Semi-proximal Mirror-Prox for Nonsmooth Composite Minimization

Niao He, Zaid Harchaoui

Semi-proximal Mirror-Prox for Nonsmooth Composite Minimization *

Niao He
nhe6@gatech.edu
GeorgiaTech

Zaid Harchaoui
zaid.harchaoui@inria.fr
NYU, Inria

July 5, 2015

Abstract

We propose a new first-order optimisation algorithm to solve high-dimensional non-smooth composite minimisation problems. Typical examples of such problems have an objective that decomposes into a non-smooth empirical risk part and a non-smooth regularisation penalty. The proposed algorithm, called Semi-Proximal Mirror-Prox, leverages the Fenchel-type representation of one part of the objective while handling the other part of the objective via linear minimization over the domain. The algorithm stands in contrast with more classical proximal gradient algorithms with smoothing, which require the computation of proximal operators at each iteration and can therefore be impractical for high-dimensional problems. We establish the theoretical convergence rate of Semi-Proximal Mirror-Prox, which exhibits the optimal complexity bounds, i.e. $O(1/\epsilon^2)$, for the number of calls to linear minimization oracle. We present promising experimental results showing the interest of the approach in comparison to competing methods.

1 Introduction

A wide range of machine learning and signal processing problems can be formulated as the minimization of a composite objective:

$$\min_{x \in X} F(x) := f(x) + \|\mathcal{B}x\| \tag{1}$$

where X is closed and convex, f is convex and can be either smooth, or nonsmooth yet enjoys a particular structure. The term $\|\mathcal{B}x\|$ defines a regularization penalty through a norm $\|\cdot\|$, and $x \mapsto \mathcal{B}x$ a linear mapping on a closed convex set X. The function f usually corresponds to an empirical risk, that is an empirical average of a possibly non-smooth loss function evaluated on a set of data-points, while x encodes the learning parameters. All in all, the objective F has a doubly non-smooth structure.

In many situations, the objective function F of interest enjoys a favorable structure, namely a so-called *Fenchel-type representation* [7, 12, 14]:

$$f(x) = \max_{z \in Z} \{\langle x, Az\rangle - \psi(z)\} \tag{2}$$

where Z is convex compact subset of a Euclidean space, and $\psi(\cdot)$ is a convex function. Sec. 4 will give several examples of such situations. Fenchel-type representations can then be leveraged to use first-order optimisation algorithms.

*The authors would like to thank Anatoli Juditsky and Arkadi Nemirovski for fruitful discussions. This work was supported by the NSF Grant CMMI-1232623, the LabEx Persyval-Lab (ANR-11-LABX-0025), the project Titan (CNRS-Mastodons), the project Macaron (ANR-14-CE23-0003-01), the MSR-Inria joint centre, and the Moore-Sloan Data Science Environment at NYU.

A simple first option to minimise F is using the so-called Nesterov smoothing technique [21] along with a proximal gradient algorithm [23], assuming that the proximal operator associated with X is computationally tractable and cheap to compute. However, this is certainly not the case when considering problems with norms acting in the spectral domain of high-dimensional matrices, such as the matrix nuclear-norm [13] and structured extensions thereof [6, 2]. In the latter situation, another option is to use a smoothing technique now with a conditional gradient or Frank-Wolfe algorithm to minimize F, assuming that a *a linear minimization oracle* associated with X is cheaper to compute than the proximal operator [7, 15, 24]. Neither option takes advantage of the composite structure of the objective (1) or handles the case when the linear mapping $\mathcal{B}$ is nontrivial.

Contributions Our goal in this paper is to propose a new first-order optimization algorithm, called Semi-Proximal Mirror-Prox , designed to solve the difficult non-smooth composite optimisation problem (1), which does not require the exact computation of proximal operators. Instead, the Semi-Proximal Mirror-Prox relies upon i) Fenchel-type representability of f; ii) Linear minimization oracle associated with $\|\cdot\|$ in the domain X. While the Fenchel-type representability of f allows to cure the non-smoothness of f, the linear minimisation over the domain X allows to tackle the non-smooth regularisation penalty $\|\cdot\|$. We establish the theoretical convergence rate of Semi-Proximal Mirror-Prox, which exhibits the *optimal complexity bounds*, i.e. $O(1/\epsilon^2)$, for the number of calls to linear minimization oracle. Furthermore, Semi-Proximal Mirror-Prox generalizes previously proposed approaches and improves upon them in special cases:

1. Case $\mathcal{B} \equiv 0$: Semi-Proximal Mirror-Prox does not require assumptions on favorable geometry of dual domains Z or simplicity of $\psi(\cdot)$ in (2).

2. Case $\mathcal{B} = \mathbb{I}$: Semi-Proximal Mirror-Prox is competitive with previously proposed approaches [16, 24] based on smoothing techniques.

3. Case of non-trivial $\mathcal{B}$: Semi-Proximal Mirror-Prox is the first proximal-free or conditional-gradient-type optimization algorithm for (1).

Related work The Semi-Proximal Mirror-Prox algorithm belongs the family of conditional gradient algorithms, whose most basic instance is the Frank-Wolfe algorithm for constrained smooth optimization using a linear minimization oracle; see [13, 1, 4]. Recently, in [7, 14], the authors consider constrained non-smooth optimisation when the domain Z has a "favorable geometry", *i.e.* the domain is amenable to linear minimisation (favorable geometry), and establish a complexity bound with $O(1/\epsilon^2)$ calls to the linear minimization oracle. Recently, in [16], a method called conditional gradient sliding is proposed to solve similar problems, using a smoothing technique, with a complexity bound in $O(1/\epsilon^2)$ for the calls to the linear minimization oracle (LMO) and additionally a $O(1/\epsilon)$ bound for the linear operator evaluations. Actually, this $O(1/\epsilon^2)$ bound for the LMO complexity can be shown to be indeed *optimal* for conditional-gradient-type or LMO-based algorithms, when solving general[1] non-smooth convex problems [15].

However, these previous approaches are appropriate for objective with a non-composite structure. When applied to our problem (1), the smoothing would be applied to the objective taken as a whole, ignoring its composite structure. Conditional-gradient-type algorithms were recently proposed for composite objectives [8, 10, 26, 24, 17], but cannot be applied for our problem. In [10], f is smooth and $\mathcal{B}$ is identity matrix, whereas in [24], f is non-smooth and $\mathcal{B}$ is also the identity matrix. The proposed Semi-Proximal Mirror-Prox can be seen as a blend of the successful components resp. of the Composite Conditional Gradient algorithm [10] and the Composite Mirror-Prox [12], that enjoys the optimal complexity bound $O(1/\epsilon^2)$ on the total number of LMO calls, yet solves a broader class of convex problems than previously considered.

[1]Related research extended such approaches to stochastic or online settings [11, 9, 16]; such settings are beyond the scope of this work.

Outline The paper is organized as follows. In Section 2, we describe the norm-regularized nonsmooth problem of interest and illustrate it with several examples. In Section 3, we present the conditional gradient type method based on an inexact Mirror-Prox framework for structured variational inequalities. In Section 4, we present promising experimental results showing the interest of the approach in comparison to competing methods, resp. on a collaborative filtering for movie recommendation and link prediction for social network analysis applications.

2 Framework and assumptions

We present here our theoretical framework, which hinges upon a smooth convex-concave saddle point reformulation of the norm-regularized non-smooth minimization (3). We shall use the following notations throughout the paper. For a given norm $\|\cdot\|$, we define the dual norm as $\|s\|_* = \max_{\|x\|\leq 1}\langle s, x\rangle$. For any $x \in \mathbf{R}^{m\times n}$, $\|x\|_2 = \|x\|_F = (\sum_{i=1}^m \sum_{j=1}^n |x_{ij}|^2)^{1/2}$.

Problem We consider the composite minimization problem

$$\mathrm{Opt} = \min_{x\in X} \; f(x) + \|\mathcal{B}x\| \tag{3}$$

where X is a closed convex set in the Euclidean space E_x; $x \mapsto \mathcal{B}x$ is a linear mapping from X to $Y(\supset \mathcal{B}X)$, where Y is a closed convex set in the Euclidean space E_y. We make two important assumptions on the function f and the norm $\|\cdot\|$ defining the regularization penalty, explained below.

Fenchel-type Representation The non-smoothness of f can be challenging to tackle. However, in many cases of interest, the function f enjoys a favorable structure that allows to tackle it with smoothing techniques. We assume that the norm $f(x)$ is a non-smooth convex function given by

$$f(x) = \max_{z\in Z} \; \Phi(x, z) \tag{4}$$

a where $\Phi(x, z)$ is a smooth convex-concave function and Z is a convex and compact set in the Euclidean space E_z. Such representation was introduced and developed in [7, 12, 14], for the purpose of non-smooth optimisation. Fenchel-type representability can be interpreted as a general form of the smoothing-favorable structure of non-smooth functions used in the Nesterov smoothing technique [21]. Representations of this type are readily available for a wide family of "well-structured" nonsmooth functions f; see Sec. 4 for examples.

Composite Linear Minimization Oracle Proximal-gradient-type algorithms require the computation of a proximal operator at each iteration, i.e.

$$\min_{y\in Y} \left\{ \frac{1}{2}\|y\|_2^2 + \langle \eta, y\rangle + \alpha\|y\| \right\}. \tag{5}$$

For several cases of interest, described below, the computation of the proximal operator can be expensive or intractable. A classical example is the nuclear norm, whose proximal operator boils down to singular value thresholding, therefore requiring a full singular value decomposition. In contrast to the proximal operator, the linear minimization oracle can much cheaper. The linear minimization oracle (LMO) is a routine which, given an input $\alpha > 0$ and $\eta \in E_y$, returns a point

$$\min_{y\in Y} \; \{\langle \eta, y\rangle + \alpha\|y\|\} \tag{6}$$

In the case of the nuclear-norm, the LMO only requires the computation of the top pair of eigenvectors/eigenvalues, which is an order of magnitude fast in time-complexity.

