_k^*(\cdot)$,
- Ψ_k *are Lipschitz continuous on* Y_k *with constants* H_k *w.r.t.* $\pi_k^*(\cdot)$.

Given a feasible solution $\overline{x} = [\overline{x}^1; \overline{x}^2]$, $\overline{x}^1 := \{[\overline{y}^k; \overline{\tau}^k] \in Y_k^+\}_{k=0}^K$ to the saddle point problem (48b), let us set

$$\widehat{y}^0 = \overline{y}^0;\ \widehat{y}^k = A_k\overline{y}^0 + b_k,\ 1 \le k \le K;\ \widehat{\tau}^k = \Psi_k(\widehat{y}^k),\ 0 \le k \le K,$$

thus getting another feasible (by assumption **B1**) solution $\widehat{x}=\left[\widehat{x}^1=\{[\widehat{y}^k; \widehat{\tau}^k]\}_{k=0}^K; \overline{x}^2\right]$ to (48b). We call $\widehat{x}^1$ *correction* of $\overline{x}^1$. For $1 \le k \le K$ we clearly have

$$\begin{aligned}\psi_k\left(\widehat{y}^k\right) \le \psi_k\left(\overline{y}^k\right) + G_k\pi_k^*\left(\widehat{y}^k - \overline{y}^k\right) = \psi_k\left(\overline{y}^k\right) + G_k\pi_k^*\left(\overline{y}^k - A_k\overline{y}^0 - b_k\right),\\ \widehat{\tau}^k = \Psi_k(\widehat{y}^k) \le \Psi_k\left(\overline{y}^k\right) + H_k\pi_k^*\left(\widehat{y}^k - \overline{y}^k\right) \le \overline{\tau}^k + H_k\pi_k^*\left(\overline{y}^k - A_k\overline{y}^0 - b_k\right),\end{aligned}$$

and $\widehat{\tau}^0 = \Psi_0(\overline{y}^0) \le \overline{\tau}^0$. Hence for $\overline{\Phi}(x^1) = \max\limits_{x^2 \in X_2} \Phi(x^1, x^2)$ we have

$$\overline{\Phi}(\widehat{x}^1) \le \overline{\Phi}(\overline{x}^1) + \sum_{k=1}^K [H_k + G_k]\pi_k^*(\overline{y}^k - A_k\overline{y}^0 - b_k) - \sum_{k=1}^K \rho_k\pi_k^*(\overline{y}^k - A_k\overline{y}^0 - b_k).$$

We see that under the condition

$$\rho_k \ge G_k + H_k,\ 1 \le k \le K, \tag{52}$$

correction does not increase the value of the primal objective of (48b), whence the saddle point value $\widehat{\mathrm{Opt}}$ of (48b) is $\ge$ the optimal value Opt in the problem of interest (47a). Since the opposite inequality is evident, we arrive at the following

Proposition 4 *In the situation of Sect. 4.1, let assumptions* **B1**, **B2**, **C** *and (52) hold true. Then*

(i) the optimal value $\widehat{\mathrm{Opt}}$ *in* (48a) *coincides with the optimal value Opt in the problem of interest (47a);*

(ii) consequently, if $\overline{x} = [\overline{x}^1; \overline{x}^2]$ *is a feasible solution of the saddle point problem in (48b), then the correction* $\widehat{x}^1 = \{[\widehat{y}^k; \widehat{\tau}^k]\}_{k=0}^K$ *of* $\overline{x}^1$ *is a feasible solution to the problem of interest (47c), and*

$$f(\widehat{y}^0) - Opt \le \epsilon_{\text{Sad}}(\overline{x}\big|X_1, X_2, \Phi), \tag{53}$$

where $\widehat{y}^0 (= y^0(\widehat{x}^1))$ *is the "*y^0*-component" of* $\widehat{x}^1$*;*

As a corollary, under the premise of Proposition 4, when applying to the saddle point problem (48b) the CoMP algorithm induced by the above setup and passing "at no cost" from the approximate solutions $x^t = [x^{1,t}; x^{2,t}]$ generated by CoMP to the corrections $\widehat{x}^{1,t}$ of $x^{1,t}$'s, we get feasible solutions to the problem of interest (47a) satisfying the error bound

$$f(y^0(\widehat{x}^{1,t})) - \text{Opt} \le \frac{\Theta[x^1_* \times X_2]L}{t}, \ t = 1, 2, \ldots \tag{54}$$

where L is the Lipschitz constant of $F_u(\cdot)$ induced by the norm $\|\cdot\|$ given by (51), and $\Theta[\cdot]$ is induced by the d.g.f. given by the same (51) and the $u = [y^0; \ldots; y^K; z^0; \ldots; z^K; w^1; \ldots; w^K]$ -component of the starting point. Note that W_k and Z_k are compact, whence $\Theta[x^1_* \times X_2]$ is finite.

Remark In principle, we can use the result of Proposition 4 "as is", that is, to work from the very beginning with values of ρ_k satisfying (52); this option is feasible, provided that we know in advance the corresponding Lipschitz constants and they are not too large (which indeed is the case in some applications). This being said, when our objective is to ensure the validity of the bound (53), selecting ρ_k's according to (52) could be very conservative. From our experience, usually it is better to adjust the penalization coefficients ρ_k on-line. Specifically, let $\overline{\Phi}(\overline{x}^1) = \sup_{x^2 \in X_2} \Phi(\overline{x}^1, x^2)$ (cf (15)). We always have $\widehat{\text{Opt}} \le \text{Opt}$. It follows that independently of how ρ_k are selected, we have

$$f(\widehat{y}^0) - \text{Opt} \le \underbrace{[f(\widehat{y}^0) - \overline{\Phi}(\overline{x}^1)]}_{\epsilon_1} + \underbrace{\left[\overline{\Phi}(\overline{x}^1) - \widehat{\text{Opt}}\right]}_{\epsilon_2} \tag{55}$$

for every feasible solution $\overline{x}^1 = \{[\overline{y}^k; \overline{\tau}^k]\}_{k=0}^K$ to (48b) and the same inequality holds for its correction $\widehat{x}^1 = \left\{[\widehat{y}^k; \widehat{\tau}^k]\right\}_{k=0}^K$. When $\overline{x}^1$ is a component of a good (with small ϵ_{Sad}) approximate solution to the saddle point problem (48b), ϵ_2 is small. If ϵ_1 also is small, we are done; otherwise we can either increase in a fixed ratio the current values of all ρ_k, or only of those ρ_k for which passing from $[\overline{y}^k; \overline{\tau}^k]$ to $[\widehat{y}^k; \widehat{\tau}^k]$ results in "significant" quantities

$$[\psi_k(\widehat{y}^k) + \widehat{\tau}^k] - [\psi_k(\overline{y}^k) + \overline{\tau}^k + \rho_k \pi^*_k(\overline{y}^k - A_k \overline{y}^0 - b_k)]$$

and solve the updated saddle point problem (48b).

4.4 Numerical illustrations

4.4.1 Matrix completion

Problem of interest In the experiments to be reported, we applied the just outlined approach to Example 1, that is, to the problem

$$\mathrm{Opt} = \min_{y^0 \in \mathbf{R}^{n\times n}} \left[\upsilon(y^0) = \underbrace{\frac{1}{2}\|P_\Omega y^0 - b\|_2^2}_{\psi_0(y^0)} + \underbrace{\lambda\|y^0\|_1}_{\Psi_0(y^0)} + \underbrace{\mu\|y^0\|_{\mathrm{nuc}}}_{\Psi_1(y^0)} \right]. \tag{56}$$

where Ω is a given set of cells in an $n \times n$ matrix, and $P_\Omega y$ is the restriction of $y \in \mathbf{R}^{n\times n}$ onto Ω; this restriction is treated as a vector from $\mathbf{R}^M$, $M = \mathrm{Card}(\Omega)$. Thus, (56) is a kind of matrix completion problem where we want to recover a sparse and low rank $n \times n$ matrix given noisy observations b of its entries in cells from Ω. Note that (56) is a special case of (47b) with $K = 1$, $Y_0 = Y_1 = E_0 = E_1 = \mathbf{R}^{n\times n}$, the identity mapping $y^0 \mapsto A_1 y^0$, and $\phi_0(y^0, z^0) \equiv \psi_0(y^0)$, $\phi_1 \equiv 0$ (so that Z_k can be defined as singletons, and $\overline{\Psi}_k(\cdot)$ set to 0, $k = 0, 1$).

Implementing the CoMP algorithm When implementing the CoMP algorithm, we used the Frobenius norm $\|\cdot\|_F$ on $\mathbf{R}^{n\times n}$ in the role of $p_0(\cdot)$, $p_1(\cdot)$ and $\pi_1(\cdot)$, and the function $\frac{1}{2}\|\cdot\|_F^2$ in the role of d.g.f.'s $\omega_0(\cdot)$, $\omega_1(\cdot)$, $\widehat{\omega}_1(\cdot)$.

The aggregation weights in (51) were chosen as $\alpha_0 = \alpha_1 = 1/D$ and $\gamma_1 = 1$, where D is a *guess* of the quantity $D_* := \|y_*^0\|_F$, where y_*^0 is the optimal solution (56). With $D = D_*$, our aggregation would roughly optimize the right hand side in (54), provided the starting point is the origin.

The coefficient ρ_1 in (48b) was adjusted dynamically as explained at the end of Sect. 4.3. Specifically, we start with a small (0.001) value of ρ_1 and restart the solution process, increasing by factor 3 the previous value of ρ_1, each time when the x^1-component $\overline{x}$ of current approximate solution and its correction $\widehat{x}$ violate the inequality $\upsilon(y^0(\widehat{x})) \le (1+\kappa)\overline{\Phi}(\overline{x})$ for some small tolerance κ (we used κ=1.e−4), cf. (55).

The stepsizes γ_t in the CoMP algorithm were adjusted dynamically, specifically, as follows. At a step τ, given a current guess γ for the stepsize, we set $\gamma_\tau = \gamma$, perform the step and check whether $\delta_\tau \le 0$. If this is the case, we pass to step $\tau + 1$, the new guess for the stepsize being 1.2 times the old one. If δ_τ is positive, we decrease γ_τ in a fixed proportion (in our implementation—by factor 0.8), repeat the step, and proceed in this fashion until the resulting value of δ_τ becomes nonpositive. When it happens, we pass to step $\tau + 1$, and use the value of γ_τ we have ended up with as our new guess for the stepsize.

In all our experiments, the starting point was given by the matrix $\widehat{y} := P_\Omega^* b$ ("observations of entries in cells from Ω and zeros in all other cells") according to $y^0 = y^1 = \widehat{y}$, $\tau^0 = \lambda\|\widehat{y}\|_1$, $\tau^1 = \mu\|\widehat{y}\|_{\mathrm{nuc}}$, $w^1 = 0$.

Lower bounding the optimal value When running the CoMP algorithm, we at every step t have at our disposal an approximate solution $y^{0,t}$ to the problem of interest (59); $y^{0,t}$ is nothing but the y^0-component of the approximate solution x^t generated by CoMP as applied to the saddle point approximation of (59) corresponding to the current value of ρ_1, see (49). We have at our disposal also the value $\upsilon(y^{0,t})$ of the objective of (56) at $y^{0,t}$; this quantity is a byproduct of checking whether we should update the current value of ρ_1.[4] As a result, we have at our disposal the best found so far value $\upsilon^t = \min_{1\le\tau\le t}\upsilon(y^{0,\tau})$, along with the corresponding value $y_*^{0,t}$ of y^0: $\upsilon(y_*^{0,t}) = \upsilon^t$. In order to understand how good is the best generated so far approximate solution $y_*^{0,t}$ to the problem of interest, we need to upper bound the quantity $\upsilon^t - \mathrm{Opt}$, or, which is the same, to lower bound Opt. This is a nontrivial task, since the domain of the problem of interest is unbounded, while the usual techniques for online bounding from below the optimal value in a convex minimization problem require the domain to be bounded. We are about to describe a technique for lower bounding Opt utilizing the structure of (56).

