Semi-Proximal Mirror-Prox for Nonsmooth Composite Minimization

Authored by:

Zaid Harchaoui
Niao He

Abstract

We propose a new first-order optimization algorithm to solve high-dimensional non-smooth composite minimization problems. Typical examples of such problems have an objective that decomposes into a non-smooth empirical risk part and a non-smooth regularization penalty. The proposed algorithm, called Semi-Proximal Mirror-Prox, leverages the saddle point 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 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 Paper Body

A wide range of machine learning and signal processing problems can be formulated as the minimization of a composite objective: min F (x) := f (x) + kBxk

x?X

(1)

where X is closed and convex, f is convex and can be either smooth, or nonsmooth yet enjoys a particular structure. The term kBxk defines a regularization penalty through a norm k ? k, and x 7? Bx a linear mapping on a closed convex set X. In many situations, the objective function F of interest enjoys a favorable structure, namely a socalled saddle point representation [6, 11, 13]: f (x) = max {hx, Azi ? ?(z)} z?Z

(2)

where Z is convex compact subset of a Euclidean space, and ?(?) is a convex function. Sec. 4 will give several examples of such situations. Saddle point representations can then be leveraged to use first-order optimization algorithms. The simple first option to minimize F is using the so-called Nesterov smoothing technique [19] 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 [12] and structured extensions thereof [5, 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 [6, 14, 24]. Neither option takes advantage of the composite structure of the objective (1) or handles the case when the linear mapping B is nontrivial. 1

Contributions Our goal is to propose a new first-order optimization algorithm, called SemiProximal Mirror-Prox, designed to solve the difficult nonsmooth composite optimization problem (1), which does not require the exact computation of proximal operators. Instead, the SemiProximal Mirror-Prox relies upon i) Saddle point representability of f (a less restricted role than Fenchel-type representation); ii) Linear minimization oracle associated with k ? k in the domain X. While the saddle point representability of f allows to cure the non-smoothness of f , the linear minimization over the domain X allows to tackle the non-smooth regularization penalty k ? k. We establish the theoretical convergence rate of Semi-Proximal Mirror-Prox, which exhibits the optimal complexity bounds, i.e. O(1/?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 B ? 0: Semi-Proximal Mirror-Prox does not require assumptions on favorable geometry of dual domain Z or simplicity of ?(?) in (2). 2. Case B = I: Semi-Proximal Mirror-Prox is competitive with previously proposed approaches [15, 24] based on smoothing techniques. 3. Case of non-trivial B: Semi-Proximal Mirror-Prox is the first proximal-free or conditionalgradient-type optimization algorithm for (1). Related work The Semi-Proximal Mirror-Prox algorithm belongs to 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 [12, 1, 4]. Recently, in [6, 13], the authors consider constrained non-smooth optimization when the domain Z has a ?favorable geometry?, i.e. the domain is amenable to proximal setups (favorable geometry), and establish a complexity bound with O(1/?2) calls to the linear minimization oracle. Recently, in [15], a method called conditional gradient sliding is proposed to solve similar problems, using a smoothing technique, with a complexity bound in O(1/?2) for the calls to the linear minimization oracle (LMO) and additionally a O(1/?) bound for the linear operator evaluations. Actually, this O(1/?2) bound for the LMO complexity can be shown to be indeed optimal for conditional-gradient-type or LMObased algorithms, when solving general1 non-smooth convex problems [14]. 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 [7, 9, 26, 24, 16], but cannot be applied for our problem. In [9], f is smooth and B is identity matrix, whereas in [24], f is non-smooth and 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 [9] and the Composite Mirror-Prox [11], that enjoys the optimal complexity bound O(1/?2) on the total number of LMO calls, yet solves a broader class of convex problems than previously considered.