Saddle Point Reformulation. The crux of our approach is a smooth convex-concave saddle point reformulation of (3). After massaging the saddle-point reformulation, we consider the variational inequality associated with the obtained saddle-point problem. For a constrained smooth optimisation problem, the corresponding variational inequality provides the sufficient and necessary condition for an optimal solution to the problem [3, 4]. For non-smooth optimization problems, the corresponding variational inequality is directly related to the accuracy certificate used to guarantee the accuracy of a solution to the optimisation problem; see Sec. 2.1 in [12] and [19]. We shall present then an algorithm to solve the variational inequality established below, that leverages its particular structure.

Assuming that f admits a Fenchel-type representation (4), we rewrite (3) in epigraph form

$$\min_{x\in X, y\in Y, \tau\geq\|y\|} \max_{z\in Z} \{\Phi(x,z) + \tau : y = \mathcal{B}x\},$$

which, with a properly selected $\rho > 0$, can be further approximated by

$$\widehat{\mathrm{Opt}} = \min_{x\in X, y\in Y, \tau\geq\|y\|} \max_{z\in Z} \{\Phi(x,z) + \tau + \rho\|y - \mathcal{B}x\|_2\} \tag{7}$$

$$= \min_{x\in X, y\in Y, \tau\geq\|y\|} \max_{z\in Z, \|w\|_2\leq 1} \{\Phi(x,z) + \tau + \rho\langle y - \mathcal{B}x, w\rangle\}\ . \tag{8}$$

In fact, when ρ is large enough one can always guarantee $\widehat{\mathrm{Opt}} = \mathrm{Opt}$. It is indeed sufficient to set ρ as the Lipschitz constant of $\|\cdot\|$ with respect to $\|\cdot\|_2$.

Introduce the variables $u := [x, y; z, w]$ and $v := \tau$. The variational inequality associated with the above saddle point problem is fully described by the domain

$$X_+ \quad = \quad \{x_+ = [u; v] : x \in X, y \in Y, z \in Z, \|w\|_2 \leq 1, \tau \geq \|y\|\}$$

and the monotone vector field

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

where

$$F_u\left(u = \begin{bmatrix} x \\ y \\ z \\ w \end{bmatrix}\right) = \begin{bmatrix} \nabla_x \Phi(x,z) - \rho\mathcal{B}^T w \\ \rho w \\ -\nabla_z \Phi(x,z) \\ \rho(\mathcal{B}x - y) \end{bmatrix}, \qquad F_v(v = \tau) = 1.$$

In the next section, we present an efficient algorithm to solve this type of variational inequality, which enjoys a particular structure; we call such an inequality *semi-structured*.

3 Semi-Proximal Mirror-Prox for Semi-structured Variational Inequalities

Semi-structured variational inequalities (Semi-VI) enjoy a particular product structure, that allows to get the best of two worlds, namely the proximal setup (where the proximal operator can be computed) and the LMO setup (where the linear minimization oracle can be computed). Basically, the domain X is decomposed as a Cartesian product over two sets $X = X_1 \times X_2$, such that X_1 admits a proximal-mapping while X_2 admits a linear minimization oracle. We now describe the main theoretical and algorithmic components of the Semi-Proximal Mirror-Prox algorithm, resp. in Sec. 3.1 and in Sec. 3.2, and finally describe the overall algorithm in Sec. 3.3.

3.1 Composite Mirror-Prox with Inexact Prox-mappings

We first present a new algorithm, which can be seen as an extension of the composite Mirror Prox algorithm, denoted CMP for brevity, that allows inexact computation of the Prox-mappings, and can solve a broad class of variational inequalites. The original Mirror Prox algorithm was introduced in [18], and was extended to composite minimization in [12] assuming exact computations of Prox-mappings.

Structured Variational Inequalities. We consider the variational inequality VI(X, F):

$$\text{Find } x_* \in X : \langle F(x), x - x_* \rangle \geq 0, \forall x \in X$$

with domain X and operator F that satisfy the assumptions (**A**.1)–(**A**.4) below.

(**A**.1) Set $X \subset E_u \times E_v$ is closed convex and its projection $PX = \{u : x = [u; v] \in X\} \subset U$, where U is convex and closed, E_u, E_v are Euclidean spaces;

(**A**.2) The function $\omega(\cdot) : U \to \mathbf{R}$ is continuously differentiable and also 1-strongly convex w.r.t. some norm[2] $\|\cdot\|$. This defines the Bregman distance

$$V_u(u') = \omega(u') - \omega(u) - \langle \omega'(u), u' - u \rangle \geq \frac{1}{2}\|u' - u\|^2 .$$

(**A**.3) The operator $F(x = [u, v]) : X \to E_u \times E_v$ is monotone and of form $F(u, v) = [F_u(u); F_v]$ with $F_v \in E_v$ being a constant and $F_u(u) \in E_u$ satisfying the condition

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

for some $L < \infty, M < \infty$;

(**A**.4) The linear form $\langle F_v, v \rangle$ of $[u; v] \in E_u \times E_v$ is bounded from below on X and is coercive on X w.r.t. v: whenever $[u^t; v^t] \in X$, $t = 1, 2, ...$ is a sequence such that $\{u^t\}_{t=1}^{\infty}$ is bounded and $\|v^t\|_2 \to \infty$ as $t \to \infty$, we have $\langle F_v, v^t \rangle \to \infty$, $t \to \infty$.

ϵ-Prox-mapping In the Composite Mirror Prox with exact Prox-mappings [12], the quality of an iterate, in the course of the algorithm, is measured through the so-called dual gap function

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

We give in Appendix A a refresher on dual gap functions, for the reader's convenience. We shall establish the complexity bounds in terms this dual gap function for our algorithm, which directly provides an accuracy certificate along the iterations. However, we first need to define what we mean by an inexact prox-mapping. Inexact proximal mapping were recently considered in the context of accelerated proximal gradient algorithms [25]. The definition we give below is more general, allowing for non-Euclidean proximal-mappings.

We introduce here the notion of ϵ-prox-mapping ($\epsilon \geq 0$). For $\xi = [\eta; \zeta] \in E_u \times E_v$ and $x = [u; v] \in X$, let us define the subset $P_x^\epsilon(\xi)$ of X as

$$P_x^\epsilon(\xi) = \{\widehat{x} = [\widehat{u}; \widehat{v}] \in X : \langle \eta + \omega'(\widehat{u}) - \omega'(u), \widehat{u} - s \rangle + \langle \zeta, \widehat{v} - w \rangle \leq \epsilon \ \forall [s; w] \in X\}.$$

When $\epsilon = 0$, this reduces to the exact prox-mapping, in the usual setting, that is

$$P_x(\xi) = \underset{[s;w] \in X}{\text{Argmin}} \ \{\langle \eta, s \rangle + \langle \zeta, w \rangle + V_u(s)\} .$$

When $\epsilon > 0$, this yields our definition of an inexact prox-mapping, with inexactness parameter ϵ. Note that for any $\epsilon \geq 0$, the set $P_x^\epsilon(\xi = [\eta; \gamma F_v])$ is well defined whenever $\gamma > 0$. The Composite Mirror-Prox with Inexact Prox-mappings is outlined in Algorithm 1.

[2]There is a slight abuse of notation here. The norm here is not the same as the one in problem (3)

Algorithm 1 Composite Mirror Prox Algorithm (CMP) for VI(X, F)

Input: stepsizes $\gamma_t > 0$, inexactness $\epsilon_t \geq 0$, $t = 1, 2, \ldots$
Initialize $x^1 = [u^1; v^1] \in X$
for $t = 1, 2, \ldots, T$ **do**

$$\begin{array}{rcl} y^t := [\widehat{u}^t; \widehat{v}^t] & \in & P^{\epsilon_t}_{x^t}(\gamma_t F(x^t)) = P^{\epsilon_t}_{x^t}(\gamma_t [F_u(u^t); F_v]) \\ x^{t+1} := [u^{t+1}; v^{t+1}] & \in & P^{\epsilon_t}_{x^t}(\gamma_t F(y^t)) = P^{\epsilon_t}_{x^t}(\gamma_t [F_u(\widehat{u}^t); F_v]) \end{array} \tag{9}$$

end for
Output: $\overline{x}_T := [\bar{u}_T; \bar{v}_T] = \left(\sum_{t=1}^T \gamma_t\right)^{-1} \sum_{t=1}^T \gamma_t y^t$

Note that this composite version of Mirror Prox algorithm works essentially *as if* there were no v-component at all. Therefore, the proposed algorithm is a not-trivial extension of the Composite Mirror-Prox with *exact prox-mappings*, both from a theoretical and algorithmic point of views. We establish below the theoretical convergence rate; see Appendix for the proof.

Theorem 3.1. *Assume that the sequence of step-sizes (γ_t) in the CMP algorithm satisfy*

$$\sigma_t := \gamma_t \langle F_u(\widehat{u}^t) - F_u(u^t), \widehat{u}^t - u^{t+1} \rangle - V_{\widehat{u}^t}(u^{t+1}) - V_{u^t}(\widehat{u}^t) \leq \gamma_t^2 M^2, \quad t = 1, 2, \ldots, T. \tag{10}$$

Then, denoting $\Theta[X] = \sup_{[u;v] \in X} V_{u^1}(u)$, for a sequence of inexact prox-mappings with inexactness $\epsilon_t \geq 0$, we have

$$\epsilon_{\mathrm{VI}}(\bar{x}_T | X, F) := \sup_{x \in X} \langle F(x), \bar{x}_T - x \rangle \leq \frac{\Theta[X] + M^2 \sum_{t=1}^T \gamma_t^2 + 2\sum_{t=1}^T \epsilon_t}{\sum_{t=1}^T \gamma_t}. \tag{11}$$

Remarks Note that the assumption on the sequence of step-sizes (γ_t) is clearly satisfied when $\gamma_t \leq (\sqrt{2}L)^{-1}$. When $M = 0$, it is satisfied as long as $\gamma_t \leq L^{-1}$.