Let y_*^0 be an optimal solution to (56) (it clearly exists since $\psi_0 \ge 0$ and $\lambda, \mu > 0$). Assume that at a step t we have at our disposal an upper bound $R = R_t$ on $\|y_*^0\|_1$, and let

$$R^+ = \max\left[R, \|y^{0,t}\|_1\right].$$

Let us look at the saddle point approximation of the problem of interest

$$\begin{aligned}
\widehat{\mathrm{Opt}} &= \min_{x^1=[y^0;\tau^0;y^1;\tau^1]\in\widehat{X}_1}\ \max_{x^2\in X_2}\left[\Phi(x^1,x^2) := \psi_0(y^0)+\tau^0+\tau^1+\rho_1\langle y^1-y^0, x^2\rangle\right],\\
X_1 &= \left\{[y^0;\tau^0;y^1;\tau^1] : \tau^0 \ge \lambda\|y^0\|_1, \tau^1 \ge \mu\|y^1\|_{\mathrm{nuc}}\right\},\\
X_2 &= \left\{x^2 : \|x^2\|_F \le 1\right\}.
\end{aligned} \tag{57}$$

associated with current value of ρ_1, and let

$$\widehat{X}_1 = \left\{[y^0;\tau^0;y^1;\tau^1] \in X_1 : \tau^0 \le \lambda R^+, \tau^1 \le \mu R^+\right\}.$$

Observe that the point $x^{1,*} = [y_*^0; \lambda\|y_*^0\|_1; y_*^0; \mu\|y_*^0\|_{\mathrm{nuc}}]$ belongs to $\widehat{X}_1$ (recall that $\|\cdot\|_{\mathrm{nuc}} \le \|\cdot\|_1$) and that

$$\mathrm{Opt} = \upsilon(y_*^0) \ge \overline{\Phi}(x^{1,*}),\ \ \overline{\Phi}(x^1) = \max_{x^2\in X_2}\Phi(x^1,x^2).$$

It follows that

$$\widehat{\mathrm{Opt}} := \min_{x^1\in\widehat{X}_1}\overline{\Phi}(x^1) \le \mathrm{Opt}.$$

[4] With our implementation, we run this test for both search points and approximate solutions generated by the algorithm.

Further, by Proposition 2 as applied to $X_1' = \widehat{X}_1$ and $X_2' = X_2$ we have[5]

$$\overline{\Phi}(x^{1,t}) - \widehat{\mathrm{Opt}} \le \mathrm{Res}(\widehat{X}_1 \times X_2 \big| \mathcal{I}_t, \lambda^t),$$

where $\mathcal{I}_t$ is the execution protocol generated by CoMP *as applied to the saddle point problem* (57) (i.e., since the last restart preceding step t till this step), and λ^t is the associated accuracy certificate. We conclude that

$$\ell_t := \overline{\Phi}(x^{1,t}) - \mathrm{Res}(\widehat{X}_1 \times X_2 \big| \mathcal{I}_t, \lambda^t) \le \widehat{\mathrm{Opt}} \le \mathrm{Opt},$$

and ℓ_t is easy to compute (since the resolution is just the maximum of a readily given by $\mathcal{I}_t, \lambda^t$ affine function over $\widehat{X}_1 \times X_2$). Setting $\upsilon_t = \max_{\tau \le t} \ell_\tau$, we get nondecreasing with t lower bounds on Opt. Note that this component of our lower bounding is independent of the particular structure of ψ_0.

It remains to explain how to get an upper bound R on $\|y_*^0\|_1$, and this is where the special structure of $\psi_0(y) = \frac{1}{2}\|P_\Omega y - b\|_2^2$ is used. Recalling that $b \in \mathbf{R}^M$, let us set

$$\vartheta(r) = \min_{v \in \mathbf{R}^M} \left\{ \frac{1}{2}\|v - b\|_2^2 : \|v\|_1 \le r \right\}, \ r \ge 0,$$

It is immediately seen that replacing the entries in b by their magnitudes, $\vartheta(\cdot)$ remains intact, and that for $b \ge 0$ we have

$$\vartheta(r) = \min_{v \in \mathbf{R}^M} \left\{ \frac{1}{2}\|v - b\|_2^2 : v \ge 0, \sum_i v_i \le r \right\},$$

so that $\vartheta(\cdot)$ is an easy to compute nonnegative and nonincreasing convex function of $r \ge 0$. Now, by definition of P_Ω, the function $\vartheta^+(\|y^0\|_1)$ where

$$\vartheta^+(r) = \lambda r + \vartheta(r)$$

is a lower bound on $\upsilon(y^0)$. As a result, given an upper bound υ^t on $\mathrm{Opt} = \upsilon(y_*)$, the easy-to-compute quantity

$$R_t := \max\left\{ r : \vartheta^+(r) \le \upsilon^t \right\}$$

is an upper bound on $\|y_*^0\|_1$. Since υ^t is nonincreasing in t, R_t is nonincreasing in t as well.

Generating the data In the experiments to be reported, the data of (56) were generated as follows. Given n, we build "true" $n \times n$ matrix $y_\# = \sum_{i=1}^k e_i f_i^T$, with $k = \lfloor n/4 \rfloor$ and vectors $e_i, f_i \in \mathbf{R}^n$ sampled, independently of each other, as follows: we draw a vector from the standard Gaussian distribution $\mathcal{N}(0, I_n)$, and then zero out part of

[5] Note that the latter relation implies that what was denoted by $\widetilde{\Phi}$ in Proposition 2 is nothing but $\overline{\Phi}$.

the entries, with probability of replacing a particular entry with zero selected in such a way that the sparsity of $y_\#$ is about a desired level (in our experiments, we wanted $y_\#$ to have about 10% of nonzero entries). The set Ω of "observed cells" was built at random, with probability 0.25 for a particular cell to be in Ω. Finally, b was generated as $P_\Omega(y_\# + \sigma\xi)$, where the entries of $\xi \in \mathbf{R}^{n\times n}$ were independently of each other drawn from the standard Gaussian distribution, and

$$\sigma = 0.1\frac{\sum_{i,j}|[y_\#]_{ij}|}{n^2}.$$

We used $\lambda = \mu = 10\sigma$.[6] Finally, our guess for the Frobenius norm of the optimal solution to (56) is defined as follows. Note that the quantity $\|b\|_2^2 - M\sigma^2$ is an estimate of $\|P_\Omega y_\#\|_2^2$. We define the estimate D of $D_* := \|y_*\|_F$ "as if" the optimal solution were $y_\#$, and all entries of $y_\#$ were of the same order of magnitude

$$D = \sqrt{\frac{n^2}{M}\max[\|b\|_2^2 - M\sigma^2, 1]}, \quad M = \mathrm{Card}(\Omega).$$

Numerical results The results of the first series of experiments are presented in Table 1. The comments are as follows.

In the "small" experiment ($n = 128$, the largest n where we were able to solve (56) in a reasonable time by `CVX` [13] using the state-of-the-art `mosek` [1] Interior-Point solver and thus knew the "exact" optimal value), CoMP exhibited fast convergence: relative accuracies 1.1e−3 and 6.2e−6 are achieved in 64 and 4,096 steps (1.2 and 74.9 s, respectively, as compared to 4,756.7 s taken by `CVX`).

In larger experiments ($n = 512$ and $n = 1,024$, meaning design dimensions 262,144 and 1,048,576, respectively), the running times look moderate, and the convergence pattern of the CoMP still looks promising.[7] Note that our lower bounding, while somehow working, is very conservative: it overestimates the "optimality gap" $\upsilon^t - \upsilon_t$ by 2–3 orders of magnitude for moderate and large values of t in the 128×128 experiment. More accurate performance evaluation would require a less conservative lower bounding of the optimal value (as of now, we are not aware of any alternative).

In the second series of experiments, the data of (56) were generated in such a way that the true optimal solution and optimal value to the problem were known from the very beginning. To this end we take as Ω the collection of all cells of an $n \times n$ matrix, which, via optimality conditions, allows to select b making our "true" matrix $y_\#$ the optimal solution to (56). The results are presented in Table 2.

In the third series of experiments, we compared our algorithm with the basic version of ADMM as presented in [5]; this version is capable to handle straightforwardly the

[6] If the goal of solving (56) were to recover $y_\#$, our λ and μ would, perhaps, be too large. Our goal, however, was solving (56) as an "optimization beast," and we were interested in "meaningful" contribution of Ψ_0 and Ψ_1 to the objective of the problem, and thus in not too small λ and μ.

[7] Recall that we do not expect linear convergence, just $O(1/t)$ one.

Table 1 Composite Mirror Prox algorithm on problem (56) with $n \times n$ matrices

t	8	16	32	64	128	256	512	1,024	2,048	4,096
(a) $n = 128$, Opt $= 13.28797$ (CVX CPU 4756.7 s)										
CPU (s)	0.1	0.3	0.6	1.2	2.3	4.7	9.4	18.7	37.5	74.9
$\upsilon^t - \text{Opt}$	2.0e−2	1.8e−2	1.8e−2	1.4e−2	5.3e−3	5.0e−3	1.3e−3	7.8e−4	3.2e−4	8.3e−5
$\upsilon^t - \upsilon_t$	4.8e0	4.5e0	4.2e0	3.7e0	2.1e0	6.3e−1	2.1e−1	1.3e−1	6.0e−2	3.4e−2
$\frac{\upsilon^t - \text{Opt}}{\text{Opt}}$	1.5e−3	1.3e−3	1.3e−3	1.1e−3	4.0e−4	3.7e−4	9.5e−5	5.8e−5	2.4e−5	6.2e−6
$\frac{\upsilon^t - \upsilon_t}{\upsilon_{4,096}}$	3.6e−1	3.4e−1	3.2e−1	2.8e−1	1.5e−1	4.7e−2	1.6e−2	9.4e−3	4.5e−3	2.6e−3
$\frac{\upsilon^1 - \text{Opt}}{\upsilon^t - \text{Opt}}$	4.8e1	5.4e1	5.4e1	6.7e1	1.8e2	1.9e2	7.5e2	1.2e3	2.9e3	1.1e4
$\frac{\upsilon^1 - \upsilon_1}{\upsilon^t - \upsilon_t}$	3.0e0	3.2e0	3.7e0	3.9e0	6.9e0	2.3e1	6.7e1	1.1e2	2.4e2	4.1e2
(b) $n = 512$, $\upsilon_{2,048} = 175.445 \le \text{Opt} \le \upsilon^{2,048} = 180.503$ (CVX not tested)										
CPU (s)	3.7	7.5	15.0	29.9	59.8	119.6	239.2	478.4	992.0	
$\upsilon^t - \upsilon_t$	4.4e1	4.4e1	4.3e1	4.2e1	4.1e1	3.7e1	2.3e1	1.2e1	5.1e0	
$\frac{\upsilon^t - \upsilon_t}{\upsilon_{1,024}}$	2.4e−1	2.4e−1	2.4e−1	2.4e−1	2.2e−1	2.0e−1	1.3-1	6.4e−2	2.8e−2	
$\frac{\upsilon^1 - \upsilon_1}{\upsilon^t - \upsilon_t}$	4.4e0	4.4e0	4.5e0	4.6e0	4.8e0	5.5e0	8.5e0	1.7e1	3.8e1	
(c) $n = 1,024$, $\upsilon_{1,024} = 655.422 \le \text{Opt} \le \upsilon^{1,024} = 660.786$ (CVX not tested)										
CPU (s)	23.5	46.9	93.8	187.6	375.3	750.6	1501.2	3,002.3		
$\upsilon^t - \upsilon_t$	1.5e2	1.5e2	1.3e2	1.2e2	1.1e2	8.0e1	1.6e1	5.4e0		
$\frac{\upsilon^t - \upsilon_t}{\upsilon_{1,024}}$	2.4e−1	2.2e−1	2.2e−1	1.9e−1	1.7e−01	1.2e−1	2.4e−2	8.1e−3		
$\frac{\upsilon^1 - \upsilon_1}{\upsilon^t - \upsilon_t}$	4.6e0	4.8e0	5.3e0	5.7e0	6.3e0	8.9e0	4.5e1	1.3e2		