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 kP ? k, we define the dual norm as m Pn ksk? = maxkxk?1 hs, xi. For any x ? Rm?n , kxk2 = kxkF = (i=1 j=1 —xij —2)1/2 . Problem We consider the composite minimization problem Opt = min f (x) + kBxk x?X

(3)

where X is a closed convex set in the Euclidean space Ex ; x 7? Bx is a linear mapping from X to Y (? BX), where Y is a closed convex set in the Euclidean space Ey . We make two important assumptions on the function f and the norm k?k defining the regularization penalty, explained below. 1 Related research extended such approaches to stochastic or online settings [10, 8, 15]; such settings are beyond the scope of this work.

2

Saddle Point 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 f (x) is a non-smooth convex function given by f (x) = max ?(x, z)

(4)

z?Z

where ?(x, z) is a smooth convex-concave function and Z is a convex and compact set in the Euclidean space Ez . Such representation was introduced and developed in [6, 11, 13], for the purpose of non-smooth optimization. Saddle point representability can be interpreted as a general form of the smoothing-favorable structure of non-smooth functions used in the Nesterov smoothing technique [19]. Representations of this type are readily available for a wide family of ?well-structured? nonsmooth functions f (see Sec. 4 for examples), and actually for all empirical risk functions with convex loss in machine learning, up to our knowledge. Composite Linear Minimization Oracle Proximal-gradient-type algorithms require the compu tation of a proximal operator at each iteration, i.e. miny?Y 12 kyk22 + h?, yi + ?kyk . 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 be much cheaper. The linear minimization oracle (LMO) is a routine which, given an input ? ¿ 0 and ? ? Ey , returns a point LMO(?, ?) := argmin {h?, yi + ?kyk}

(5)

y?Y

In the case of nuclear-norm, the LMO only requires the computation of the leading pair of singular vectors, which is an order of magnitude faster 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 associated variational inequality, which provides the sufficient and necessary condition for an optimal solution to the saddle point problem [3, 4]. For any optimization problem with convex structure (including convex minimization, convex-concave saddle point problem, convex Nash equilibrium), the corresponding variational inequality is directly related to the accuracy certificate used to guarantee the accuracy of a solution to the optimization problem; see Sec. 2.1 in [11] and [18]. We shall present then an algorithm to solve the variational inequality established below, that exploits its particular structure. Assuming that f admits a saddle point representation (4), we write (3) in epigraph form Opt =

min

max {?(x, z) + ? : y = Bx} .

x?X,y?Y,? ?kyk z?Z

where Y (? BX) is a convex set. We can approximate Opt by d= Opt

min

max

x?X,y?Y,? ?kyk z?Z,kwk2 ?1

{?(x, z) + ? + ?hy ? Bx, wi} .

(6)

d = Opt (see details in [11]). By introducing the variables For properly selected ? ¿ 0, one has Opt u := [x, y; z, w] and v := ? , the variational inequality associated with the above saddle point problem is fully described by the domain X+

=

{x+ = [u; v] : x ? X, y ? Y, z ? Z, kwk2 ? 1, ? ? kyk}

and the monotone vector field

F (x+ = [u; v]) = [Fu (u); Fv] , where

? ?? ? ? x ?x ?(x, z) ? ?B T w ? y ?? ? ? ? ?w Fu ?u = ? ?? = ? ?, z ??z ?(x, z) w ?(Bx ? y) ?

Fv (v = ?) = 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

3

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

Semi-structured variational inequalities (Semi-VI) enjoy a particular mixed 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 = X1 ? X2 , such that X1 admits a proximal-mapping while X2 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 prox-mappings and can solve a broad class of variational inequalites. The original Mirror Prox algorithm was introduced in [17] and was extended to composite settings in [11] assuming exact computations of prox-mappings. Structured Variational Inequalities. We consider the variational inequality VI(X, F): Find x? ? X : hF (x), x ? x? i ? 0, ?x ? X