Corollary 3.1. *Assume further that $X = X_1 \times X_2$, and let F be the monotone vector field associated with the saddle point problem*

$$\mathrm{SadVal} = \min_{x^1 \in X_1} \max_{x^2 \in X_2} \Phi(x^1, x^2), \tag{12}$$

two induced convex optimization problems

$$\begin{array}{rcll} \mathrm{Opt}(P) & = & \min_{x^1 \in X_1} \left[\overline{\Phi}(x^1) = \sup_{x^2 \in X_2} \Phi(x^1, x^2)\right] & (P) \\ \mathrm{Opt}(D) & = & \max_{x^2 \in X_2} \left[\underline{\Phi}(x^2) = \inf_{x^1 \in X_1} \Phi(x^1, x^2)\right] & (D) \end{array} \tag{13}$$

with convex-concave locally Lipschitz continuous cost function Φ. In addition, assuming that problem (P) in (13) is solvable with optimal solution x^1_* and denoting by $\bar{x}^1_T$ the projection of $\bar{x}_T \in X = X_1 \times X_2$ onto X_1, we have

$$\overline{\Phi}(\bar{x}^1_T) - \mathrm{Opt}(P) \leq \left[\sum_{t=1}^T \gamma_t\right]^{-1} \left[\Theta[\{x^1_*\} \times X_2] + M^2 \sum_{t=1}^T \gamma_t^2 + 2\sum_{t=1}^T \epsilon_t\right]. \tag{14}$$

The theoretical convergence rate established in Theorem 3.1 and Corollary 3.1 generalizes the previous result established in Corollary 3.1 in [12] for CMP with exact prox-mappings. Indeed, when exact prox-mappings are used, we recover the result of [12]. When inexact prox-mappings are used, the errors due to the inexactness of the prox-mappings accumulates and is reflected in the bound (34) and (14).

3.2 Composite Conditional Gradient

We now turn to a variant of the composite conditional gradient algorithm, denoted CCG, tailored for a particular class of problems, which we call *smooth semi-linear problems*. The composite conditional gradient algorithm was introduced in [10]. We present an extension here which will turn to be especially tailored for sub-problems that will be solved in Sec. 3.3.

Minimizing Smooth Semi-linear Problems. We consider the smooth semi-linear problem

$$\min_{x=[u;v]\in X} \left\{\phi^+(u,v) = \phi(u) + \langle \theta, v\rangle\right\} \tag{15}$$

represented by the pair $(X;\phi^+)$ such that the following assumptions are satisfied. We assume that

i) $X \subset E_u \times E_v$ is closed convex and its projection $PX \subset U$, where U is convex and compact;

ii) $\phi(u) : U \to \mathbf{R}$ be a convex continuously differentiable function, and there exists $1 < \kappa \leq 2$ and $L < \infty$ such that

$$\phi(u') \leq \phi(u) + \langle \nabla\phi(u), u' - u\rangle + \frac{L_0}{\kappa}\|u' - u\|^\kappa \ \forall u, u' \in U; \tag{16}$$

iii) $\theta \in E_v$ be such that every linear function on $E_u \times E_v$ of the form

$$[u;v] \mapsto \langle \eta, u\rangle + \langle \theta, v\rangle \tag{17}$$

with $\eta \in E_u$ attains its minimum on X at some point $x[\eta] = [u[\eta]; v[\eta]]$; we have at our disposal a *Composite Linear Minimization Oracle* (LMO) which, given on input $\eta \in E_u$, returns $x[\eta]$.

Algorithm 2 Composite Conditional Gradient Algorithm $\mathbf{CCG}(X, \phi(\cdot), \theta; \epsilon)$

Input: accuracy $\epsilon > 0$ and $\gamma_t = 2/(t+1), t = 1, 2, \ldots$
Initialize $x^1 = [u^1; v^1] \in X$ and
for $t = 1, 2, \ldots$ **do**
  Compute $\delta_t = \langle g_t, u^t - u^t[g_t]\rangle + \langle \theta, v^t - v^t[g_t]\rangle$, where $g_t = \nabla\phi(u^t)$;
  if $\delta_t \leq \epsilon$ **then**
    Return $x^t = [u^t; v^t]$
  else
    Update $x^{t+1} = [u^{t+1}; v^{t+1}] \in X$ such that $\phi^+(x^{t+1}) \leq \phi^+\left(x^t + \gamma_t(x^t[g_t] - x^t)\right)$
  end if
end for

The algorithm is outlined in Algorithm 2. Note that CCG works essentially *as if* there were no v-component at all. The CCG algorithm enjoys a convergence rate in $O(t^{-(\kappa-1)})$ in the evaluations of the function ϕ^+, and the accuracy certificates (δ_t) enjoy the same rate $O(t^{-(\kappa-1)})$ as well, for solving problems of type (15). See Appendix for details and the proof.

Proposition 3.1. *Denote D the $\|\cdot\|$-diameter of U. When solving problems of type (15), the sequence of iterates (x^t) of CCG satisfies*

$$\epsilon_t := \phi^+(x^t) - \min_{x\in X}\phi^+(x) \leq \frac{2L_0D^\kappa}{\kappa(3-\kappa)}\left(\frac{2}{t+1}\right)^{\kappa-1}, \ t \geq 2 \tag{18}$$

In addition, the accuracy certificates (δ_t) satisfy

$$\min_{1\leq s\leq t}\delta_s \leq O(1)L_0D^\kappa\left(\frac{2}{t+1}\right)^{\kappa-1}, \ t \geq 2. \tag{19}$$

3.3 Semi-Proximal Mirror-Prox for Semi-structured Variational Inequality

We now give the full description of a special class of variational inequalities, called *semi-structured variational inequalities*. This family of problems encompasses both cases that we discussed so far in Section 3.1 and 3.2. But most importantly, it also covers many other problems that do not fall into these two regimes and in particular, our essential problem of interest (3).

Semi-structured Variational Inequalities. The class of semi-structured variational inequalities allows to go beyond Assumptions (**A**.1) − (**A**.4), by assuming more structure. This structure is consistent with what we call a *semi-proximal* setup, which encompasses both the regular *proximal setup* and the regular *linear minimization setup* as special cases. Indeed, we consider a class of variational inequality VI(X, F) that satisfies, in addition to Assumptions (**A**.1) − (**A**.4), the following assumptions:

(**S**.1) *Proximal setup for X*: we assume that $E_u = E_{u_1} \times E_{u_2}$, $E_v = E_{v_1} \times E_{v_2}$, and $U \subset U_1 \times U_2$, $X = X_1 \times X_2$ with $X_i \in E_{u_i} \times E_{v_i}$ and $P_i X = \{u_i : [u_i; v_i] \in X_i\} \subset U_i$ for $i = 1, 2$, where U_1 is convex and closed, U_2 is convex and compact. We also assume that $\omega(u) = \omega_1(u_1) + \omega_2(u_2)$ and $\|u\| = \|u_1\|_{E_{u_1}} + \|u_2\|_{E_{u_2}}$, with $\omega_2(\cdot) : U_2 \to \mathbf{R}$ continuously differentiable such that

$$\omega_2(u_2') \le \omega_2(u_2) + \langle \nabla\omega_2(u_2), u_2' - u_2\rangle + \frac{L_0}{\kappa}\|u_2' - u_2\|^{\kappa}_{E_{u_2}}, \forall u_2, u_2' \in U_2;$$

for a particular $1 < \kappa \le 2$ and $L_0 < \infty$. Furthermore, we assume that the $\|\cdot\|_{E_{u_2}}$-diameter of U_2 is bounded by some $D > 0$..

(**S**.2) *Proximal mapping on X_1*: we assume that for any $\eta_1 \in E_{u_1}$ and $\alpha > 0$, we have at disposal easy-to-compute prox-mappings of the form,

$$\mathrm{Prox}_{\omega_1}(\eta_1, \alpha) := \min_{x_1=[u_1;v_1]\in X_1} \left\{\omega_1(u_1) + \langle \eta_1, u_1\rangle + \alpha\langle F_{v_1}, v_1\rangle\right\}.$$

(**S**.3) *Linear minimization on X_2*: we assume that we we have at our disposal Composite Linear Minimization Oracle (LMO), which given any input $\eta_2 \in E_{u_2}$ and $\alpha > 0$, returns an optimal solution to the minimization problem with linear form, that is,

$$\mathrm{LMO}(\eta_2, \alpha) := \min_{x_2=[u_2;v_2]\in X_2} \left\{\langle \eta_2, u_2\rangle + \alpha\langle F_{v_2}, v_2\rangle\right\}.$$

Semi-proximal setup We denote such problems as Semi-VI(X, F). On the one hand, when U_2 is a singleton, we get the *full-proximal setup*. On the other hand, when U_1 is a singleton, we get the *full linear-minimization-oracle setup* (full LMO setup). In the gray zone in between, we get the *semi-proximal setup*.

The Semi-Proximal Mirror-Prox algorithm. We finally present here our main contribution, the Semi-Proximal Mirror-Prox algorithm, which solves the semi-structured variational inequality under (**A**.1) − (**A**.4) and (**S**.1) − (**S**.3). The Semi-Proximal Mirror-Prox algorithm blends both CMP and CCG. Basically, for sub-domain X_2 given by LMO, instead of computing exactly the prox-mapping, we mimick inexactly the prox-mapping via a conditional gradient algorithm in the composite Mirror Prox algorithm. For the sub-domain X_1, we compute the prox-mapping as it is.

Course of the Semi-Proximal Mirror-Prox algorithm Basically, at step t, we first update $y_1^t = [\widehat{u}_1^t; \widehat{v}_1^t]$ by computing the exact prox-mapping and update $y_2^t = [\widehat{u}_2^t; \widehat{v}_2^t]$ by running the composite conditional gradient algorithm to problem (15) specifically with

$$X = X_2, \phi(\cdot) = \omega_2(\cdot) + \langle \gamma_t F_{u_2}(u_2^t) - \omega_2'(u_2^t), \cdot\rangle, \text{ and } \theta = \gamma_t F_{v_2},$$

until $\delta(y_2^t) = \max_{y_2\in X_2}\langle \nabla\phi^+(y_2^t), y_2^t - y_2\rangle \le \epsilon_t$. We then update $x_1^{t+1} = [u_1^{t+1}; v_1^{t+1}]$ and $x_2^{t+1} = [u_2^{t+1}; v_2^{t+1}]$ similarly except this time taking the value of the operator at point y^t. Combining the results in Theorem 3.1 and Proposition 3.1, we arrive at the following complexity bound.