υ^t are the best values of $\upsilon(\cdot)$, and υ_t are lower bounds on the optimal value found in course of t steps. Platform: `MATLAB` on 3.40 GHz Intel Core i7-3770 desktop with 16 GB RAM, 64 bit Windows 7

Table 2 Composite Mirror Prox algorithm on problem (56) with $n \times n$ matrices and known optimal value Opt

t	1	7	8	12	128	256	512	1,024
(a) $n = 512$, Opt $= 607.9854$								
CPU (s)	1.3	8.3	9.3	11.0	65.9	125.0	244.7	486.0
$\upsilon^t - \text{Opt}$	92.9	1.58	0.30	0.110	0.095	0.076	0.069	0.069
$\upsilon^t - \upsilon_t$	700.9	92.4	69.5	54.6	52.8	44.2	21.2	3.07
$\frac{\upsilon^t - \text{Opt}}{\text{Opt}}$	0.153	2.6e−3	5.0e−4	1.8e−4	1.6e−4	1.3e−4	1.1e−4	1.1e−4
$\frac{\upsilon^t - \upsilon_t}{\text{Opt}}$	1.153	0.152	0.114	0.090	0.087	0.073	0.035	0.005
(b) $n = 1{,}024$, Opt=2,401.168								
CPU (s)	8.9	48.1	51.9	392.7	752.1	1,464.9		
$\upsilon^t - \text{Opt}$	371.4	3.48	0.21	0.21	0.19	0.16		
$\upsilon^t - \upsilon_t$	2772	241.7	201.2	147.3	146.5	122.9		
$\frac{\upsilon^t - \text{Opt}}{\text{Opt}}$	0.154	1.5e−3	9e−5	9e−5	8e−5	7e−5		
$\frac{\upsilon^t - \upsilon_t}{\text{Opt}}$	1.155	0.101	0.084	0.061	0.061	0.051		

υ^t are the best values of $\upsilon(\cdot)$, and υ_t are lower bounds on the optimal value found in course of t steps. Platform: MATLAB on 3.40 GHz Intel Core i7-3770 desktop with 16 GB RAM, 64 bit Windows 7

Table 3 Number of steps and CPU time for Composite Mirror Prox algorithm and ADMM algorithm to achieve relative error $\epsilon = 10^{-4}$ on problem (56)

$n \times n$	Composite Mirror Prox		ADMM	
	Step	CPU (s)	Step	CPU (s)
128×128	34	0.77	11	0.13
256×256	94	8.02	9	0.37
512×512	38	15.06	9	1.42
1024×1024	34	81.76	8	8.74

Platform: MATLAB on Intel i5-2400S @2.5GHz CPU with 4GB RAM, 64-bit Windows 7

matrix completion with noisy observations of part of the entries.[8] The data in these experiments were generated in the same way as in the aforementioned experiments with known optimal solutions. The results are presented in Table 3. We see that ADMM is essentially faster than our algorithm, suggesting that ADMM, *when applicable in its basic form*, typically outperforms CoMP. However, this is not the case when ADMM is not directly applicable; we consider one example of the sort in the next section.

[8] Note that in a more complicated matrix recovery problem, where noisy linear combinations of the matrix entries rather than just some of these entries are observed, applying ADMM becomes somehow problematic, whilethe proposed algorithm still is applicable "as is."

It should be mentioned that in these experiments the value of ρ_1 resulting in negligibly small, as compared to ϵ_2, values of ϵ_1 in (55) was found in the first 10–30 steps of the algorithm, with no restarts afterwards.

Remark For the sake of simplicity, so far we were considering problem (56), where minimization is carried out over y^0 running through the entire space $\mathbf{R}^{n\times n}$ of $n \times n$ matrices. What happens if we restrict y^0 to reside in a given closed convex domain Y_0?

It is immediately seen that the construction we have presented can be straightforwardly modified for the cases when Y_0 is a centered at the origin ball of the Frobenius or $\|\cdot\|_1$ norm, or the intersection of such a set with the space of symmetric $n \times n$ matrices. We could also handle the case when Y_0 is the centered at the origin nuclear norm ball (or intersection of this ball with the space of symmetric matrices, or with the cone of positive semidefinite symmetric matrices), but to this end one needs to "swap the penalties"—to write the representation (47c) of problem (56) as

$$\begin{aligned}
&\min_{\substack{\{y^k;\tau^k]\}_{k=0}^1\\ \in Y_0^+\times Y_1^+}} \left\{\Upsilon(y^0,y^1,\tau^0,\tau^1) := \underbrace{\frac{1}{2}\|P_\Omega y^0 - b\|_2^2}_{\psi_0(y^0)} + \tau^0 + \tau^1 : y^0 = y^1\right\},\\
&Y_0^+ = \{[y^0;\tau^0] : y^0 \in Y_0, \tau^0 \geq \mu\|y^0\|_{\rm nuc}\},\\
&Y_1^+ = \{[y^1;\tau^1] : y^1 \in Y_1, \tau^1 \geq \lambda\|y^1\|_1\},
\end{aligned}$$

where $Y_1 \supset Y_0$ "fits" $\|\cdot\|_1$ (meaning that we can point out a d.g.f. $\omega_1(\cdot)$ for Y_1 which, taken along with $\Psi_1(y^1) = \lambda\|y^1\|_1$, results in easy-to-solve auxiliary problems (45)). We can take, e.g. $\omega_1(y^1) = \frac{1}{2}\|y^1\|_F^2$ and define Y_1 as the entire space, or a centered at the origin Frobenius/$\|\cdot\|_1$ norm ball large enough to contain Y_0.

4.4.2 Image decomposition

Problem of interest In the experiments to be reported, we applied the just outlined approach to Example 2, that is, to the problem

$$\begin{aligned}
{\rm Opt} = \min_{y^1,y^2,y^3\in\mathbf{R}^{n\times n}} \Big\{&\|A(y^1+y^2+y^3) - b\|_2 + \mu_1\|y^1\|_{\rm nuc}\\
&+ \mu_2\|y^2\|_1 + \mu_3\|y^3\|_{\rm TV}\Big\},
\end{aligned} \tag{58}$$

where $A(y) : \mathbf{R}^{n\times n} \to \mathbf{R}^M$ is a given linear mapping.

Problem reformulation We first rewrite (58) as a saddle point optimization problem

$$\begin{aligned}
{\rm Opt} = \min_{y^1,y^2,y^3\in\mathbf{R}^{n\times n}} \Big\{&\|A(y^1+y^2+y^3) - b\|_2 + \mu_1\|y^1\|_{\rm nuc}\\
&+ \mu_2\|y^2\|_1 + \mu_3\|Ty^3\|_1\Big\}
\end{aligned}$$

$$= \min_{y^1,y^2,y^3\in\mathbf{R}^{n\times n}} \left\{ \max_{\|z\|_2\le 1} \langle z,\, A(y^1+y^2+y^3)-b\rangle + \quad \mu_1\|y^1\|_{\mathrm{nuc}} + \mu_2\|y^2\|_1 + \mu_3\|Ty^3\|_1 \right\}, \tag{59}$$

where $T : \mathbf{R}^{n\times n} \to \mathbf{R}^{2n(n-1)}$ is the mapping $y \mapsto Ty = \begin{bmatrix} \{(\nabla_i y)_{n(j-1)+i}\}_{i=1,\dots,n-1,\ j=1,\dots,n} \\ \{(\nabla_j y)_{n(i-1)+j}\}_{i=1,\dots,n,\ j=1,\dots,n-1} \end{bmatrix}$.

Next we rewrite (59) as a linearly constrained saddle-point problem with "simple" penalties:

$$\mathrm{Opt} = \min_{\substack{y^3\in Y_3 \\ [y^k;\tau_k]\in Y_k^+,\ 0\le k\le 2}} \max_{z\in Z} \left\{ \langle z,\, A(y^1+y^2+y^3)-b\rangle + \tau_1+\tau_2+\tau_0,\ y^0 = Ty^3 \right\},$$

where

$$\begin{aligned} Y_0^+ &= \left\{ [y^0;\tau_0] :\ y^0 \in Y_0 = \mathbf{R}^{2n(n-1)} : \|y^0\|_1 \le \tau_0/\mu_3 \right\}, \\ Y_1^+ &= \left\{ [y^1;\tau_1] :\ y^1 \in Y_1 = \mathbf{R}^{n\times n} : \|y^1\|_{\mathrm{nuc}} \le \tau_1/\mu_1 \right\}, \\ Y_2^+ &= \left\{ [y^2;\tau_2] :\ y^2 \in Y_2 = \mathbf{R}^{n\times n} : \|y^2\|_1 \le \tau_2/\mu_2 \right\} \\ Y_3 &= \mathbf{R}^{n\times n}, \quad Z = \left\{ z\in\mathbf{R}^M : \|z\|_2 \le 1 \right\}, \end{aligned}$$

and further approximate the resulting problem with its penalized version:

$$\widehat{\mathrm{Opt}} = \min_{\substack{y^3\in Y_3 \\ [y^k;\tau_k]\in Y_k^+,\ 0\le k\le 2}} \max_{\substack{z\in Z \\ w\in W}} \left\{ \begin{array}{l} \langle z,\, A(y^1+y^2+y^3)-b\rangle \\ +\tau_1+\tau_2+\tau_0+\rho\langle w,\, y^0 - Ty^3\rangle \end{array} \right\}, \tag{60}$$

with

$$W = \left\{ w\in\mathbf{R}^{2n(n-1)},\ \|w\|_2 \le 1 \right\}.$$

Note that the function $\psi(y^1,y^2,y^3) := \|A(y^1+y^2+y^3)-b\|_2 = \max_{\|z\|_2\le 1} \langle z,\, A(y^1+y^2+y^3)-b\rangle$ is Lipschitz continuous in y^3 with respect to the Euclidean norm on $\mathbf{R}^{n\times n}$ with corresponding Lipschitz constant $G = \|A\|_{2,2}$, which is the spectral norm (the principal singular value) of A. Further, $\Psi(y^0) = \mu_3\|y^0\|_1$ is Lipschitz-continuous in y^0 with respect to the Euclidean norm on $\mathbf{R}^{2n(n-1)}$ with the Lipschitz constant $H \le \mu_3\sqrt{2n(n-1)}$. With the help of the result of Proposition 4 we conclude that to ensure the "exact penalty" property it suffices to choose $\rho \ge \|A\|_{2,2} + \mu_3\sqrt{2n(n-1)}$. Let us denote

$$U = \left\{ \begin{array}{l} u = [y^0;\dots;y^3;z;w] :\ y^k\in Y^k,\ 0\le k\le 3, \\ z\in\mathbf{R}^M,\ \|z\|_2\le 1,\ w\in\mathbf{R}^{2n(n-1)},\ \|w\|_2\le 1 \end{array} \right\}.$$

Table 4 Composite Mirror Prox algorithm on problem (58) with $n \times n$ matrices

t	8	16	32	64	128	256	512	1,024	2,048
(a) $n = 64$, Opt $= 15.543$ (CVX CPU 4525.5 sec)									
CPU (s)	0.1	0.2	0.4	0.8	1.6	3.1	6.3	12.6	25.2
$\upsilon_t - \upsilon_{2,048}$	1.5e1	2.8e0	6.2e−1	2.3e−1	1.1e−1	4.2e−2	1.5e−2	4.4e−3	0.0e0
$\frac{\upsilon_t - \upsilon_{2,048}}{\upsilon_{2,048}}$	9.5e−1	1.8e−1	4.0e−2	1.5e−2	7.0e−3	2.7e−3	9.9e−4	2.8e−4	0.0e0
$\upsilon_t - \text{Opt}$	1.5e1	2.8e0	6.2e−1	2.3e−1	1.1e−1	4.5e−2	1.8e−2	6.6e−3	2.2e−3
$\frac{\upsilon_t - \text{Opt}}{\text{Opt}}$	9.5e−1	1.8e−1	4.0e−2	1.5e−2	7.1e−3	2.9e−3	1.1e−3	4.2e−4	1.4e−4
(b) $n = 512$ (CVX not tested)									
CPU (s)	6.2	12.3	24.7	49.3	98.6	197.2	394.4	788.9	1,577.8
$\upsilon_t - \upsilon_{2,048}$	1.1e2	5.8e1	2.7e1	1.3e1	6.2e0	2.9e0	1.2e0	3.9e−1	0.0e0
$\frac{\upsilon_t - \upsilon_{2,048}}{\upsilon_{2,048}}$	9.0e−1	4.9e−1	2.3e−1	1.1e−1	5.2e−2	2.5e−2	1.0e−2	3.3e−3	0.0e0