with domain X and operator F that satisfy the assumptions (A.1)?(A.4) below. (A.1) Set X ? Eu ? Ev is closed convex and its projection P X = {u : x = [u; v] ? X} ? U , where U is convex and closed, Eu , Ev are Euclidean spaces; (A.2) The function ?(?) : U ? R is continuously differentiable and also 1-strongly convex w.r.t. some norm2 k ? k. This defines the Bregman distance Vu (u?) = ?(u?) ? ?(u) ? h? ? (u), u? ? ui ? 12 ku? ? uk2 . (A.3) The operator F (x = [u, v]) : X ? Eu ? Ev is monotone and of form F (u, v) = [Fu (u); Fv] with Fv ? Ev being a constant and Fu (u) ? Eu satisfying the condition ?u, u? ? U : kFu (u) ? Fu (u?)k? ? Lku ? u? k + M

for some L ¡ ?, M ¡ ?; (A.4) The linear form hFv , vi of [u; v] ? Eu ? Ev is bounded from below on X and is coercive on X w.r.t. v: whenever [ut ; v t] ? X, t = 1, 2, ... is a sequence such that {ut }? t=1 is bounded and kv t k2 ? ? as t ? ?, we have hFv , v t i ? ?, t ? ?. The quality of an iterate, in the course of the algorithm, is measured through the so-called dual gap function

?VI (xX, F) = sup hF (y), x ? yi . y?X

We give in Appendix A a refresher on dual gap functions, for the reader?s convenience. We shall establish the complexity bounds in terms of 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. ?-Prox-mapping Inexact proximal mappings 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 ?-prox-mapping for ? ? 0. For ? = [?; ?] ? Eu ? Ev and x = [u; v] ? X, let us define the subset Px? (?) of X as Px? (?) = {b x = [b u; vb] ? X : h? + ? ? (b u) ? ? ? (u), u b ? si + h?, vb ? wi ? ? ?[s; w] ? X}.

When ? = 0, this reduces to the exact prox-mapping, in the usual setting, that is Px (?) = Argmin {h?, si + h?, wi + Vu (s)} . [s;w]?X

2

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

4

When ? ¿ 0, this yields our definition of an inexact prox-mapping, with inexactness parameter ?. Note that for any ? ? 0, the set Px? (? = [?; ?Fv]) is well defined whenever ? ¿ 0. The Composite Mirror Prox with inexact prox-mappings is outlined in Algorithm 1. Algorithm 1 Composite Mirror Prox Algorithm (CMP) for VI(X, F) Input: stepsizes ?t ¿ 0, inexactness ?t ? 0, t = 1, 2, . . . Initialize x1 = [u1 ; v 1] ? X for t = 1, 2, . . . , T do y t := [b ut ; vbt] ? Px?tt (?t F (xt)) = Px?tt (?t [Fu (ut); Fv]) t+1 t+1 t+1 x := [u ; v] ? Px?tt (?t F (y t)) = Px?tt (?t [Fu (b ut); Fv]) end for PT ?1 PT t Output: xT := [? uT ; v?T] = (t=1 ?t) t=1 ?t y

(7)

The proposed algorithm is a non-trivial extension of the Composite Mirror Prox with exact proxmappings, both from a theoretical and algorithmic point of views. We establish below the theoretical convergence rate; see Appendix B for the proof. Theorem 3.1. Assume that the sequence of step-sizes (?t) in the CMP algorithm satisfy ut) ? ?t2 M 2 , ?t := ?t hFu (b ut) ? Fu (ut), u bt ? ut+1 i ? Vubt (ut+1) ? Vut (b

t = 1, 2, . . . , T . (8)

Then, denoting ?[X] = sup[u;v]?X Vu1 (u), for a sequence of inexact prox-mappings with inexactness ?t ? 0, we have PT PT

?[X] + M 2 t=1 ?t2 + 2 t=1 ?t ?T ? xi ? ?VI (? xT X, F) := sup hF (x), x . (9) PT x?X t=1 ?t