Algorithm 3 Semi-Proximal Mirror-Prox Algorithm for Semi-VI(X, F)

Input: stepsizes $\gamma_t > 0$, accuracies $\epsilon_t \geq 0$, $t = 1, 2, \ldots$
[1] Initialize $x^1 = [x_1^1; x_2^1] \in X$, where $x_1^1 = [u_1^1; v_1^1]; x_2^1 = [u_2^1, ; v_2^1]$.
for $t = 1, 2, \ldots, T$ **do**
[2] Compute $y^t = [y_1^t; y_2^t]$ that

$$\begin{array}{rcl} y_1^t := [\widehat{u}_1^t; \widehat{v}_1^t] & = & \mathrm{Prox}_{\omega_1}(\gamma_t F_{u_1}(u_1^t) - \omega_1'(u_1^t), \gamma_t) \\ y_2^t := [\widehat{u}_2^t; \widehat{v}_2^t] & = & \mathbf{CCG}(X_2, \omega_2(\cdot) + \langle \gamma_t F_{u_2}(u_2^t) - \omega_2'(u_2^t), \cdot \rangle, \gamma_t F_{v_2}; \epsilon_t) \end{array}$$

[3] Compute $x^{t+1} = [x_1^{t+1}; x_2^{t+1}]$ that

$$\begin{array}{rcl} x_1^{t+1} := [u_1^{t+1}; v_1^{t+1}] & = & \mathrm{Prox}_{\omega_1}(\gamma_t F_{u_1}(\widehat{u}_1^t) - \omega_1'(u_1^t), \gamma_t) \\ x_2^{t+1} := [u_2^{t+1}; v_2^{t+1}] & = & \mathbf{CCG}(X_2, \omega_2(\cdot) + \langle \gamma_t F_{u_2}(\widehat{u}_2^t) - \omega_2'(u_2^t), \cdot \rangle, \gamma_t F_{v_2}; \epsilon_t) \end{array}$$

end for
Output: $\overline{x}_T := [\overline{u}_T; \overline{v}_T] = \left(\sum_{t=1}^T \gamma_t\right)^{-1} \sum_{t=1}^T \gamma_t y^t$

Proposition 3.2. *Under the assumption* $(\boldsymbol{A}.1) - (\boldsymbol{A}.4)$ *and* $(\boldsymbol{S}.1) - (\boldsymbol{S}.3)$ *with* $M = 0$*, for the outlined algorithm to return an* ϵ*-solution to the variational inequality* $VI(X, F)$*, the total number of Mirror Prox steps required does not exceed*

$$\textit{Total number of steps} = O\left(\frac{L\Theta[X]}{\epsilon}\right)$$

and the total number of calls to the Linear Minimization Oracle does not exceed

$$\mathcal{N} = O(1)\left(\frac{L_0 L^\kappa D^\kappa}{\epsilon^\kappa}\right)^{\frac{1}{\kappa-1}} \Theta[X].$$

In particular, if we use Euclidean proximal setup on U_2 *with* $\omega_2(\cdot) = \frac{1}{2}\|x_2\|^2$*, which leads to* $\kappa = 2$ *and* $L_0 = 1$*, then the number of LMO calls does not exceed* $\mathcal{N} = O(1)\left(L^2 D^2 (\Theta[X_1] + D^2)/\epsilon^2\right.$.

Discussion The proposed Semi-Proximal Mirror-Prox algorithm enjoys the *optimal complexity bounds*, i.e. $O(1/\epsilon^2)$, in the number of calls to LMO; see [15] for the optimal complexity bounds for general non-smooth optimisation with LMO. Furthermore, Semi-Proximal Mirror-Prox generalizes previously proposed approaches and improves upon them in special cases of problem (3); see Appendix.

4 Experiments

We present here illustrations of the proposed approach. We report the experimental results obtained with the proposed Semi-Proximal Mirror-Prox, denoted **Semi-MP** here, and state-of-the-art competing optimization algorithms. We consider three different models, all with a non-smooth loss function and a nuclear-norm regularization penalty: i) matrix completion with ℓ_2 data fidelity term; ii) robust collaborative filtering for movie recommendation; iii) link prediction for social network analysis. For i) & ii), we compare to two competing approaches: a) smoothing conditional gradient proposed in [24] (denoted Smooth-CG); b) smoothing proximal gradient ([20, 6]) equipped semi-proximal setup (Semi-SPG). For iii), we compare to Semi-LPADMM, using [22], and solving proximal mapping through conditional gradient routines. Additional experiments and implementation details are given in Appendix E.

Matrix completion on synthetic data We consider the matrix completion problem, with a nuclear-norm regularisation penalty and an ℓ_2 data-fidelity term. We first investigate the convergence patterns of our Semi-MP and Semi-SPG under two different strategies of the inexactness, a) fixed inner CG steps and

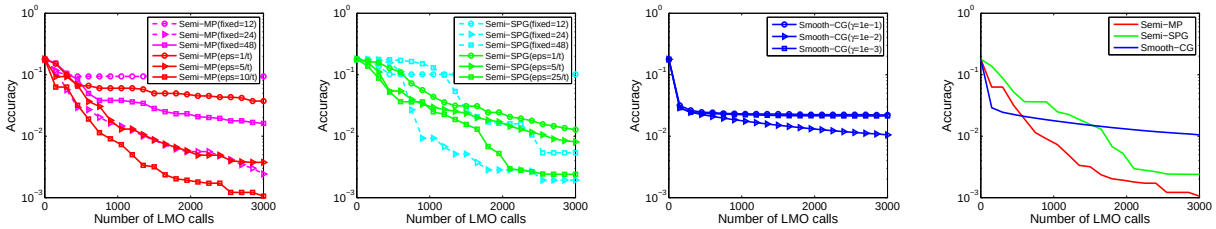

Figure 1: Matrix completion on synthetic data (1024×1024): optimality gap vs the LMO calls. From left to right: (a) Semi-MP; (b) Semi-SPG ; (c) Smooth-CG; (d) best of three algorithms.

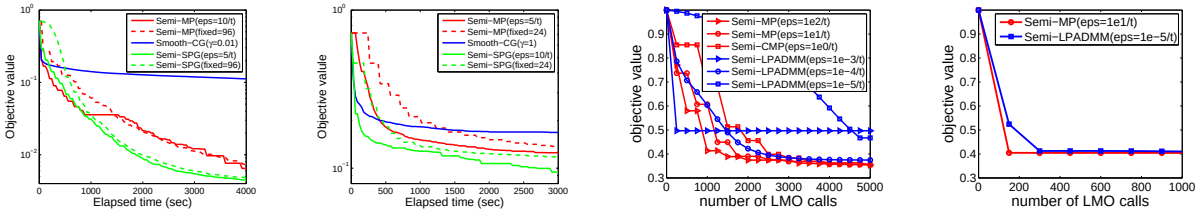

Figure 2: Robust collaborative filtering and link prediction: objective function vs elapsed time. From left to right: (a) MovieLens 100K; (b) MovieLens 1M; (c) Wikivote(1024); (d) Wikivote(full)

b) decaying $\epsilon_t = c/t$ as the theory suggested. The plots in Fig. 3 indicate that using the second strategy with $O(1/t)$ decaying inexactness provides better and more reliable performance than using fixed number of inner steps. Similar trends are observed for the Semi-SPG. One can see that these two algorithms based on inexact proximal mappings are notably faster than applying conditional gradient on the smoothed problem.

Robust collaborative filtering We consider the collaborative filtering problem, with a nuclear-norm regularisation penalty and an ℓ_1-loss function. We run the above three algorithms on the the small and medium MovieLens datasets. The small-size dataset consists of 943 users and 1682 movies with about 100K ratings, while the medium-size dataset consists of 3952 users and 6040 movies with about 1M ratings. We follow [24] to set the regularisation parameters. In Fig. 2, we can see that Semi-MP clearly outperforms Smooth-CG, while it is competitive with Semi-SPG.

Link prediction We consider now the link prediction problem, where the objective consists a hinge-loss for the empirical risk part and multiple regularization penalties, namely the ℓ_1-norm and the nuclear-norm. For this example, applying the Smooth-CG or Semi-SPG would require two smooth approximations, one for hinge loss term and one for ℓ_1 norm term. Therefore, we consider another alternative approach, Semi-LPADMM, where we apply the linearized preconditioned ADMM algorithm [22] by solving proximal mapping through conditional gradient routines. Up to our knowledge, ADMM with early stopping is not fully theoretically analysed in literature. However, from an intuitive point of view, as long as the accumulated error is controlled sufficiently, such variant of ADMM should converge.

We conduct experiments on a binary social graph data set called Wikivote, which consists of 7118 nodes and 103,747 edges. Since the computation cost of these two algorithms mainly come from the LMO calls, we present in below the performance in terms of number of LMO calls. For the first set of experiments, we select top 1024 highest degree users from Wikivote and run the two algorithms on this small dataset with different strategies for the inner LMO calls.

In Fig. 2, we observe that the Semi-MP is less sensitive to the inner accuracies of prox-mappings compared to the ADMM variant, which sometimes stops progressing if the prox-mapping of early iterations are not solved with sufficient accuracy. The results on the full dataset corroborate the fact that Semi-MP outperforms the semi-proximal variant of the ADMM algorithm.

References

[1] Francis Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 2015.

[2] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.*, 4(1):1–106, 2012.

[3] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

[4] D. P. Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, 2015.

[5] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[6] Xi Chen, Qihang Lin, Seyoung Kim, Jaime G Carbonell, and Eric P Xing. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2):719–752, 2012.

[7] Bruce Cox, Anatoli Juditsky, and Arkadi Nemirovski. Dual subgradient algorithms for large-scale nonsmooth learning problems. *Mathematical Programming*, pages 1–38, 2013.

[8] M. Dudik, Z. Harchaoui, and J. Malick. Lifted coordinate descent for learning with trace-norm regularization. *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[9] Dan Garber and Elad Hazan. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. *arXiv preprint arXiv:1301.4666*, 2013.

[10] Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, pages 1–38, 2013.

[11] E. Hazan and S. Kale. Projection-free online learning. In *ICML*, 2012.

[12] Niao He, Anatoli Juditsky, and Arkadi Nemirovski. Mirror prox algorithm for multi-term composite minimization and semi-separable problems. *arXiv preprint arXiv:1311.1098*, 2013.

[13] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, pages 427–435, 2013.