υ^t are the best values of $\upsilon(\cdot)$ in course of t steps. Platform: `MATLAB` on Intel i5-2400S @2.5GHz CPU with 4GB RAM, 64-bit Windows 7

We equip the embedding space E_u of U with the norm

$$\|u\| = \left(\alpha_0 \|y^0\|_2^2 + \sum_{k=1}^{3} \alpha_k \|y^k\|_2^2 + \beta \|z\|_2^2 + \gamma \|w\|_2^2 \right)^{1/2},$$

and U with the proximal setup $(\|\cdot\|, \omega(\cdot))$ with

$$\omega(u) = \frac{\alpha_0}{2} \|y^0\|_2^2 + \sum_{k=1}^{3} \frac{\alpha_k}{2} \|y^k\|_2^2 + \frac{\beta}{2} \|z\|_2^2 + \frac{\gamma}{2} \|w\|_2^2.$$

Implementing the CoMP algorithm When implementing the CoMP algorithm, we use the above proximal setup with adaptive aggregation parameters $\alpha_0 = \cdots = \alpha_4 = 1/D^2$ where D is our guess for the upper bound of $||y_*||_2$, that is, whenever the norm of the current solution exceeds 20 % of the guess value, we increase D by factor 2 and update the scales accordingly. The penalty ρ and stepsizes γ_t are adjusted dynamically the same way as explained in the last experiment.

Numerical results In the first series of experiments, we build the $n \times n$ observation matrix b by first generating a random matrix with rank $r = \lfloor \sqrt{n} \rfloor$ and another random matrix with sparsity $p = 0.01$, so that the observation matrix is a sum of these two matrices and of random noise of level $\sigma = 0.01$; we take $y \mapsto Ay$ as the identity mapping. We use $\mu_1 = 10\sigma, \mu_2 = \sigma, \mu_3 = \sigma$. The very preliminary results of this series of experiments are presented in Table 4. Note that unlike the matrix completion problem, discussed in Sect. 4.4.1, here we are not able to generate the problem with known optimal solutions. Better performance evaluation would require good lower

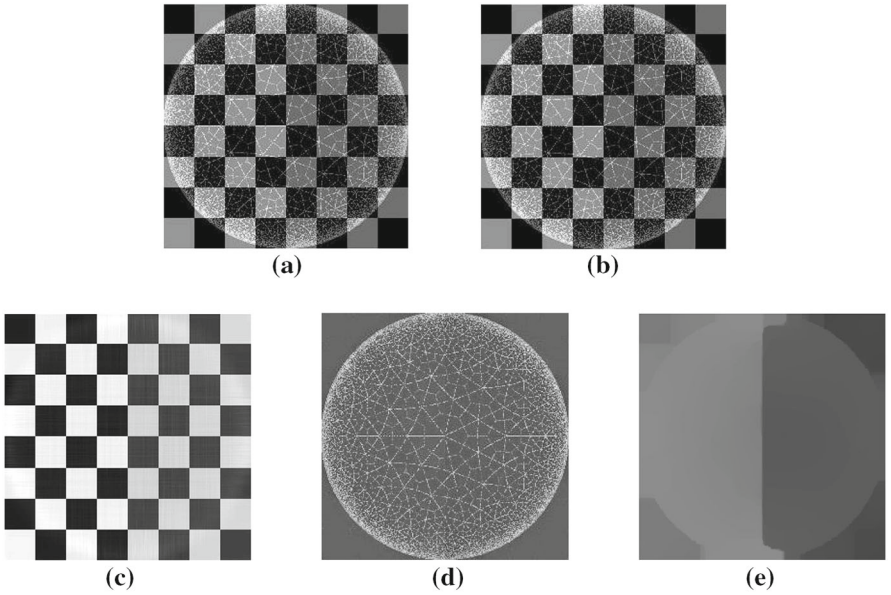

Fig. 1 Observed and reconstructed images (size 256 × 256), **a** observation b, **b** recovery $y_1 + y_2 + y_3$, **c** low-rank component, **d** sparse component, **e** smooth component

bounding of the true optimal value, which is however problematic due to unbounded problem domain.

In the second series of experiments, we implement the CoMP algorithm to decompose real images and extract the underlying low rank/sparse singular distortion/smooth background components. The purpose of these experiments is to illustrate how the algorithm performs with the choice of small regularization parameters which is meaningful from the point of view of applications to image recovery. Image decomposition results for two images are provided on Figs. 1 and 2. On Fig. 1, we present the decomposition of the observed image of size 256 × 256. We apply the model (59) with regularization parameters $\mu_1 = 0.03, \mu_2 = 0.001, \mu_3 = 0.005$. We run 2,000 iterations of CoMP (total of 393.5 s MATLAB, Intel i5-2400S@2.5GHz CPU). The first component y_1 has approximate rank ≈ 1; the relative reconstruction error is $\|y_1 + y_2 + y_3 - b\|_2/\|b\|_2 \approx 2.8 \times 10^{-4}$. Figure 2 shows the decomposition of the observed image of size 480 × 640 after 1,000 iterations of CoMP (total of 873.6 sec). The regularization parameters of the problem (58) were set to $\mu_1 = 0.06, \mu_2 = 0.002, \mu_3 = 0.005$. The relative reconstruction error is $\|y_1 + y_2 + y_3 - b\|_2/\|b\|_2 \approx 8.4 \times 10^{-3}$.

In the third series of experiments, we compare the CoMP algorithm with some other first-order methods. To the best of our knowledge, a quite limited set of known methods are readily applicable to problems of the form (58), where the "observation-fitting" component in the objective is nonsmooth and the penalty terms involve different components of the observed image. As a result, we compared CoMP to just two alternatives. The first, below referred to as smoothing-APG, applies Nesterov's smoothing tech-

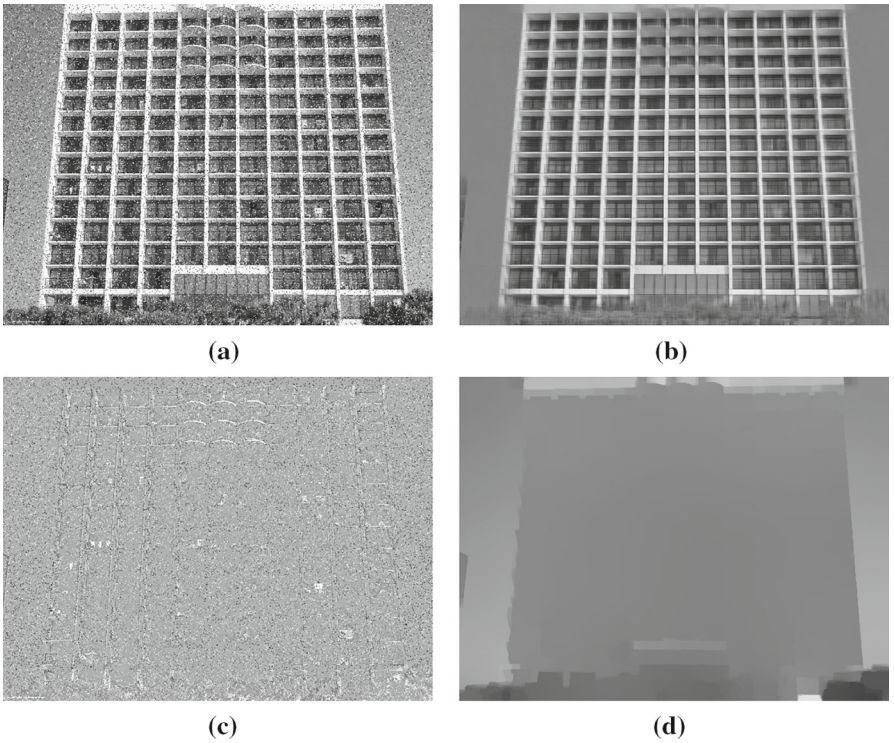

Fig. 2 Observed and decomposed images (size 480 × 640) **a** observation b, **b** low-rank component, **c** sparse component, **d** smooth component

niques to both the first $\|\cdot\|_2$ term and the total variation term in the objective of (58) and then uses the Accelerated Proximal Gradient method (see [20,21] for details) to solve the resulting problem which takes the form

$$\min_{y^1,y^2,y^3\in\mathbf{R}^{m\times n}} \left\{ f_{\rho_1}(y^1, y^2, y^3) + \mu_1\|y^1\|_{\text{nuc}} + \mu_2\|y^2\|_1 + f_{\rho_2}(y^3) \right\} \tag{61}$$

with

$$f_{\rho_1}(y^1, y^2, y^3) = \max_{z:\|z\|_2\le 1} \left\{ \langle P_\Omega(y^1 + y^2 + y^3) - b, z\rangle - \frac{\rho_1}{2}\|z\|_2^2 \right\}$$
$$f_{\rho_2}(y^3) = \max_{w:\|w\|_\infty\le 1} \left\{ \mu_3\langle Ty^3, w\rangle - \frac{\rho_2}{2}\|w\|_2^2 \right\}$$

where $\rho_1 > 0$, $\rho_2 > 0$. In the experiment, we specified the smoothing parameters as $\rho_1 = \epsilon$, $\rho_2 = \frac{\epsilon}{2(n-1)n}$, $\epsilon = 10^{-3}$.

The second alternative, referred to as smoothing-ADMM, applies smoothing technique to the first term in the objective of (58) and uses the ADMM algorithm to solve the resulting problem

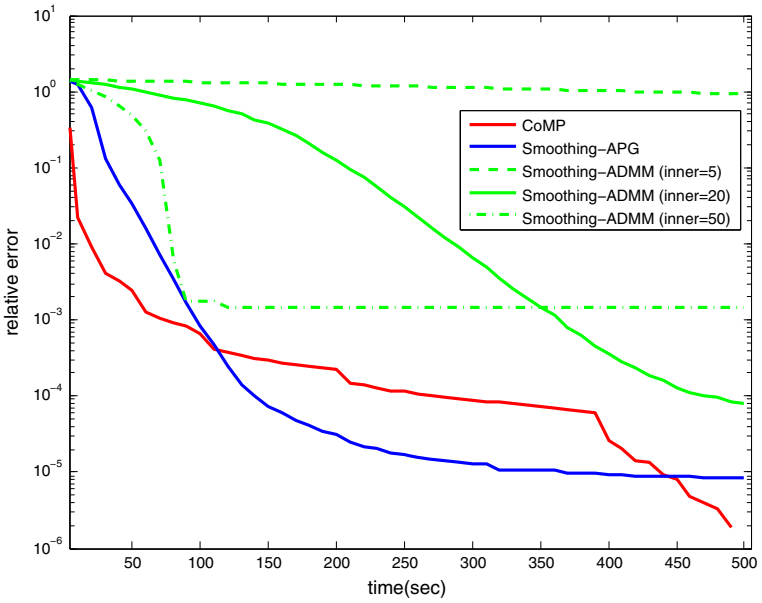

Fig. 3 Comparing CoMP, smoothing-APG, and smoothing-ADMM on problem (58) with 128×128 matrix. x-axis: CPU time; y-axis: relative inaccuracy in terms of the objective. Platform: MATLAB on Intel i5-2400S @2.5GHz CPU with 4GB RAM, 64-bit Windows 7

$$\min_{y^1, y^2, y^3 \in \mathbf{R}^{m \times n}} \left\{ f_{\rho_1}(y^1, y^2, y^3) + \mu_1 \|y^1\|_{\text{nuc}} + \mu_2 \|y^2\|_1 + \mu_3 \|z\|_1 \right\} \quad \text{s.t. } Ty^3 - z = 0 \tag{62}$$

the associated augmented Lagrangian being

$$L_\nu(x = [y^1, y^2, y^3], z; w) = f_{\rho_1}(y^1, y^2, y^3) + \mu_1 \|y^1\|_{\text{nuc}} + \mu_2 \|y^2\|_1 + \mu_3 \|z\|_1 + \langle w, Ty^3 - z \rangle + \frac{\nu}{2} \|Ty^3 - z\|_2^2$$

where $\nu > 0$ is a parameter. The basic version of ADMM would require performing alternatingly $x = (y^1, y^2, y^3)$-updates and z-updates. Since minimizing L_ν in x in a closed analytic form is impossible, we are enforced to perform x-update iteratively and hence inexactly. In our experiment, we used for this purpose the Accelerated Proximal Gradient method, with three implementations differing by the allowed number of inner iterations (5, 20, 50, respectively).