Remarks. Note that the assumption on the sequence of step-sizes (?t) is clearly satisfied when ? ?t ? (2L)?1 . When M = 0 (which is essentially the case for the problem described in Section 2), it suffices as long as ?t ? L?1 . When (?t) is summable, we achieve the same O(1/T) convergence rate as when there is no error. If (?t) decays with a rate of O(1/t), then the overall convergence is only affected by a log(T) factor. Convergence results on the sequence of projections of (? xT) onto X1 when F stems from saddle point problem minx1 ?X1 supx2 ?X2 ?(x1 , x2) is established in Appendix B. The theoretical convergence rate established in Theorem 3.1 and Corollary B.1 generalizes the previous result established in Corollary 3.1 in [11] for CMP with exact prox-mappings. Indeed, when exact prox-mappings are used, we recover the result of [11]. When inexact prox-mappings are used, the errors due to the inexactness of the prox-mappings accumulate and is reflected in (9) and (37). 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 first introduced in [9] and also developed in [21]. We present an extension here which turns to be well-suited for sub-problems that will be solved in Sec. 3.3. Minimizing Smooth Semi-linear Functions. We consider the smooth semi-linear problem +

min ? (u, v) = ?(u) + h?, vi (10) x=[u;v]?X

+

represented by the pair (X; ?) such that the following assumptions are satisfied. We assume that i) X ? Eu ? Ev is closed convex and its projection P X on Eu belongs to U , where U is convex and compact; ii) ?(u) : U ? R is a convex continuously differentiable function, and there exist 1 ¡ ? ? 2 and L0 ¡ ? such that L0 ? ku ? uk? ?u, u? ? U ; (11) ?(u?) ? ?(u) + h??(u), u? ? ui + ? 5

iii) ? ? Ev is such that every linear function on Eu ? Ev of the form [u; v] 7? h?, ui + h?, vi

(12)

with ? ? Eu attains its minimum on X at some point x[?] = [u[?]; v[?]]; we have at our disposal a Composite Linear Minimization Oracle (LMO) which, given on input ? ? Eu , returns x[?]. Algorithm 2 Composite Conditional Gradient Algorithm CCG(X, ?(?), ?; ?) Input: accuracy ? ¿ 0 and ?t = 2/(t + 1), t = 1, 2, . . . Initialize x1 = [u1 ; v 1] ? X for t = 1, 2, . . . do Compute ?t = hgt , ut ? ut [gt]i + h?, v t ? v t [gt]i, where gt = ??(ut); if ?t ? ? then Return xt = [ut ; v t] else Find xt+1 = [ut+1 ; v t+1] ? X such that ?+ (xt+1) ? ?+ (xt + ?t (xt [gt] ? xt)) end if end for The algorithm is outlined in Algorithm 2. Note that CCG works essentially as if there were no vcomponent at all. The CCG algorithm enjoys a convergence rate in O(t?(??1)) in the evaluations of the function ?+ , and the accuracy certificates (?t) enjoy the same rate O(t?(??1)) as well. Proposition 3.1. Denote D the k?k-diameter of U . When solving problems of type (10), the sequence of iterates (xt) of CCG satisfies

??1 2 2L0 D? (13) ?+ (xt) ? min ?+ (x) ? , t?2 x?X ?(3 ? ?) t + 1

In addition, the accuracy certificates (?t) satisfy

min ?s ? O(1)L0 D?

1?s?t

3.3

2 t+1

??1

, t?2

(14)

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 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 Eu = Eu1 ? Eu2 , Ev = Ev1 ? Ev2 , and U ? U1 ? U2 , X = X1 ? X2 with Xi ? Eui ? Evi and Pi X = {ui : [ui ; vi] ? Xi } ? Ui for i = 1, 2, where U1 is convex and closed, U2 is convex and compact. We also assume that ?(u) = ?1 (u1) + ?2 (u2) and kuk = ku1 kEu1 + ku2 kEu2 , with ?2 (?) : U2 ? R continuously differentiable such that L0 ? ?2 (u?2) ? ?2 (u2) + h??2 (u2), u?2 ? u2 i + ku ? u2 k?Eu2 , ?u2 , u?2 ? U2 ; ? 2 for a particular 1 ¡ ? ? 2 and L0 ¡ ?. Furthermore, we assume that the k ? kEu2 -diameter of U2 is bounded by some D ¿ 0. (S.2) Partition of F : the operator F induced by the above partition of X1 and X2 can be written as F (x) = [Fu (u); Fv] with Fu (u) = [Fu1 (u1 , u2); Fu2 (u1 , u2)], Fv = [Fv1 ; Fv2]. 6