[14] Anatoli Juditsky and Arkadi Nemirovski. Solving variational inequalities with monotone operators on domains given by linear minimization oracles. *arXiv preprint arXiv:1312.107*, 2013.

[15] Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv*, 2013.

[16] Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *arXiv*, 2014.

[17] Cun Mu, Yuqian Zhang, John Wright, and Donald Goldfarb. Scalable robust matrix recovery: Frank-wolfe meets proximal methods. *arXiv preprint arXiv:1403.7588*, 2014.

[18] Arkadi Nemirovski. Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[19] Arkadi Nemirovski, Shmuel Onn, and Uriel G Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1):52–78, 2010.

[20] Y. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Math. Program.*, 110(2):245–259, 2007.

[21] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[22] Yuyuan Ouyang, Yunmei Chen, Guanghui Lan, and Eduardo Pasiliao Jr. An accelerated linearized alternating direction method of multipliers, 2014. `http://arxiv.org/abs/1401.6607`.

[23] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, pages 1–96, 2013.

[24] Federico Pierucci, Zaid Harchaoui, and Jérôme Malick. A smoothing approach for composite conditional gradient with nonsmooth loss. In *Conférence dApprentissage Automatique–Actes CAP14*, 2014.

[25] Mark Schmidt, Nicolas L. Roux, and Francis R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Adv. NIPS*. 2011.

[26] X. Zhang, Y. Yu, and D. Schuurmans. Accelerated training for matrix-norm regularization: A boosting approach. In *NIPS*, 2012.

In this Appendix, we provide additional material on variational inequalities and non-smooth optimisation algorithms, give the proofs on the main theorems, and provide additional information regarding the competing algorithms based on smoothing techniques and the implementation details for different models.

A Preliminaries: Variational Inequalities and Accuracy Certificates

For the reader's convenience, we recall here the relationship between variational inequalities, accuracy certificates, and execution protocols, for non-smooth optimization algorithms. The exposition below is directly taken from [12], and recalled here for the reader's convenience.

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 $x_t \in X$, $t = 1, 2, ...$, 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* $X' \neq \emptyset$ of X given by

$$\mathrm{Res}(X'|\mathcal{I}_t, \lambda^t) = \sup_{x \in X'} \sum_{\tau=1}^t \lambda_\tau^t \langle F(x_\tau), x_\tau - x \rangle. \tag{20}$$

The role of those notions for non-smooth optimization is explained below.

Variational inequalities. Assume that F is *monotone*, i.e.,VI(X,F)

$$\langle F(x) - F(y), x - y \rangle \geq 0, \ \ \forall x, y \in X \ . \tag{21}$$

Our goal is to approximate a weak solution to the variational inequality (v.i.) $\mathrm{VI}(X, F)$ associated with (X, F). A weak solution is defined as a point $x_* \in X$ such that

$$\langle F(y), y - x_* \rangle \geq 0 \ \forall y \in X. \tag{22}$$

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

$$\epsilon_{\mathrm{VI}}(x|X, F) = \sup_{y \in X} \langle F(y), x - y \rangle \tag{23}$$

This inaccuracy is a convex nonnegative function which vanishes exactly at the set of weak solutions to the $\mathrm{VI}(X, F)$.

Proposition A.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*

$$\epsilon_{\mathrm{VI}}(x^t|X', F) \leq \mathrm{Res}(X'|\mathcal{I}_t, \lambda^t). \tag{24}$$

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

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

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

$$\mathrm{SadVal} = \min_{x^1 \in X_1} \max_{x^2 \in X_2} \Phi(x^1, x^2), \tag{25}$$

two induced convex optimization problems

$$\begin{array}{rcll} \mathrm{Opt}(P) & = & \min\limits_{x^1 \in X_1} \left[\overline{\Phi}(x^1) = \sup\limits_{x^2 \in X_2} \Phi(x^1, x^2)\right] & (P) \\ \mathrm{Opt}(D) & = & \max\limits_{x^2 \in X_2} \left[\underline{\Phi}(x^2) = \inf\limits_{x^1 \in X_1} \Phi(x^1, x^2)\right] & (D) \end{array} \tag{26}$$

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

$$\begin{array}{rcl} \epsilon_{\mathrm{Sad}}(x \big| X_1, X_2, \Phi) & = & \overline{\Phi}(x^1) - \underline{\Phi}(x^2) \\ & = & [\overline{\Phi}(x^1) - \mathrm{Opt}(P)] + [\mathrm{Opt}(D) - \underline{\Phi}(x^2)] + \underbrace{[\mathrm{Opt}(P) - \mathrm{Opt}(D)]}_{\geq 0} \end{array} \tag{27}$$

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 A.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

$$x^t \in X' := X'_1 \times X'_2. \tag{28}$$

Then

$$\epsilon_{\mathrm{Sad}}(x^t \big| X'_1, X'_2, \Phi) = \sup_{x^2 \in X'_2} \Phi(x^{1,t}, x^2) - \inf_{x^1 \in X'_1} \Phi(x^1, x^{2,t}) \leq \mathrm{Res}(X' \big| \mathcal{I}_t, \lambda^t). \tag{29}$$

In addition, setting $\widetilde{\Phi}(x^1) = \sup_{x^2 \in X'_2} \Phi(x^1, x^2)$, for every $\bar{x}^1 \in X'_1$ we have

$$\widetilde{\Phi}(x^{1,t}) - \widetilde{\Phi}(\bar{x}^1) \leq \widetilde{\Phi}(x^{1,t}) - \Phi(\bar{x}^1, x^{2,t}) \leq \mathrm{Res}(\{\bar{x}^1\} \times X'_2 \big| \mathcal{I}_t, \lambda^t). \tag{30}$$

In particular, when the problem $\mathrm{Opt} = \min_{x^1 \in X'_1} \widetilde{\Phi}(x^1)$ is solvable with an optimal solution x^1_*, we have

$$\widetilde{\Phi}(x^{1,t}) - \mathrm{Opt} \leq \mathrm{Res}(\{x^1_*\} \times X'_2 \big| \mathcal{I}_t, \lambda^t). \tag{31}$$

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

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

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

$$\sup_{[p;q]\in Y} \left[\Phi(x^{1,t}, q) - \Phi(p, x^{2,t})\right] \leq \mathrm{Res}(Y|\mathcal{I}_t, \lambda^t). \tag{32}$$

Now assume that Condition (28) is satisfied. Setting $Y = X' := X_1' \times X_2'$, and recalling what ϵ_{Sad} is, (32) yields (29). With $Y = \{\bar{x}^1\} \times X_2'$ (32) yields the second inequality in (30); the first inequality in (30) is clear since $x^{2,t} \in X_2'$. □

B Theoretical analysis of composite Mirror Prox with inexact proximal mappings

We restate the Theorem 3.1 below and the proof below. The theoretical convergence rate established in Theorem 3.1 and Corollary 3.1 extends the previous result established in Corollary 3.1 in [12] for CMP with exact prox-mappings. Indeed, when exact prox-mappings are used, we recover the result of [12]. When inexact prox-mappings are used, the errors due to the inexactness of the prox-mappings accumulates and is reflected in the bound (34) and (14).

Theorem 3.1. *Assume that the sequence of step-sizes (γ_t) in the CMP algorithm satisfy*

$$\sigma_t := \gamma_t \langle F_u(\widehat{u}^t) - F_u(u^t), \widehat{u}^t - u^{t+1}\rangle - V_{\widehat{u}^t}(u^{t+1}) - V_{u^t}(\widehat{u}^t) \leq \gamma_t^2 M^2, \quad t = 1, 2, \ldots, T. \tag{33}$$

Then, denoting $\Theta[X] = \sup_{[u;v]\in X} V_{u^1}(u)$, for a sequence of inexact prox-mappings with inexactness $\epsilon_t \geq 0$, we have

$$\epsilon_{\mathrm{VI}}(\bar{x}_T|X, F) := \sup_{x\in X} \langle F(x), \bar{x}_T - x\rangle \leq \frac{\Theta[X] + M^2 \sum_{t=1}^T \gamma_t^2 + 2\sum_{t=1}^T \epsilon_t}{\sum_{t=1}^T \gamma_t}. \tag{34}$$

Remarks Note that the assumption on the sequence of step-sizes (γ_t) is clearly satisfied when $\gamma_t \leq (\sqrt{2}L)^{-1}$. When $M = 0$, it is satisfied as long as $\gamma_t \leq L^{-1}$.

Proof. The proofs builds upon and extends the proof in [12]. For all $u, u', w \in U$, we have the well-known identity

$$\langle V_u'(u'), w - u'\rangle = V_u(w) - V_{u'}(w) - V_u(u'). \tag{35}$$

Indeed, the right hand side writes as

$$\begin{array}{rl}& [\omega(w) - \omega(u) - \langle \omega'(u), w - u\rangle] - [\omega(w) - \omega(u') - \langle \omega'(u'), w - u'\rangle] - [\omega(u') - \omega(u) - \langle \omega'(u), u' - u\rangle] \\ = & \quad \langle \omega'(u), u - w\rangle + \langle \omega'(u), u' - u\rangle + \langle \omega'(u'), w - u'\rangle = \langle \omega'(u') - \omega'(u), w - u'\rangle = \langle V_u'(u'), w - u'\rangle.\end{array}$$

For $x = [u; v] \in X$, $\xi = [\eta; \zeta]$, $\epsilon \geq 0$, let $[u'; v'] \in P_x^\epsilon(\xi)$. By definition, for all $[s; w] \in X$, the inequality holds

$$\langle \eta + V_u'(u'), u' - s\rangle + \langle \zeta, v' - w\rangle \leq \epsilon,$$

which by (35) implies that

$$\langle \eta, u' - s\rangle + \langle \zeta, v' - w\rangle \leq \langle V_u'(u'), s - u'\rangle + \epsilon = V_u(s) - V_{u'}(s) - V_u(u') + \epsilon. \tag{36}$$