In the experiment, we generated synthetic data in the same fashion as in the first series of experiments and compared the performances of the three algorithms (CoMP and two just described alternatives) by computing accuracies in terms of the objective achieved with a prescribed time budget. The results are presented in Fig. 3. One can see that the performance of ADMM heavily depends on the allowed number of

inner iterations and is not better than the performance of the Accelerated Proximal Gradient algorithm as applied to smooth approximation of the problem of interest. Our algorithm, although not consistently outperforming the Smoothing-APG approach, could still be very competitive, especially when only low accuracy is required.

5 Semi-separable convex problems

5.1 Preliminaries

Our problem of interest in this section is problem (4), (6), namely,

$$\begin{aligned}\mathrm{Opt} &= \min_{[y^1;\ldots;y^K]\in Y_1\times\cdots\times Y_K}\left\{f\left([y^1;\ldots;y^K]\right) := \sum_{k=1}^K[\psi_k(y^k)\right.\\ &\qquad \left.+\Psi_k(y^k)] : \sum_{k=1}^K A_k y^k = b\right\}\\ &= \min_{[y^1;\ldots;y^K]\in Y_1\times\cdots\times Y_K}\left\{\sum_{k=1}^K\left[\psi_k(y^k)+\Psi_k(y^k)\right] :\ g\left([y^1;\ldots;y^K]\right)\le 0\right\},\\ &\qquad g\left([y^1;\ldots;y^K]\right) = \pi^*\left(\sum_{k=1}^K A_k y^k - b\right) = \max_{\pi(w)\le 1}\sum_{k=1}^K\langle A_k y^k - b, w\rangle,\end{aligned} \tag{63}$$

where $\pi(\cdot)$ is some norm and $\pi^*(\cdot)$ is the conjugate norm. A straightforward approach to (63) would be to rewrite it as a saddle point problem

$$\min_{[y^1;\ldots;y^K]\in Y_1\times\cdots\times Y_K}\max_w\left\{\sum_{k=1}^K\left[\psi_k(y^k)+\Psi_k(y^k)\right]+\left\langle\sum_{k=1}^K A_k z^k - b, w\right\rangle\right\} \tag{64}$$

and solve by the Mirror-Prox algorithm from Sect. 3.2 adjusted to work with an unbounded domain U, or, alternatively, we could replace $\max_w$ with $\max_{w:\,\pi(w)\le R}$ with "large enough" R and use the above algorithm "as is." The potential problem with this approach is that if the w-component w^* of the saddle point of (64) is of large π-norm (or "large enough" R is indeed large), the (theoretical) efficiency estimate would be bad since it is proportional to the magnitude of w^* (resp., to R). To circumvent this difficulty, we apply to (63) the sophisticated policy originating from [15]. This policy requires the set $Y = Y_1 \times \cdots \times Y_K$ to be bounded, which we assume below.

Course of actions Note that our problem of interest is of the generic form

$$\mathrm{Opt} = \min_{y\in Y}\{f(y) :\ g(y)\le 0\} \tag{65}$$

where Y is a convex compact set in a Euclidean space E, f and $g : Y \to \mathbf{R}$ are convex and Lipschitz continuous functions. For the time being, we focus on (65) and assume that the problem is feasible and thus solvable.

We intend to solve (65) by the generic algorithm presented in [15]; for our now purposes, the following description of the algorithm will do:

1. The algorithm works in *stages*. Stage $s = 1, 2, ...$ is associated with *working parameter* $\alpha_s \in (0, 1)$. We set $\alpha_1 = \frac{1}{2}$.
2. At stage s, we apply a first order method $\mathcal{B}$ to the problem

$$(P_s) \quad \mathrm{Opt}_s = \min_{y \in Y} \{f_s(y) = \alpha_s f(y) + (1 - \alpha_s) g(y)\} \tag{66}$$

The only property of the algorithm $\mathcal{B}$ which matters here is its ability, when run on (P_s), to produce in course of $t = 1, 2, \ldots$ steps iterates $y_{s,t}$, upper bounds $\overline{f}_s^t$ on Opt_s and lower bounds $\underline{f}_{s,t}$ on Opt_s in such a way that

(a) for every $t = 1, 2, \ldots$, the t-th iterate $y_{s,t}$ of $\mathcal{B}$ as applied to (P_s) belongs to Y;
(b) the upper bounds $\overline{f}_s^t$ are nonincreasing in t (this is "for free") and "are achievable," that is, they are of the form

$$\overline{f}_s^t = f_s(y^{s,t}),$$

where $y^{s,t} \in Y$ is a vector which we have at our disposal at step t of stage s;
(c) the lower bounds $\underline{f}_{s,t}$ should be nondecreasing in t (this again is "for free");
(d) for some nonincreasing sequence $\epsilon_t \to +0$, $t \to \infty$, we should have

$$\overline{f}_s^t - \underline{f}_{s,t} \leq \epsilon_t$$

for all t and s.

Note that since (65) is solvable, we clearly have $\mathrm{Opt}_s \leq \alpha_s \mathrm{Opt}$, implying that the quantity $\underline{f}_{s,t}/\alpha_s$ is a lower bound on Opt. Thus, at step t of stage s we have at our disposal a number of valid lower bounds on Opt; we denote the best (the largest) of these bounds $\underline{\mathrm{Opt}}_{s,t}$, so that

$$\mathrm{Opt} \geq \underline{\mathrm{Opt}}_{s,t} \geq \underline{f}_{s,t}/\alpha_s \tag{67}$$

for all s, t, and $\underline{\mathrm{Opt}}_{s,t}$ is nondecreasing in time.[9]

3. When the First Order oracle is invoked at step t of stage s, we get at our disposal a triple $(y_{s,t} \in Y, f(y_{s,t}), g(y_{s,t}))$. We assume that all these triples are somehow memorized. Thus, after calling First Order oracle at step t of stage s, we have at our disposal a finite set $Q_{s,t}$ on the 2D plane such that for every point $(p, q) \in Q_{s,t}$ we

[9] In what follows, we call a collection $a_{s,t}$ of reals nonincreasing in time, if $a_{s',t'} \leq a_{s,t}$ whenever $s' \geq s$, same as whenever $s = s'$ and $t' \geq t$. "Nondecreasing in time" is defined similarly.

have at our disposal a vector $y_{pq} \in Y$ such that $f(y_{pq}) \le p$ and $g(y_{pq}) \le q$; the set $Q_{s,t}$ (in today terminology, *a filter*) is comprised of all pairs $(f(y_{s',t'}), g(y_{s',t'}))$ generated so far. We set

$$h_{s,t}(\alpha) = \min_{(p,q)\in Q_{s,t}} \Big[\alpha(p - \underline{\mathrm{Opt}}_{s,t}) + (1-\alpha)q\Big] : [0,1] \to \mathbf{R},$$
$$\mathrm{Gap}(s,t) = \max_{0\le\alpha\le 1} h_{s,t}(\alpha). \tag{68}$$

4. Let $\Delta_{s,t} = \{\alpha \in [0,1] : h_{s,t}(\alpha) \ge 0\}$, so that $\Delta_{s,t}$ is a segment in $[0,1]$. Unless we have arrived at $\mathrm{Gap}(s,t) = 0$ (i.e., got an optimal solution to (65), see (69)), $\Delta_{s,t}$ is not a singleton (since otherwise $\mathrm{Gap}(s,t)$ were 0). Observe also that $\Delta_{s,t}$ are nested: $\Delta_{s',t'} \subset \Delta_{s,t}$ whenever $s' \ge s$, same as whenever $s' = s$ and $t' \ge t$.

We continue iterations of stage s while α_s is "well-centered" in $\Delta_{s,t}$, e.g., belongs to the mid-third of the segment. When this condition is violated, we start stage $s+1$, specifying α_{s+1} as the midpoint of $\Delta_{s,t}$.

The properties of the aforementioned routine are summarized in the following statement (cf. [15]).

Proposition 5 (i) *$\mathrm{Gap}(s,t)$ is nonincreasing in time. Furthermore, at step t of stage s, we have at our disposal a solution $\widehat{y}^{s,t} \in Y$ to 65 such that*

$$f(\widehat{y}^{s,t}) \le Opt + \mathrm{Gap}(s,t), \quad \text{and} \quad g(\widehat{y}^{s,t}) \le \mathrm{Gap}(s,t), \tag{69}$$

so that $\widehat{y}^{s,t}$ belongs to the domain Y of problem (65) and is both $\mathrm{Gap}(s,t)$-feasible and $\mathrm{Gap}(s,t)$-optimal.

(ii) *For every $\epsilon > 0$, the number $s(\epsilon)$ of stages until a pair (s,t) with $\mathrm{Gap}(s,t) \le \epsilon$ is found obeys the bound*

$$s(\epsilon) \le \frac{\ln\left(3L\epsilon^{-1}\right)}{\ln(4/3)}, \tag{70}$$

where $L < \infty$ is an a priori upper bound on $\max_{y\in Y} \max[|f(y)|, |g(y)|]$. Besides this, the number of steps at each stage does not exceed

$$T(\epsilon) = \min\left\{t \ge 1 : \epsilon_t \le \frac{\epsilon}{3}\right\} + 1. \tag{71}$$

5.2 Composite Mirror Prox algorithm for semi-separable optimization

We are about to apply the approach above to the semi-separable problem (63), (6).

Problem setup we consider now is as follows (cf. Sect. 4.1). For every k, $1 \le k \le K$, we are given

1. Euclidean spaces E_k and $\overline{E}_k$ along with their nonempty closed and bounded convex subsets Y_k and Z_k, respectively;

2. proximal setups for (E_k, Y_k) and $(\overline{E}_k, Z_k)$, that is, norms $p_k(\cdot)$ on E_k, norms q_k on $\overline{E}_k$, and d.g.f.'s $\omega_k(\cdot) : Y_k \to \mathbf{R}$, $\overline{\omega}_k(\cdot) : Z_k \to \mathbf{R}$, which are compatible with $p_k(\cdot)$ and $q_k(\cdot)$, respectively;
3. linear mapping $y^k \mapsto A_k y^k : E_k \to E$, where E is a Euclidean space;
4. Lipschitz continuous convex functions $\psi_k(y^k) : Y_k \to \mathbf{R}$ along with their *saddle point representations*

$$\psi_k(y^k) = \sup_{z^k \in Z_k} [\phi_k(y^k, z^k) - \overline{\Psi}_k(z^k)], \quad 1 \le k \le K, \tag{72}$$

where $\phi_k(y^k, z^k) : Y_k \times Z_k \to \mathbf{R}$ are smooth (with Lipschitz continuous gradients) functions convex in $y^k \in Y_k$ and concave in $z^k \in Z_k$, and $\overline{\Psi}_k(z^k) : Z_k \to \mathbf{R}$ are Lipschitz continuous convex functions such that the problems of the form

$$\min_{z^k \in Z_k} \left[\overline{\omega}_k(z^k) + \langle \xi^k, z^k \rangle + \alpha \overline{\Psi}_k(z^k)\right] \quad [\alpha > 0] \tag{73}$$

are easy to solve;
5. Lipschitz continuous convex functions $\Psi_k(y^k) : Y_k \to \mathbf{R}$ such that the problems of the form

$$\min_{y^k \in Y_k} \left[\omega_k(y^k) + \langle \xi^k, y^k \rangle + \alpha \Psi_k(y^k)\right] \quad [\alpha > 0]$$

are easy to solve;
6. a norm $\pi^*(\cdot)$ on E, with conjugate norm $\pi(\cdot)$, along with a d.g.f. $\widehat{\omega}(\cdot) : W := \{w \in E : \pi(w) \le 1\} \to \mathbf{R}$ compatible with $\pi(\cdot)$ and is such that problems of the form

$$\min_{w \in W} [\widehat{\omega}(w) + \langle \xi, w \rangle]$$

are easy to solve.