(S.3) Proximal mapping on X1 : we assume that for any ?1 ? Eu1 and ? ¿ 0, we have at our disposal easy-to-compute prox-mappings of the form, argmin Prox?1 (?1 , ?) := {?1 (u1) + h?1 , u1 i + ?hFv1 , v1 i} . x1 =[u1 ;v1]?X1

(S.4) Linear minimization oracle for X2 : we assume that we we have at our disposal Composite Linear Minimization Oracle (LMO), which given any input ?2 ? Eu2 and ? ¿ 0, returns an optimal solution to the minimization problem with linear form, that is, LMO(?2 , ?) := argmin {h?2 , u2 i + ?hFv2 , v2 i} . x2 =[u2 ;v2]?X2

Semi-proximal setup We denote such problems as Semi-VI(X, F). On the one hand, when U2 is a singleton, we get the full-proximal setup. On the other hand, when U1 is a singleton, we get the full linear-minimization-oracle setup (full LMO setup). The semi-proximal setup allows to cover both setups and all the ones in between as well. 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.4). The Semi-Proximal Mirror-Prox algorithm blends both CMP and CCG. Basically, for sub-domain X2 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 X1 , we compute the prox-mapping as it is. Algorithm 3 Semi-Proximal Mirror-Prox Algorithm for Semi-VI(X, F) Input: stepsizes ?t ¿ 0, accuracies ?t ? 0, t = 1, 2, . . . [1] Initialize x1 = [x11 ; x12] ? X, where x11 = [u11 ; v11]; x12 = [u12 , ; v21]. for t = 1, 2, . . . , T do [2] Compute y t = [y1t ; y2t] that y1t := [b ut1 ; vb1t] t ut2 ; vb2t] y2 := [b

= =

Prox?1 (?t Fu1 (ut1 , ut2) ? ?1? (ut1), ?t) CCG(X2 , ?2 (?) + h?t Fu2 (ut1 , ut2) ? ?2? (ut2), ?i, ?t Fv2 ; ?t)

t+1 [3] Compute xt+1 = [xt+1 1 ; x2] that t+1 xt+1 := [ut+1 1 1 ; v1] t+1 xt+1 := [ut+1 2 2 ; v2]

= =

ut1 , u bt2) ? ?1? (ut1), ?t) Prox?1 (?t Fu1 (b ut1 , u bt2) ? ?2? (ut2), ?i, ?t Fv2 ; ?t) CCG(X2 , ?2 (?) + h?t Fu2 (b

end for PT ?1 PT t uT ; v?T] = (t=1 ?t) Output: xT := [? t=1 ?t y

At step t, we first update y1t = [b ut1 ; vb1t] by computing the exact prox-mapping and build y2t = [b ut2 ; vb2t] by running the composite conditional gradient algorithm to problem (10) specifically with X = X2 , ?(?) = ?2 (?) + h?t Fu2 (ut1 , ut2) ? ?2? (ut2), ?i, and ? = ?t Fv2 ,

t+1 t+1 = [ut+1 = until ?(y2t) = maxy2 ?X2 h??+ (y2t), y2t ? y2 i ? ?t . We then build xt+1 1 1 ; v1] and x2 t+1 t+1 t [u2 ; v2] similarly except this time taking the value of the operator at point y . Combining the results in Theorem 3.1 and Proposition 3.1, we arrive at the following complexity bound. Proposition 3.2. Under the assumption (A.1) ? (A.4) and (S.1) ? (S.4) with M = 0, and choice of stepsize ?t = L?1 , t = 1, . . . , T , for the outlined algorithm to return an ?-solution to the variational inequality V I(X, F), the total number of Mirror Prox steps required does not exceed L?[X] Total number of steps = O(1) ? and the total number of calls to the Linear Minimization Oracle does not exceed 1

L0 L? D? ??1 ?[X]. N = O(1) ??