When applying (36) with $\epsilon = \epsilon_t$, $[u;v] = [u^t;v^t] = x^t$, $\xi = \gamma_t F(x^t) = [\gamma_t F_u(u^t); \gamma_t F_v]$, $[u';v'] = [\widehat{u}^t;\widehat{v}^t] = y^t$, and $[s;w] = [u^{t+1};v^{t+1}] = x^{t+1}$ we obtain

$$\gamma_t[\langle F_u(u^t), \widehat{u}^t - u^{t+1}\rangle + \langle F_v, \widehat{v}^t - v^{t+1}\rangle] \le V_{u^t}(u^{t+1}) - V_{\widehat{u}^t}(u^{t+1}) - V_{u^t}(\widehat{u}^t) + \epsilon_t \; ; \tag{37}$$

and applying (36) with $\epsilon = \epsilon_t$, $[u;v] = x^t$, $\xi = \gamma_t F(y^t)$, $[u';v'] = x^{t+1}$, and $[s;w] = z \in X$ we get

$$\gamma_t[\langle F_u(\widehat{u}^t), u^{t+1} - s\rangle + \langle F_v, v^{t+1} - w\rangle] \le V_{u^t}(s) - V_{u^{t+1}}(s) - V_{u^t}(u^{t+1}) + \epsilon_t \; . \tag{38}$$

Adding (38) to (37), we obtain for every $z = [s;w] \in X$

$$\begin{aligned} \gamma_t\langle F(y^t), y^t - z\rangle &= \gamma_t[\langle F_u(\widehat{u}^t), \widehat{u}^t - s\rangle + \langle F_v, \widehat{v}^t - w\rangle] \\ &\le V_{u^t}(s) - V_{u^{t+1}}(s) + \sigma_t + 2\epsilon_t \; , \end{aligned} \tag{39}$$

with

$$\sigma_t := \gamma_t\langle F_u(\widehat{u}^t) - F_u(u^t), \widehat{u}^t - u^{t+1}\rangle - V_{\widehat{u}^t}(u^{t+1}) - V_{u^t}(\widehat{u}_t) \; .$$

Due to the strong convexity, with modulus 1, of $V_u(\cdot)$ w.r.t. $\|\cdot\|$, we have for all $u, \widehat{u}$

$$V_u(\widehat{u}) \ge \frac{1}{2}\|u - \widehat{u}\|^2 \; .$$

Therefore,

$$\begin{array}{rcl} \sigma_t & \le & \gamma_t\|F_u(\widehat{u}^t) - F_u(u^t)\|_*\|\widehat{u}^t - u^{t+1}\| - \frac{1}{2}\|\widehat{u}^t - u^{t+1}\|^2 - \frac{1}{2}\|u^t - \widehat{u}^t\|^2 \\ & \le & \frac{1}{2}\left[\gamma_t^2\|F_u(\widehat{u}^t) - F_u(u^t)\|_*^2 - \|u^t - \widehat{u}^t\|^2\right] \\ & \le & \frac{1}{2}\left[\gamma_t^2[M + L\|\widehat{u}^t - u^t\|]^2 - \|u^t - \widehat{u}^t\|^2\right] , \end{array}$$

where the last inequality follows from Assumption **A**.3. Note that $\gamma_t L < 1$ implies that

$$\gamma_t^2[M + L\|\widehat{u}^t - u^t\|]^2 - \|\widehat{u}^t - u^t\|^2 \le \max_r \left[\gamma_t^2[M + Lr]^2 - r^2\right] = \frac{\gamma_t^2 M^2}{1 - \gamma_t^2 L^2}.$$

Let us assume that the step-sizes $\gamma_t > 0$ are chosen so that (33) holds, that is $\sigma_t \le \gamma_t^2 M^2$. It is indeed the case when $0 < \gamma_t \le \frac{1}{\sqrt{2}L}$; when $M = 0$, we can take also $\gamma_t \le \frac{1}{L}$. Summing up inequalities (39) over $t = 1, 2, ..., t$, and taking into account that $V_{u^{t+1}}(s) \ge 0$, we finally conclude that for all $z = [s;w] \in X$,

$$\sum_{t=1}^T \lambda_T^t\langle F(y^t), y^t - z\rangle \le \frac{V_{u^1}(s) + M^2\sum_{t=1}^T \gamma_t^2 + 2\sum_{t=1}^T \epsilon_t}{\sum_{t=1}^T \gamma_t}, \text{ where } \lambda_T^t = (\sum_{i=1}^T \gamma_i)^{-1}\gamma_t \; .$$

□

C Theoretical analysis of composite conditional gradient

C.1 Convergence rate

The CCG algorithm enjoys a convergence rate in $O(t^{-(\kappa-1)})$ in the evaluations of the function ϕ^+, and the accuracy certificates (δ_t) enjoy the same rate $O(t^{-(\kappa-1)})$ as well, for solving problems of type (15).

Proposition 3.1. *Denote D the $\|\cdot\|$-diameter of U. When solving problems of type (15), the sequence of iterates (x^t) of CCG satisfies*

$$\epsilon_t := \phi^+(x^t) - \min_{x\in X}\phi^+(x) \le \frac{2L_0D^\kappa}{\kappa(3-\kappa)}\left(\frac{2}{t+1}\right)^{\kappa-1}, \; t \ge 2 \tag{40}$$

In addition, the accuracy certificates (δ_t) satisfy

$$\min_{1\le s\le t} \delta_s \le O(1)L_0D^\kappa\left(\frac{2}{t+1}\right)^{\kappa-1}, \; t \ge 2 \tag{41}$$

C.2 Proof of Proposition 3.1

1^0. The projection of X_2 onto E_{u_2} is contained in U_2, whence

$$\|u_2[\nabla\phi(u_2^s)] - u_2^s\| \le D.$$

This observation, due to the structure of ϕ^+, implies that whenever $x, x' \in X$ and $\gamma \in [0,1]$, we have

$$\phi^+(x+\gamma(x^+-x)) \le \phi^+(x) + \gamma\langle\nabla\phi^+(x), x'-x\rangle + \frac{L_0D^\kappa}{\kappa}\gamma^\kappa. \tag{42}$$

Setting $x_+^s = x_2^s + \gamma_s(x_2[\nabla\phi(u^s)] - x_2^s)$ and $\gamma_s 2/(s+1)$, we have

$$\delta_{t+1} \le \phi^+(x_+^s) - \min_{x_2\in X_2}\phi^+(x_2) \tag{43}$$

$$\le \delta_s + \gamma_s\langle\nabla\phi(x_2^s), x[\nabla\phi^+(x_2^s)] - x_2\rangle + \frac{L_0D^\kappa}{\kappa}\gamma_s^\kappa \tag{44}$$

$$= \delta_s - \gamma_s\Delta^s + \frac{L_0D^\kappa}{\kappa}\gamma_s^\kappa, \tag{45}$$

whence, due to $\Delta_s \ge \delta_s \ge 0$,

$$\begin{aligned} &(i) \quad \delta_{t+1} \le (1-\gamma_s)\delta_s + \frac{L_0D^\kappa}{\kappa}\gamma_s^\kappa,\ s = 1,2,..., \\ &(ii) \quad \gamma_\tau\Delta_\tau \le \delta_\tau - \delta_{\tau+1} + \frac{L_0D^\kappa}{\kappa}\gamma_\tau^\kappa,\ \tau = 1,2,... \end{aligned} \tag{46}$$

2^0. Let us prove (40) by induction on $s \ge 2$. By (46.i) and due to $\gamma_1 = 1$ we have $\delta_2 \le \frac{L_0D^\kappa}{\kappa}$, whence $\delta_2 \le \frac{2L_0D^\kappa}{\kappa(3-\kappa)}\gamma_2^{\kappa-1}$ due to $\gamma_2 = 2/3$ and $1 < \kappa \le 2$. Now assume that $\delta_s \le \frac{2L_0D^\kappa}{\kappa(3-\kappa)}\gamma_s^{\kappa-1}$ for some $t \ge 2$. Then, invoking (46.i),

$$\begin{aligned} \delta_{s+1} &\le \frac{2L_0D^\kappa}{\kappa(3-\kappa)}\gamma_s^{\kappa-1}(1-\gamma_s) + \frac{L_0D^\kappa}{\kappa}\gamma_s^\kappa \\ &\le \frac{2L_0D^\kappa}{\kappa(3-\kappa)}\left[\gamma_s^{\kappa-1} - \frac{\kappa-1}{2}\gamma_s^\kappa\right] \\ &\le \frac{2L_0D^\kappa}{\kappa(3-\kappa)}2^{\kappa-1}\left[(t+1)^{1-\kappa} + (1-\kappa)(t+1)^{-\kappa}\right] \end{aligned}$$

Therefore, by convexity of $(t+1)^{1-\kappa}$ in t

$$\delta_{s+1} \le \frac{2L_0D^\kappa}{\kappa(3-\kappa)}2^{\kappa-1}(t+2)^{1-\kappa} = \frac{2L_0D^\kappa}{\kappa(3-\kappa)}\gamma_{t+1}^{\kappa-1}$$

The induction is completed.

3^0. To prove (41), given $s \ge 2$, let $s_- = \mathrm{Ceil}(\max[2, s/2])$. Summing up inequalities (46.ii) over $s_- \le \tau \le s$, we get

$$\left(\min_{\tau\le s}\Delta_\tau\right)\sum\nolimits_{\tau=s_-}^s\gamma_\tau \le \sum_{\tau=s_-}^s\gamma_\tau\Delta_\tau \le \delta_{s_-} - \delta_{s+1} + \frac{L_0D^\kappa}{2}\sum\nolimits_{\tau=s_-}^s\gamma_\tau^\kappa \le O(1)L_0D^\kappa\gamma_s^{\kappa-1}$$

and $\sum_{\tau=s_-}^s\gamma_\tau \ge O(1)$, and (41) follows. □

D Semi-Proximal Mirror-Prox

D.1 Theoretical analysis for Semi-Proximal Mirror-Prox

We first restate Proposition 3.2 and provide the proof below.