The outlined data define the sets

$$Y_k^+ = \left\{[y^k; \tau^k] : \ y^k \in Y_k, \tau^k \ge \Psi_k(y^k)\right\} \subset E_k^+ := E_k \times \mathbf{R}, \ 1 \le k \le K,$$
$$Z_k^+ = \left\{[z^k; \sigma^k] : \ z^k \in Z_k, \sigma^k \ge \overline{\Psi}_k(z^k)\right\} \subset \overline{E}_k^+ := \overline{E}_k \times \mathbf{R}, \ 1 \le k \le K.$$

The problem of interest here is problem (63), (72):

$$\mathrm{Opt} = \min_{[y^1;\ldots;y^K]} \max_{[z^1;\ldots;z^K]} \left\{\sum_{k=1}^K [\phi_k(y^k, z^k) + \Psi_k(y^k) - \overline{\Psi}_k(z^k)] : \ \pi^*\left(\sum_{k=1}^K A_k y^k - b\right) \le 0,\right.$$

$$
\begin{aligned}
&[y^1;\ldots;y^K]\in Y_1\times\cdots\times Y_K,\ \ [z^1;\ldots;z^k]\in Z_1\times\cdots\times Z_K\Bigg\}\\
&=\min_{\{[y^k;\tau^k]\}_{k=1}^K}\ \max_{\{[z^k;\sigma^k]\}_{k=1}^K}\Bigg\{\sum_{k=1}^K[\phi_k(y^k,z^k)+\tau^k\\
&\quad-\sigma^k]:\ \max_{w\in W}\sum_{k=1}^K\langle A_k y^k-b,w\rangle\le 0,\\
&\quad\Big\{[y^k;\tau^k]\in Y_k^+\Big\}_{k=1}^K,\ \ \Big\{[z^k;\sigma^k]\in Z_k^+\Big\}_{k=1}^K,\ w\in W\Bigg\}. \qquad (74)
\end{aligned}
$$

Solving (74) using the approach in the previous section amounts to resolving a sequence of problems (P_s) as in (66) where, with a slight abuse of notation,

$$
\begin{aligned}
Y&=\Big\{y=\Big\{[y^k;\tau^k]\Big\}_{k=1}^K:\ [y^k;\tau^k]\in Y_k^+,\ \tau^k\le C_k,\ 1\le k\le K\Big\};\\
f(y)&=\max_{z=\{[z^k;\sigma^k]\}_{k=1}^K}\Bigg\{\sum_{k=1}^K\Big[\phi_k(y^k,z^k)+\tau^k-\sigma^k\Big]:\ z\in Z=\Big\{[z^k;\sigma^k]\in Z_k^+\Big\}_{k=1}^K\Bigg\};\\
g(y)&=\max_w\Bigg\{\sum_{k=1}^K\langle A_k y^k-b,w\rangle:\ w\in W\Bigg\}.
\end{aligned}
$$

Here $C_k\ge\max_{y^k\in Y_k}\Psi_k(y^k)$ are finite constants introduced to make Y compact, as required in the premise of Proposition 5; it is immediately seen that the magnitudes of these constants (same as their very presence) does not affect the algorithm $\mathcal{B}$ we are about to describe.

The algorithm $\mathcal{B}$ we intend to use will solve (P_s) by reducing the problem to the saddle point problem

$$
\begin{aligned}
\overline{\mathrm{Opt}}=\min_y\ \max_{[z;w]}\Bigg\{&\Phi(y,[z;w]):=\alpha\sum_{k=1}^K\Big[\phi_k(y^k,z^k)+\tau^k-\sigma^k\Big]\\
&+(1-\alpha)\sum_{k=1}^K\langle A_k y^k-b,w\rangle:\\
&y=\Big\{[y^k;\tau^k]\Big\}_{k=1}^K\in Y,\ [z=\Big\{[z^k;\sigma^k]\Big\}_{k=1}^K\in Z;\ w\in W]\Bigg\},
\end{aligned}
$$

where $\alpha=\alpha_s$.

Setting

$$
\begin{aligned}
U&=\Big\{u=\Big[y^1;\ldots;y^K;z^1;\ldots;z^K;w\Big]:\ y^k\in Y_k,\ z^k\in Z_k,\ 1\le k\le K,w\in W\Big\},\\
X&=\Big\{\Big[u;v=[\tau^1;\ldots;\tau^K;\sigma^1;\ldots;\sigma^K]\Big]:\ u\in U,\ \Psi_k(y^k)\le\tau^k\le C_k,
\end{aligned}
$$

$$\overline{\Psi}_k(z^k) \le \sigma^k,\ 1 \le k \le K\Big\},$$

X can be thought of as the domain of the variational inequality associated with (75), the monotone operator in question being

$$\begin{aligned} F(u,v) &= [F_u(u); F_v], \\ F_u(u) &= \begin{bmatrix} \{\alpha\nabla_y\phi_k(y^k,z^k) + (1-\alpha)A_k^T w\}_{k=1}^K \\ \{-\alpha\nabla_z\phi_k(y^k,z^k)\}_{k=1}^K \\ (1-\alpha)[b - \sum_{k=1}^K A_k y^k] \end{bmatrix}, \\ F_v &= \alpha[1;\ldots;1]. \end{aligned} \tag{75}$$

By exactly the same reasons as in Sect. 4, with properly assembled norm on the embedding space of U and d.g.f., (75) can be solved by the MP algorithm from Sect. 3.2. Let us denote

$$\zeta^{s,t} = \left[\widehat{y}^{s,t} = \left\{[\widehat{y}^k;\widehat{\tau}^k]\right\}_{k=1}^K \in Y; \left[z^{s,t} \in Z; w^{s,t} \in W\right]\right]$$

the approximate solution obtained in course of $t = 1, 2, \ldots$ steps of CoMP when solving (P_s), and let

$$\widehat{f}_s^t := \max_{z\in Z, w\in W} \Phi(\widehat{y}^{s,t}, [z; w]) = \alpha\sum_{k=1}^K [\psi_k(\widehat{y}^k) + \widehat{\tau}^k] + (1-\alpha)\pi^*\left(\sum_{k=1}^K A_k\widehat{y}^k - b\right)$$

be the corresponding value of the objective of (P_s). It holds

$$\widehat{f}_s^t - \overline{\mathrm{Opt}} \le \epsilon_{\mathrm{Sad}}(\zeta^{s,t}\big|Y, Z\times W, \Phi) \le \epsilon_t := O(1)\mathcal{L}/t, \tag{76}$$

where $\mathcal{L} < \infty$ is explicitly given by the proximal setup we use and by the related Lipschitz constant of $F_u(\cdot)$ (note that this constant can be chosen to be independent of $\alpha \in [0, 1]$). We assume that computing the corresponding objective value is a part of step t (these computations increase the complexity of a step by factor at most $O(1)$), and thus that $\overline{f}_s^t \le \widehat{f}_s^t$. By (76), the quantity $\widehat{f}_s^t - \epsilon_t$ is a valid lower bound on the optimal value of (P_s), and thus we can ensure that $\underline{f}_{s,t} \ge \widehat{f}_s^t - \epsilon_t$. The bottom line is that with the outlined implementation, we have

$$\overline{f}_s^t - \underline{f}_{s,t} \le \epsilon_t$$

for all s, t, with ϵ_t given by (76). Consequently, by Proposition (5), the total number of CoMP steps needed to find a belonging to the domain of the problem of interest (63) ϵ-feasible and ϵ-optimal solution to this problem can be upper-bounded by

$$O(1)\ln\left(\frac{3L}{\epsilon}\right)\left(\frac{\mathcal{L}}{\epsilon}\right),$$

where L and $\mathcal{L}$ are readily given by the smoothness parameters of ϕ_k and by the proximal setup we use.

5.3 Numerical illustration: ℓ_1-minimization

Problem of interest We consider the simple ℓ_1 minimization problem

$$\min_{x \in X} \{\|x\|_1 : \ Ax = b\} \tag{77}$$

where $x \in \mathbf{R}^n$, $A \in \mathbf{R}^{m \times n}$ and $m < n$. Note that this problem can also be written in the semi-separable form

$$\min_{x \in X} \left\{ \sum_{k=1}^{K} \|x_k\|_1 : \ \sum_{k=1}^{K} A_k x_k = b \right\}$$

if the data is partitioned into K blocks: $x = [x_1; x_2; \ldots; x_K]$ and $A = [A_1, A_2, \ldots, A_K]$.

Our main purpose here is to test the approach described in 5.1 and compare it to the simplest approach where we directly apply CoMP to the (saddle point reformulation of the) problem $\min_{x \in X}[\|x\|_1 + R\|Ax - b\|_2]$ with large enough value of R. For the sake of simplicity, we work with the case when $K = 1$ and $X = \{x \in \mathbf{R}^n : \|x\|_2 \leq 1\}$.

Generating the data In the experiments to be reported, the data of (77) were generated as follows. Given m, n, we first build a sparse solution x^* by drawing random vector from the standard Gaussian distribution $\mathcal{N}(0, I_n)$, zeroing out part of the entries and scaling the resulting vector to enforce $x^* \in X$. We also build a dual solution λ^* by scaling a random vector from distribution $\mathcal{N}(0, I_m)$ to satisfy $\|\lambda^*\|_2 = R_*$ for a prescribed R_*. Next we generate A and b such that x^* and λ^* are indeed the optimal primal and dual solutions to the ℓ_1 minimization problem (77), i.e. $A^T\lambda^* \in \partial\big|_{x=x^*}\|x\|_1$ and $Ax^* = b$. To achieve this, we set

$$A = \frac{1}{\sqrt{n}}\widehat{F}_n + pq^T, \ b = Ax^*$$

where $p = \frac{\lambda^*}{\|\lambda^*\|_2^2}$, $q \in \partial\big|_{x=x^*}\|x\|_1 - \frac{1}{\sqrt{n}}\widehat{F}_n\lambda^*$, and $\widehat{F}_n$ is a $m \times n$ submatrix randomly selected from the DFT matrix F_n. We expect that the larger is the $\|\cdot\|_2$-norm R_* of the dual solution, the harder is problem (77).