Proposition 3.2. *Under the assumption* $(\boldsymbol{A}.1) - (\boldsymbol{A}.4)$ *and* $(\boldsymbol{S}.1) - (\boldsymbol{S}.3)$ *with* $M = 0$*, for the outlined algorithm to return an* ϵ*-solution to the variational inequality* $VI(X, F)$*, the total number of Mirror Prox steps required does not exceed* $O\left(\frac{L\Theta[X]}{\epsilon}\right)$*, and the total number of calls to the Linear Minimization Oracle does not exceed*

$$\mathcal{N} = O(1)\left(\frac{L_0 L^\kappa D^\kappa}{\epsilon^\kappa}\right)^{\frac{1}{\kappa-1}} \Theta[X].$$

In particular, if we use Euclidean proximal setup on U_2 *with* $\omega_2(\cdot) = \frac{1}{2}\|x_2\|^2$*, which leads to* $\kappa = 2$ *and* $L_0 = 1$*, then the number of LMO calls does not exceed* $\mathcal{N} = O(1)\left(L^2 D^2 (\Theta[X_1] + D^2)\right)/\epsilon^2$*.*

Proof. Let us fix N as the number of Mirror prox steps, and since $M = 0$, from Theorem 3.1, the efficiency estimate of the variational inequality implies that

$$\epsilon_{\mathrm{VI}}(\bar{x}^N | X, F) \leq \frac{L(\Theta[X] + 2\sum_{t=1}^N \epsilon_t)}{N}.$$

Let us fix $\epsilon_t = \frac{2\Theta[X]}{N}$ for each $t = 1, \ldots, N$, then from Proposition 3.1, it takes at most $s = O(1)(\frac{L_0 D^\kappa N}{\Theta[X]})^{1/(\kappa-1)}$ LMO oracles to generate a point such that $\Delta_s \leq \epsilon_t$. Moreover, we have

$$\epsilon_{\mathrm{VI}}(\bar{x}^N | X, F) \leq 2\frac{L\Theta[X]}{N}.$$

Therefore, to ensure $\epsilon_{\mathrm{VI}}(\bar{x}^N | X, F) \leq \epsilon$ for a given accuracy $\epsilon > 0$, the number of Mirror Prox steps N is at most $O(\frac{L\Theta[X]}{\epsilon})$ and the number of LMO calls on X_2 needed is at most

$$\mathcal{N} = O(1)\Big(\frac{L_0 D^\kappa N}{\Theta[X]}\Big)^{1/(\kappa-1)} \cdot N = O(1)\Big(\frac{L_0 L^\kappa D^\kappa}{\epsilon^\kappa}\Big)^{1/(\kappa-1)} \Theta[X].$$

In particular, if $\kappa = 2$ and $L_0 = 1$, this quantity can be reduced to

$$\mathcal{N} = O(1)\frac{L^2 D^2 \Theta[X]}{\epsilon^2}.$$

□

D.2 Discussion of Semi-Proximal Mirror-Prox

The proposed Semi-Proximal Mirror-Prox algorithm enjoys the *optimal complexity bounds*, i.e. $O(1/\epsilon^2)$, in the number of calls to linear minimization oracle. Furthermore, Semi-Proximal Mirror-Prox generalizes previously proposed approaches and improves upon them in special cases of problem (3).

When there is no regularisation penalty, Semi-Proximal Mirror-Prox is more general than previous algorithms for solving the corresponding constrained non-smooth optimisation problem. Semi-Proximal Mirror-Prox does not require assumptions on favorable geometry of dual domains Z or simplicity of $\psi(\cdot)$ in (2). When the regularisation is simply a norm (with no operator in front of the argument), Semi-Proximal Mirror-Prox is competitive with previously proposed approaches [16, 24] based on smoothing techniques.

When the regularisation penalty is non-trivial, Semi-Proximal Mirror-Prox is the first proximal-free or conditional-gradient-type optimization algorithm, up to our knowledge.

E Numerical experiments and implementation details

E.1 Matrix completion: ℓ_2-fit +nuclear norm

We first consider the the following type of matrix completion problem,

$$\min_{x\in\mathbf{R}^{m\times n}} \|P_\Omega x - b\|_2 + \lambda\|x\|_{\text{nuc}} \tag{47}$$

where $\|\cdot\|_{\text{nuc}}$ stands for the nuclear norm and $P_\Omega x$ is the restriction of x onto the cells Ω.

Competing algorithms. We compare the following three candidate algorithms, i) Semi-Proximal Mirror-Prox (**Semi-MP**) ; ii) conditional gradient after smoothing (**Smooth-CG**); iii) inexact accelerate proximal gradient after smoothing (**Semi-SPG**). We provide below the key steps of each algorithms.

1. **Semi-MP**: this is shorted for our Semi-Proximal Mirror-Prox algorithm, we solve the saddle point reformulation given by
$$\min_{x,v:\|x\|_{\text{nuc}}\le v} \max_{\|y\|_2\le 1} \langle P_\Omega x - b, y\rangle + \lambda v \tag{48}$$
which is equivalent as to the semi-structured variational inequality Semi-VI (X, F) with $X = \{[u = (x; y); v] : \|x\|_{\text{nuc}} \le v, \|y\|_2 \le 1\}$ and $F = [F_u(u); F_v] = [P_\Omega^T y; b - P_\Omega x; \lambda]$. The subdomain $X_1 = \{y : \|y\|_2 \le 1\}$ is given by full-prox setup and the subdomain $X_2 = \{(x; v) : \|x\|_{\text{nuc}} \le v\}$ is given by LMO. By setting both the distance generating functions $\omega_x(x)$ and $\omega_y(y)$ as the Euclidean distance, the update of y reduces to a gradient step, and the update of x follows the composite conditional gradient routine over a simple quadratic problem.

2. **Smooth-CG**: The algorithm ([24]) directly applies the generalized composite conditional gradient on the following smoothed problem using the Nesterov smoothing technique,
$$\min_{x,v:\|x\|_{\text{nuc}}\le v} f^\gamma(x) + \lambda v, \text{ where } f^\gamma(x) = \max_{\|y\|_2\le 1} \{\langle P_\Omega x - b, y\rangle - \frac{\gamma}{2}\|y\|_2^2\}. \tag{49}$$
Under the full memory version, the update of x at step t requires computing reoptimization problem
$$\min_{\theta_1,\ldots,\theta_t} f^\gamma(\sum_{i=1}^t \theta_i u_i v_i^T) + \lambda \sum_{i=1}^t \theta_i \tag{50}$$
where $\{u_i, v_i\}_{i=1}^t$ are the singular vectors collected from the linear minimization oracles. Same as suggested in [24], we use the quasi-Newton solver L-BFGS-B [5] to solve the above re-optimization subproblem. Notice that in this situation, solving (50) can be relatively efficient even for large t since computing the gradient of the objective in (50) does not necessarily need to compute out the full matrix representation of $x = \sum_{i=1}^t \theta_i u_i v_i^T$.

3. **Semi-SPG**: The approach is to apply the accelerated proximal gradient to the smoothed composite model as in (49) and approximately solve the proximal mappings via conditional gradient routines. In fact, Semi-SPG can be considered as a direct extension of the conditional gradient sliding to the composite setting. Same as Semi-MP, the update of x is given by the composite conditional gradient routine over a simple quadratic problem and additional interpolation step. Since the Lipschitz constant is not known, the learning rate is selected through backtracking.

For Semi-MP and Semi-SPG, we test two different strategies for the inexact prox-mappings, a)fixed inner CG steps and b)decaying $\epsilon_t = c/t$ as the theory suggested. For the sake of simplicity, we generate the synthetic data such that the magnitudes of the constant factors (i.e. Frobenius norm and nuclear norm of optimal solution) are approximately of order 1, which means the convergence rate is dominated mainly by the number of LMO calls. In Fig. 3, we evaluate the optimality gap of these algorithms with different parameters

(e.g. number of inner steps, scaling factor c, smoothness parameter γ) and compare their performance given the best-tuned parameter. As the plot shows, the Semi-MP algorithm generates a solution with $\epsilon = 10^{-3}$ accuracy within about 3000 LMO calls, which is not bad at all given the fact that the worst complexity is $O(1/\epsilon^2)$. Also, the plots indicate that using the second strategy with $O(1/t)$ decaying inexactness provides better and more reliable performance than using fixed number of inner steps. Similar trends are observed for the Semi-SPG. One can see that these two algorithms based on inexact proximal mappings are notably faster than applying conditional gradient on the smoothed problem. Moreover, since the Smooth-CG requires additional computation and memory cost for the re-optimization procedure, the actual difference in terms of CPU time could be more significant.

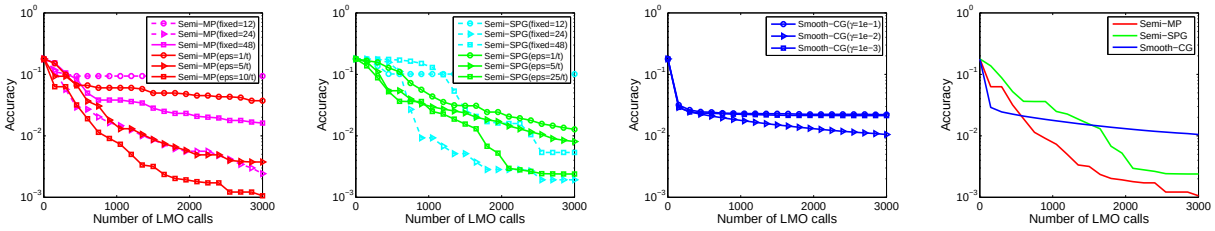

Figure 3: Matrix completion on synthetic data(1024×1024): optimality gap vs the LMO calls. From left to right: (a) Semi-MP; (b) Semi-SPG ; (c) Smooth-CG; (d) best of three algorithms.