Implementing the algorithm When implementing the algorithm from Sect. 5.2, we apply at each stage $s = 1, 2, \ldots$ CoMP to the saddle point problem

$$(P_s): \quad \min_{x,\tau:\ \|x\|_2 \leq 1, \tau \geq \|x\|_1} \ \max_{w:\|w\|_2 \leq 1} \{\alpha_s \tau + (1 - \alpha_s)\langle Ax - b, w\rangle\}.$$

The proximal setup for CoMP is given by equipping the embedding space of $U = \{u = [x; w] : x \in X, \|w\|_2 \leq 1\}$ with the norm $\|u\|_2 = \sqrt{\frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|w\|_2^2}$ and equipping U with the d.g.f. $\omega(u) = \frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|w\|_2^2$. In the sequel we refer to the

Table 5 ℓ_1-minimization

n	m	c	Sequential CoMP		Simple CoMP	
		$(R_* = c \cdot n)$	Steps	CPU (s)	Steps	CPU (s)
1,024	512	1	7,653	18.68	31,645	67.78
		5	43,130	44.66	90,736	90.67
		10	48,290	49.04	93,989	93.28
4,096	2,048	1	28,408	85.83	46,258	141.10
		5	45,825	199.96	93,483	387.88
		10	52,082	179.10	98,222	328.31
16,384	8,192	1	43,646	358.26	92,441	815.97
		5	48,660	454.70	93,035	784.05
		10	55,898	646.36	1,01,881	1,405.80
65,536	32,768	1	45,153	3,976.51	92,036	4,522.43
		5	55,684	4,138.62	1,00,341	8,054.35
		10	69,745	6,214.18	1,09,551	9,441.46
262,144	1,31,072	1	46,418	6,872.64	96,044	14,456.99
		5	69,638	10,186.51	1,09,735	16,483.62
		10	82,365	12,395.67	95,756	13,634.60

Platform: ISyE Condor Cluster

resulting algorithm as *sequential* CoMP. For comparison, we solve the same problem by applying CoMP to the saddle point problem

$$(P_R): \quad \min_{x,\tau:\ \|x\|_2\le 1, \tau\ge\|x\|_1} \ \max_{w:\|w\|_2\le 1} \{\tau + R\langle Ax - b, w\rangle\}$$

with $R = R_*$; the resulting algorithm is referred to as *simple* CoMP. Both sequential CoMP and simple CoMP algorithms are terminated when the relative nonoptimality and constraint violation are both less than $\epsilon = 10^{-5}$, namely,

$$\epsilon(x) := \max\left\{\frac{\|x\|_1 - \|x_*\|_1}{\|x_*\|_1}, \|Ax - b\|_2\right\} \le 10^{-5}.$$

Numerical results are presented in Table 5. One can immediately see that to achieve the desired accuracy, the simple CoMP with R set to R_*, i.e., to the exact magnitude of the true Lagrangian multiplier, requires almost twice as many steps as the sequential CoMP. In more realistic examples, the simple CoMP will additionally suffer from the fact that the magnitude of the optimal Lagrange multiplier is not known in advance, and the penalty R in (P_R) should be somehow tuned "online."

Acknowledgments Research of the first and the third authors was supported by the NSF Grant CMMI-1232623. Research of the second author was supported by the CNRS-Mastodons Project GARGANTUA, and the LabEx PERSYVAL-Lab (ANR-11-LABX-0025).

Appendix 1: Proof of Theorem 1

0^o. Let us verify that the prox-mapping (28) indeed is well defined whenever $\zeta = \gamma F_v$ with $\gamma > 0$. All we need is to show that whenever $u \in U$, $\eta \in E_u$, $\gamma > 0$ and $[w_t; s_t] \in X$, $t = 1, 2, \ldots$, are such that $\|w_t\|_2 + \|s_t\|_2 \to \infty$ as $t \to \infty$, we have

$$r_t := \underbrace{\langle \eta - \omega'(u), w_t \rangle + \omega(w_t)}_{a_t} + \underbrace{\gamma \langle F_v, s_t \rangle}_{b_t} \to \infty, \ t \to \infty.$$

Indeed, assuming the opposite and passing to a subsequence, we make the sequence r_t bounded. Since $\omega(\cdot)$ is strongly convex, modulus 1, w.r.t. $\|\cdot\|$, and the linear function $\langle F_v, s \rangle$ of $[w; s]$ is below bounded on X by **A4**, boundedness of the sequence $\{r_t\}$ implies boundedness of the sequence $\{w_t\}$, and since $\|[w_t; s_t]\|_2 \to \infty$ as $t \to \infty$, we get $\|s_t\|_2 \to \infty$ as $t \to \infty$. Since $\langle F_v, s \rangle$ is coercive in s on X by **A4**, and $\gamma > 0$, we conclude that $b_t \to \infty$, $t \to \infty$, while the sequence $\{a_t\}$ is bounded since the sequence $\{w_t \in U\}$ is so and ω is continuously differentiable. Thus, $\{a_t\}$ is bounded, $b_t \to \infty$, $t \to \infty$, implying that $r_t \to \infty$, $t \to \infty$, which is the desired contradiction 1°. Recall the well-known identity [9]: for all $u, u', w \in U$ one has

$$\langle V'_u(u'), w - u' \rangle = V_u(w) - V_{u'}(w) - V_u(u'). \tag{78}$$

Indeed, the right hand side is

$$\begin{aligned}
&[\omega(w) - \omega(u) - \langle \omega'(u), w - u \rangle] - [\omega(w) - \omega(u') - \langle \omega'(u'), w - u' \rangle] \\
&\quad - [\omega(u') - \omega(u) - \langle \omega'(u), u' - u \rangle] \\
&= \langle \omega'(u), u - w \rangle + \langle \omega'(u), u' - u \rangle + \langle \omega'(u'), w - u' \rangle \\
&= \langle \omega'(u') - \omega'(u), w - u' \rangle = \langle V'_u(u'), w - u' \rangle.
\end{aligned}$$

For $x = [u; v] \in X$, $\xi = [\eta; \zeta]$, let $P_x(\xi) = [u'; v'] \in X$. By the optimality condition for the problem (28), for all $[s; w] \in X$,

$$\langle \eta + V'_u(u'), u' - s \rangle + \langle \zeta, v' - w \rangle \le 0,$$

which by (78) implies that

$$\langle \eta, u' - s \rangle + \langle \zeta, v' - w \rangle \le \langle V'_u(u'), s - u' \rangle = V_u(s) - V_{u'}(s) - V_u(u'). \tag{79}$$

2°. When applying (79) with $[u; v] = [u_\tau; v_\tau] = x_\tau$, $\xi = \gamma_\tau F(x_\tau) = [\gamma_\tau F_u(u_\tau); \gamma_\tau F_v]$, $[u'; v'] = [u'_\tau; v'_\tau] = y_\tau$, and $[s; w] = [u_{\tau+1}; v_{\tau+1}] = x_{\tau+1}$ we obtain:

$$\gamma_\tau [\langle F_u(u_\tau), u'_\tau - u_{\tau+1} \rangle + \langle F_v, v'_\tau - v_{\tau+1} \rangle] \le V_{u_\tau}(u_{\tau+1}) - V_{u'_\tau}(u_{\tau+1}) - V_{u_\tau}(u'_\tau); \tag{80}$$

and applying (79) with $[u; v] = x_\tau, \xi = \gamma_\tau F(y_\tau), [u'; v'] = x_{\tau+1}$, and $[s; w] = z \in X$ we get:

$$\gamma_\tau[\langle F_u(u'_\tau), u_{\tau+1} - s\rangle + \langle F_v, v_{\tau+1} - w\rangle] \le V_{u_\tau}(s) - V_{u_{\tau+1}}(s) - V_{u_\tau}(u_{\tau+1}). \tag{81}$$

Adding (81) to (80) we obtain for every $z = [s; w] \in X$

$$\gamma_\tau\langle F(y_\tau), y_\tau - z\rangle = \gamma_\tau[\langle F_u(u'_\tau), u'_\tau - s\rangle + \langle F_v, v'_\tau - w\rangle] \le V_{u_\tau}(s) - V_{u_{\tau+1}}(s) + \underbrace{\gamma_\tau\langle F_u(u'_\tau) - F_u(u_\tau), u'_\tau - u_{\tau+1}\rangle - V_{u'_\tau}(u_{\tau+1}) - V_{u_\tau}(u'_\tau)}_{\delta_\tau}. \tag{82}$$

Due to the strong convexity, modulus 1, of $V_u(\cdot)$ w.r.t. $\|\cdot\|$, $V_u(u') \ge \frac{1}{2}\|u - u'\|^2$ for all u, u'. Therefore,

$$\begin{aligned}\delta_\tau &\le \gamma_\tau\|F_u(u'_\tau) - F_u(u_\tau)\|_*\|u'_\tau - u_{\tau+1}\| - \frac{1}{2}\|u'_\tau - u_{\tau+1}\|^2 - \frac{1}{2}\|u_\tau - u'_\tau\|^2\\ &\le \frac{1}{2}\left[\gamma_\tau^2\|F_u(u'_\tau) - F_u(u_\tau)\|_*^2 - \|u_\tau - u'_\tau\|^2\right]\\ &\le \frac{1}{2}\left[\gamma_\tau^2[M + L\|u'_\tau - u_\tau\|]^2 - \|u_\tau - u'_\tau\|^2\right],\end{aligned}$$

where the last inequality is due to (23). Note that $\gamma_\tau L < 1$ implies that

$$\gamma_\tau^2[M + L\|u'_\tau - u_\tau\|]^2 - \|u'_\tau - u_\tau\|^2 \le \max_r\left[\gamma_\tau^2[M + Lr]^2 - r^2\right] = \frac{\gamma_\tau^2 M^2}{1 - \gamma_\tau^2 L^2}.$$

Let us assume that the stepsizes $\gamma_\tau > 0$ ensure that (30) holds, meaning that $\delta_\tau \le \gamma_\tau^2 M^2$ (which, by the above analysis, is definitely the case when $0 < \gamma_\tau \le \frac{1}{\sqrt{2}L}$; when $M = 0$, we can take also $\gamma_\tau \le \frac{1}{L}$). When summing up inequalities (82) over $\tau = 1, 2, \ldots, t$ and taking into account that $V_{u_{t+1}}(s) \ge 0$, we conclude that for all $z = [s; w] \in X$,

$$\sum_{\tau=1}^t \lambda_\tau^t\langle F(y_\tau), y_\tau - z\rangle \le \frac{V_{u_1}(s) + \sum_{\tau=1}^t \delta_\tau}{\sum_{\tau=1}^t \gamma_\tau} \le \frac{V_{u_1}(s) + M^2\sum_{\tau=1}^t \gamma_\tau^2}{\sum_{\tau=1}^t \gamma_\tau},$$

$$\lambda_\tau^t = \gamma_\tau / \sum_{i=1}^t \gamma_i.$$

□

Appendix 2: Proof of Lemma 1

Proof All we need to verify is the second inequality in (38). To this end note that when $t = 1$, the inequality in (38) holds true by definition of $\widehat{\Theta}(\cdot)$. Now let $1 < t \leq N + 1$. Summing up the inequalities (82) over $\tau = 1, \ldots, t - 1$, we get for every $x = [u; v] \in X$:

$$\begin{aligned}\sum_{\tau=1}^{t-1} \gamma_\tau \langle F(y_\tau), y_\tau - [u; v] \rangle &\leq V_{u_1}(u) - V_{u_t}(u) + \sum_{\tau=1}^{t-1} \delta_\tau \\ &\leq V_{u_1}(u) - V_{u_t}(u) + \sum_{\tau=1}^{t-1} \delta_\tau \\ &\leq V_{u_1}(u) - V_{u_t}(u) + M^2 \sum_{\tau=1}^{t-1} \gamma_\tau^2\end{aligned}$$

(we have used (30)). When $[u; v]$ is z_*, the left hand side in the resulting inequality is ≥ 0, and we arrive at

$$V_{u_t}(u_*) \leq V_{u_1}(u_*) + M^2 \sum_{\tau=1}^{t-1} \gamma_\tau^2,$$

hence

$$\frac{1}{2}\|u_t - u_*\|^2 \leq V_{u_1}(u_*) + M^2 \sum_{\tau=1}^{t-1} \gamma_\tau^2$$

hence also

$$\|u_t - u_1\|^2 \leq 2\|u_t - u_*\|^2 + 2\|u_* - u_1\|^2 \leq 4\left[V_{u_1}(u_*) + M^2 \sum_{\tau=1}^{t-1} \gamma_\tau^2\right] + 4V_{u_1}(u_*)$$

and therefore

$$\|u_t - u_1\| \leq 2\sqrt{2V_{u_1}(u_*) + M^2 \sum_{\tau=1}^{t-1} \gamma_t^2} = R_N, \tag{83}$$

and (38) follows. □

Appendix 3: Proof of Proposition 3

Proof From (82) and (30) it follows that

$$\begin{aligned}\forall(x=[u;v]\in X,\tau\le N):\lambda_\tau\langle F(y_\tau),y_\tau-x\rangle\\ \le\frac{\lambda_\tau}{\gamma_\tau}[V_{u_\tau}(u)-V_{u_{\tau+1}}(u)]+M^2\lambda_\tau\gamma_\tau.\end{aligned}$$