E.2 Robust collaborative fitering: ℓ_1-empirical risk +nuclear norm

We consider the collaborative filtering problem, with a nuclear-norm regularisation penalty and an ℓ_1-empirical risk function:

$$\min_x \frac{1}{|E|} \sum_{(i,j)\in E} |x_{ij} - b_{ij}| + \lambda \|x\|_{\text{nuc}}. \tag{51}$$

Competing algorithms. We compare the above three candidate algorithm. The smoothed problem for Semi-SPG and Smooth-CG in this case becomes

$$\min_{x,v:\|x\|_{\text{nuc}}\le v} f^\gamma(x) + \lambda v, \text{ where } f^\gamma(x) = \max_{\|y\|_\infty \le 1} \left\{ \frac{1}{|E|} \sum_{(i,j)\in E} (x_{ij} - b_{ij}) y_{ij} - \frac{\gamma}{2}\|y\|_2^2 \right\}. \tag{52}$$

Note that in this case, for Smooth-CG, solving the re-optimization problem in (50) at each iteration requires computing the full matrix representation for the gradient. For large t and large-scale problems, the computation cost for re-optimization is no longer negligible. However, the Semi-MP and Semi-SPG do not suffer from this limitation since the conditional gradient routines are called for simple quadratic subproblems. For this particular example, we implement the Semi-MP slightly different from the above scheme. We solve the following saddle point reformulation with properly selected ρ,

$$\min_{\substack{x,y,v_1,v_2:\\ v_1\ge\|x\|_{\text{nuc}}, v_2\ge\|y\|_1}} \max_{\|w\|_2\le 1} v_2 + \lambda v_1 + \rho\langle \mathcal{A}x - b - y, w\rangle \tag{53}$$

where we use $\mathcal{A}$ to denote the operator $\frac{1}{|E|}P_E$. The semi-structured variational inequality Semi-VI (X, F) associated with the above saddle point problem is given by $X = \{[u = (x, y, w); v = (v_1.v_2)] : \|x\|_{\text{nuc}} \le v_1, \|y\|_1 \le v_2, \|w\|_2 \le 1\}$ and $F = [F_u(u); F_v] = [\rho\mathcal{A}w; -\rho w; \rho(y - \mathcal{A}x + b); \lambda; 1]$. The subdomain $X_1 = \{(y, w, v_2) : \|y\|_1 \le v_2, \|w\|_2 \le 1\}$ is given by full-prox setup and the subdomain $X_2 = \{(x; v_1) : \|x\|_{\text{nuc}} \le v_1\}$ is given by LMO. By setting both the distance generating functions as the Euclidean distance, the update of w reduces to the gradient step, the update of y reduces to the soft-thresholding operator, and the update

of x is given by the composite conditonal gradient routine. In our experiment, the factor ρ is updated adaptively in such a way that the back-projection step does not increase the objective function value. We set the stepsizes γ_t along the iterations using line-search. All in all, the Semi-Proximal Mirror-Prox algorithm (Semi-MP) is fully automatic, and does not require tuning of any parameter.

We run the above three algorithms on the the small and medium MovieLens datasets. The small-size dataset consists of 943 users and 1682 movies with about 100K ratings,while the medium-size dataset consists of 3952 users and 6040 movies with about 1M ratings. We follow [24] to set the regularisation parameters. We randomly pick 80% of the entries to build the training dataset, and compute the normalized mean absolute error (NMAE) on the remaining test dataset. For Smooth-CG, we carry out the algorithm with different smoothing parameters, ranging from $\{1e-3, 1e-2, 1e-1, 1e0\}$ and select the one with the best performance. For the Semi-SPG algorithm, we adopt the best smoothing parameter found in Smooth-CG. We use two different strategies to control the number of LMO calls at each iteration, i.e. the accuracy of the proximal mapping for both Semi-SPG and Semi-MP, which are a) fixed inner CG steps and b) decaying $\epsilon_t = c/t$ as the theory suggested. We report in Fig. 4 and Fig. 5 the performance of each algorithm under different choice of parameters and the overall comparison of objective value and NMAE on test data in Fig. 6.

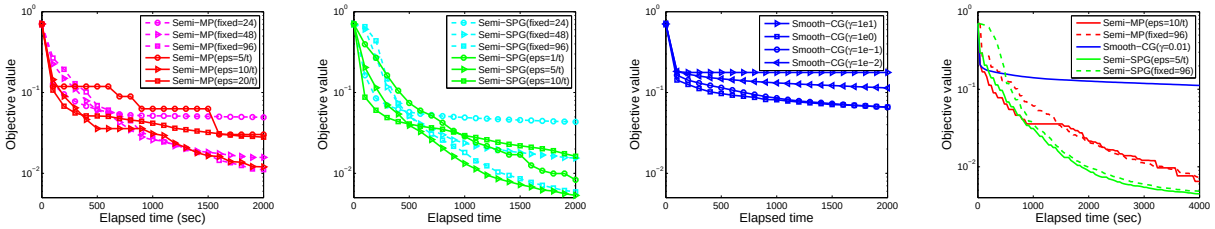

Figure 4: Robust collaborative filtering on MovieLens 100K: objective function vs elapsed time. From left to right: (a) Semi-MP; (b) Semi-SPG ; (c) Smooth-CG; (d) best of three algorithms.

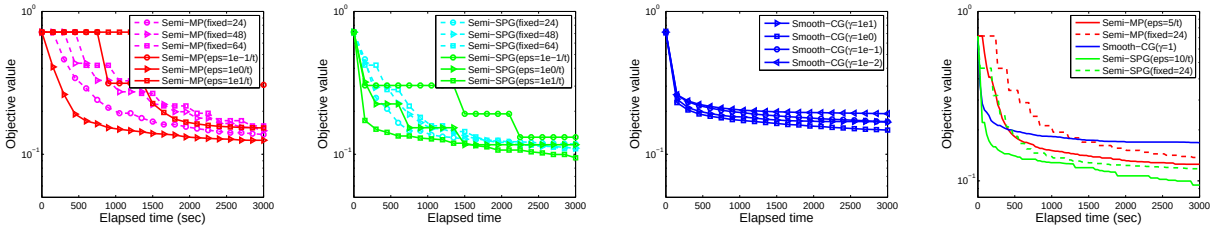

Figure 5: Robust collaborative filtering on MovieLens 1M: objective function vs elasped time. From left to right: (a) Semi-MP; (b) Semi-SPG ; (c) Smooth-CG; (d) best of three algorithms.

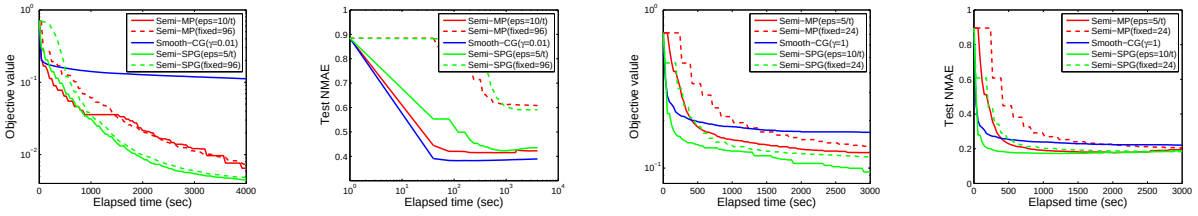

Figure 6: Robust collaborative filtering on Movie Lens: objective function and test NMAE against elapsed time. From left to right: (a) MovieLens 100K objective; (b) MovieLens 100K test NMAE; (c) MovieLens 1M objective; (d) MovieLens 1M test NMAE.

In Fig. 4 and Fig. 5, we can see that using fixed inner CG steps sometimes achieve comparable performance

as using the decaying epsilon ϵ_t. In Fig. 6, we can see that Semi-MP clearly outperforms Smooth-CG, while it is competitive with Semi-SPG. In the large-scale setting, Semi-MP achieves better objective as well as test NMAE compared to Smooth-CG.

E.3 Link prediction: hinge loss + ℓ_1-norm + nuclear norm

We consider the following model for the link prediction problem,

$$\min_{x \in \mathbf{R}^{m \times n}} \frac{1}{|E|} \sum_{(i,j) \in E} \max\left(1 - (b_{ij} - 0.5)x_{ij}, 0\right) + \lambda_1 \|x\|_1 + \lambda_2 \|x\|_{\text{nuc}} \tag{54}$$

This example is more complicated than the previous two examples since it has not only one nonsmooth loss function but also two regularization terms. Applying the smoothing-CG or Semi-SPG would require to build two smooth approximations, one for hinge loss term and one for ℓ_1 norm term. Therefore, we consider another alternative approach, Semi-LPADMM, where we apply the linearized preconditioned ADMM algorithm by solving proximal mapping through conditional gradient routines. Up to our knowledge, ADMM with early stopping is not well-analyzed in literature, but intuitively as long as the accumulated error is controlled sufficiently, the variant will converge.

We conduct experiments on a binary social graph data set called Wikivote, which consists of 7118 nodes and 103,747 edges. Since the computation cost of these two algorithms mainly come from the LMO calls, we present in below the performance in terms of number of LMO calls. For the first set of experiments, we select top 1024 highest degree users from Wikivote and run the two algorithms on this small dataset with different strategies for the inner LMO calls.

In Fig. 7, we observe that the Semi-MP is less sensitive to the inner accuracies of prox-mappings compared to the ADMM variant, which sometimes stop progressing if the prox mapping of early iterations are not solved with sufficient accuracy. Another observation is that in this example, the second strategy, which essentially saves the use of LMOs, works better in the long run than using fixed number of LMOs. The results indicate again on the full dataset again indicates that our algorithm performs better than the semi-proximal variant of ADMM algorithm.

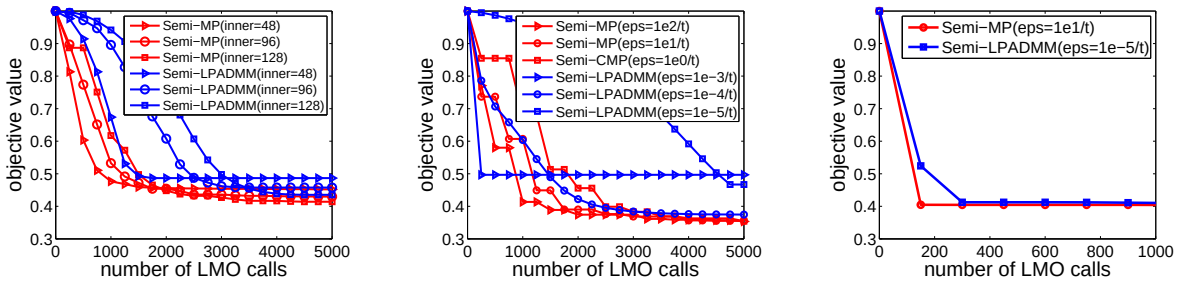

Figure 7: Link prediction on Wikivote: objective function value against the LMO calls. From left to right: (a)Wikivote(1024) with fixed inner steps; (b) Wikivote(1024) with $\epsilon_t = c/t$; (c) Wikivote(full)