Summing up these inequalities over $\tau=1,\dots,N$, we get $\forall(x=[u;v]\in X)$:

$$\begin{aligned}&\sum_{\tau=1}^N\lambda_\tau\langle F(y_\tau),y_\tau-x\rangle\\ &\le\frac{\lambda_1}{\gamma_1}[V_{u_1}(u)-V_{u_2}(u)]+\frac{\lambda_2}{\gamma_2}[V_{u_2}(u)-V_{u_3}(u)]+\cdots\\ &\quad+\frac{\lambda_N}{\gamma_N}[V_{u_N}(u)-V_{u_{N+1}}(u)]+M^2\sum_{\tau=1}^N\lambda_\tau\gamma_\tau\\ &=\underbrace{\frac{\lambda_1}{\gamma_1}}_{\ge0}V_{u_1}(u)+\underbrace{\left[\frac{\lambda_2}{\gamma_2}-\frac{\lambda_1}{\gamma_1}\right]}_{\ge0}V_{u_2}(u)+\cdots+\underbrace{\left[\frac{\lambda_N}{\gamma_N}-\frac{\lambda_{N-1}}{\gamma_{N-1}}\right]}_{\ge0}V_{u_N}(u)\\ &\quad-\frac{\lambda_N}{\gamma_N}\underbrace{V_{u_{N+1}}(u)}_{\ge0}+M^2\sum_{\tau=1}^N\lambda_\tau\gamma_\tau\\ &\le\frac{\lambda_1}{\gamma_1}\widehat{\Theta}(\max[R_N,\|u-u_1\|])+\left[\frac{\lambda_2}{\gamma_2}-\frac{\lambda_1}{\gamma_1}\right]\widehat{\Theta}(\max[R_N,\|u-u_1\|])+\cdots\\ &\quad+\left[\frac{\lambda_N}{\gamma_N}-\frac{\lambda_{N-1}}{\gamma_{N-1}}\right]\widehat{\Theta}(\max[R_N,\|u-u_1\|])+M^2\sum_{\tau=1}^N\lambda_\tau\gamma_\tau,\\ &=\frac{\lambda_N}{\gamma_N}\widehat{\Theta}(\max[R_N,\|u-u_1\|])+M^2\sum_{\tau=1}^N\lambda_\tau\gamma_\tau,\end{aligned}$$

where the concluding inequality is due to (38), and (40) follows. □

Appendix 4: Proof of Proposition 5

1^o. $h_{s,t}(\alpha)$ are concave piecewise linear functions on $[0,1]$ which clearly are pointwise nonincreasing in time. As a result, $\mathrm{Gap}(s,t)$ is nonincreasing in time. Further, we have

$$\begin{aligned}\mathrm{Gap}(s,t)=\max_{\alpha\in[0,1]}\Bigg\{\min_\lambda\sum_{(p,q)\in Q_{s,t}}\lambda_{pq}[\alpha(p-\underline{\mathrm{Opt}}_{s,t})+(1-\alpha)q]:\ \lambda_{pq}\ge0,\\ \sum_{(p,q)\in Q_{s,t}}\lambda_{pq}=1\Bigg\}\end{aligned}$$

$$= \max_{\alpha\in[0,1]} \sum_{(p,q)\in Q_{s,t}} \lambda^*_{pq}[\alpha(p - \underline{\mathrm{Opt}}_{s,t}) + (1-\alpha)q]$$

$$= \max\left[\sum_{(p,q)\in Q_{s,t}} \lambda^*_{pq}(p - \underline{\mathrm{Opt}}_{s,t}), \sum_{(p,q)\in Q_{s,t}} \lambda^*_{pq} q\right],$$

where $\lambda^*_{pq} \geq 0$ and sum up to 1. Recalling that for every $(p,q) \in Q_{s,t}$ we have at our disposal $y_{pq} \in Y$ such that $p \geq f(y_{pq})$ and $q \geq g(y_{pq})$, setting $\widehat{y}^{s,t} = \sum_{(p,q)\in Q_{s,t}} \lambda^*_{pq} y_{pq}$ and invoking convexity of f, g, we get

$$\begin{aligned} f(\widehat{y}^{s,t}) &\leq \sum_{(p,q)\in Q_{s,t}} \lambda^*_{pq} p \leq \underline{\mathrm{Opt}}_{s,t} + \mathrm{Gap}(s,t), \;\; g(\widehat{y}^{s,t}) \\ &\leq \sum_{(p,q)\in Q_{s,t}} \lambda^*_{pq} q \leq \mathrm{Gap}(s,t); \end{aligned}$$

and (69) follows, due to $\underline{\mathrm{Opt}}_{s,t} \leq \mathrm{Opt}$. 2°. We have $\overline{f}^t_s = \alpha_s f(y^{s,t}) + (1-\alpha_s) g(y^{s,t})$ for some $y^{s,t} \in Y$ which we have at our disposal at step t, implying that $(\widehat{p} = f(y^{s,t}), \widehat{q} = g(y^{s,t})) \in Q_{s,t}$. Hence by definition of $h_{s,t}(\cdot)$ it holds

$$h_{s,t}(\alpha_s) \leq \alpha_s(\widehat{p} - \underline{\mathrm{Opt}}_{s,t}) + (1-\alpha_s)\widehat{q} = \overline{f}^t_s - \alpha_s \underline{\mathrm{Opt}}_{s,t} \leq \overline{f}^t_s - \underline{f}_{s,t},$$

where the concluding inequality is given by (67). Thus, $h_{s,t}(\alpha_s) \leq \overline{f}^t_s - \underline{f}_{s,t} \leq \epsilon_t$. On the other hand, if stage s does not terminate in course of the first t steps, α_s is well-centered in the segment $\Delta_{s,t}$ where the concave function $h_{s,t}(\alpha)$ is nonnegative. We conclude that $0 \leq \mathrm{Gap}(s,t) = \max_{0\leq\alpha\leq 1} h_{s,t}(\alpha) = \max_{\alpha\in\Delta_{s,t}} h_{s,t}(\alpha) \leq 3h_{s,t}(\alpha_s)$. Thus, if a stage s does not terminate in course of the first t steps, we have $\mathrm{Gap}(s,t) \leq 3\epsilon_t$, which implies (71). Further, α_s is the midpoint of the segment $\Delta^{s-1} = \Delta_{s-1,t_{s-1}}$, where t_r is the last step of stage r (when $s = 1$, we should define Δ^0 as $[0,1]$), and α_s is not well-centered in the segment $\Delta^s = \Delta_{s,t_s} \subset \Delta_{s-1,t_{s-1}}$, which clearly implies that $|\Delta^s| \leq \frac{3}{4}|\Delta^{s-1}|$. Thus, $|\Delta^s| \leq \left(\frac{3}{4}\right)^s$ for all s. On the other hand, when $|\Delta_{s,t}| < 1$, we have $\mathrm{Gap}(s,t) = \max_{\alpha\in\Delta_{s,t}} h_{s,t}(\alpha) \leq 3L|\Delta_{s,t}|$ (since $h_{s,t}(\cdot)$ is Lipschitz continuous with constant $3L$[10] and $h_{s,t}(\cdot)$ vanishes at (at least) one endpoint of $\Delta_{s,t}$). Thus, the number of stages before $\mathrm{Gap}(s,t) \leq \epsilon$ is reached indeed obeys the bound (70). □

References

1. Andersen, E. D., Andersen, K. D.: The MOSEK optimization tools manual. http://www.mosek.com/fileadmin/products/6_0/tools/doc/pdf/tools.pdf
2. Aujol, J.-F., Chambolle, A.: Dual norms and image decomposition models. Int. J. Comput. Vis. **63**(1), 85–104 (2005)

[10] We assume w.l.o.g. that $|\underline{\mathrm{Opt}}_{s,t}| \leq L$.

3. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
4. Becker, S., Bobin, J., Candès, E.J.: Nesta: a fast and accurate first-order method for sparse recovery. SIAM J. Imaging Sci. **4**(1), 1–39 (2011)
5. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 122–122 (2010)
6. Buades, A., Coll, B., Morel, J.-M.: A review of image denoising algorithms, with a new one. Multiscale Model. Simul. **4**(2), 490–530 (2005)
7. Candés, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM (JACM) **58**(3), 11 (2011)
8. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
9. Chen, G., Teboulle, M.: Convergence analysis of a proximal-like minimization algorithm using bregman functions. SIAM J. Optim. **3**(3), 538–543 (1993)
10. Deng, W., Lai, M.-J., Peng, Z., Yin, W.: Parallel multi-block admm with o (1/k) convergence, 2013. http://www.optimization-online.org/DB_HTML/2014/03/4282.html (2013)
11. Goldfarb, D., Ma, S.: Fast multiple-splitting algorithms for convex optimization. SIAM J. Optim. **22**(2), 533–556 (2012)
12. Goldfarb, D., Ma, S., Scheinberg, K.: Fast alternating linearization methods for minimizing the sum of two convex functions. Math. Program. **141**(1–2), 349–382 (2013)
13. Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx (2013)
14. Juditsky, A., Nemirovski, A.: First-order methods for nonsmooth largescale convex minimization: I general purpose methods; ii utilizing problems structure. In: Sra, S., Nowozin, S., Wright, S. (eds.) Optimization for Machine Learning, pp. 121–183. The MIT Press, (2011)
15. Lemarchal, C., Nemirovskii, A., Nesterov, Y.: New variants of bundle methods. Math. Program. **69**(1–3), 111–147 (1995)
16. Monteiro, R.D., Svaiter, B.F.: Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. SIAM J. Optim. **23**(1), 475–507 (2013)
17. Nemirovski, A.: 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**(1), 229–251 (2004)
18. Nemirovski, A., Onn, S., Rothblum, U.G.: Accuracy certificates for computational problems with convex structure. Math. Oper. Res. **35**(1), 52–78 (2010)
19. Nemirovski, A., Rubinstein, R.: An efficient stochastic approximation algorithm for stochastic saddle point problems. In: Dror, M., L'Ecuyer, P., Szidarovszky, F. (eds.) Modeling Uncertainty and Examination of Stochastic Theory, Methods, and Applications, pp. 155–184. Kluwer Academic Publishers, Boston (2002)
20. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
21. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
22. Orabona, F., Argyriou, A., Srebro, N.: Prisma: Proximal iterative smoothing algorithm. arXiv preprint arXiv:1206.2372, (2012)
23. Ouyang, Y., Chen, Y., Lan, G., Pasiliao, E. Jr.: An accelerated linearized alternating direction method of multipliers, arXiv:1401.6607 (2014)
24. Qin, Z., Goldfarb, D.: Structured sparsity via alternating direction methods. J. Mach. Learn. Res. **13**, 1373–1406 (2012)
25. Scheinberg, K., Goldfarb, D., Bai, X.: Fast first-order methods for composite convex optimization with backtracking. http://www.optimization-online.org/DB_FILE/2011/04/3004.pdf (2011)
26. Tseng, P.: Alternating projection-proximal methods for convex programming and variational inequalities. SIAM J. Optim. **7**(4), 951–965 (1997)
27. Tseng, P.: On accelerated proximal gradient methods for convex–concave optimization. SIAM J. Optim. (2008, submitted)
28. Wen, Z., Goldfarb, D., Yin, W.: Alternating direction augmented lagrangian methods for semidefinite programming. Math. Program. Comput. **2**(3–4), 203–230 (2010)