In particular, if we use Euclidean proximal setup on U2 with ?2 (?) = 12 kx2 k2 , which leads to ? =2 and L0 = 1, then the number of LMO calls does not exceed N = O(1) L2 D2 (?[X1] + D2) /?2 . 7

0

0

10

Semi?MP(eps=1e2/t) Semi?MP(eps=1e1/t) Semi?MP(eps=1e0/t) Semi?LPADMM(eps=1e?3/t) Semi?LPADMM(eps=1e?4/t) Semi?LPADMM(eps=1e?5/t)

0.9

?2

0.8 0.7 0.6 0.5

?1

Semi?MP(eps=1e1/t) Semi?LPADMM(eps=1e?5/t)

0.8 0.7 0.6 0.5 0.4

0.4

10

0.9

objective value

?1

10

Semi?MP(eps=5/t) Semi?MP(fixed=24) Smooth?CG(?=1) Semi?SPG(eps=10/t) Semi?SPG(fixed=24)

objective value

Semi?MP(eps=10/t) Semi?MP(fixed=96) Smooth?CG(?=0.01) Semi?SPG(eps=5/t) Semi?SPG(fixed=96)

Objective valule

Objective valule

10

10 0

1000

2000

3000

Elapsed time (sec)

4000

0

500

1000

1500

2000

2500

0.3 0

3000

1000

2000

3000

4000

number of LMO calls

Elapsed time (sec)

5000

0.3 0

200

400

600

800

1000

number of LMO calls

Figure 1: Robust collaborative filtering and link prediction: objective function vs elapsed time. From left to right: (a) MovieLens100K; (b) MovieLens1M; (c) Wikivote (1024); (d) Wikivote (full) Discussion The proposed Semi-Proximal Mirror-Prox algorithm enjoys the optimal complexity bounds, i.e. O(1/?2), in the number of calls to LMO; see [14] for the optimal complexity bounds for general non-smooth optimization with LMO. Consequently, when applying the algorithm to the variational reformulation of the problem of interest (3), we are able to get an ?-optimal solution within at most O(1/?2) LMO calls. Thus, Semi-Proximal Mirror-Prox generalizes previously proposed approaches and improves upon them in special cases of problem (3); see Appendix D.2.

4

Experiments

We report the experimental results obtained with the proposed Semi-Proximal Mirror-Prox, denoted Semi-MP here, and competing algorithms. We consider two different applications: i) robust collaborative filtering for movie recommendation; ii) link prediction for social network analysis. For i), we compare to two competing approaches: a) smoothing conditional gradient proposed in [24] (denoted Smooth-CG); b) smoothing proximal gradient [20, 5] equipped with semi-proximal setup (Semi-SPG). For ii), we compare to Semi-LPADMM, using [22] equipped with semi-proximal setup. Additional experiments and implementation details are given in Appendix E. Robust collaborative filtering We consider the collaborative filtering problem, with a nuclearnorm regularization penalty and an ?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 regularization parameters. In Fig. 1, 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 ?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 ?1 norm term. Therefore, we consider an 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 analyzed in literature. However, intuitively, as long as the 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 103747 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. 1, 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-mappings 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. Acknowledgments The authors would like to thank A. Juditsky and A. Nemirovski for fruitful discussions. This work was supported by NSF Grant CMMI-1232623, LabEx Persyval-Lab (ANR-11-LABX-0025), project ?Titan? (CNRS-Mastodons), project ?Macaron? (ANR-14-CE23-0003-01), the MSR-Inria joint centre, and the Moore-Sloan Data Science Environment at NYU. 8

2 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 sparsityinducing penalties. Found. Trends Mach. Learn., 4(1):1?106, 2012. [3] Heinz H. Bauschke and Patrick L. Combettes. Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, 2011. [4] D. P. Bertsekas. Convex Optimization Algorithms. Athena Scientific, 2015. [5] 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. [6] Bruce Cox, Anatoli Juditsky, and Arkadi Nemirovski. Dual subgradient algorithms for large-scale nonsmooth learning

problems. Mathematical Programming, pages 1?38, 2013. [7] 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. [8] Dan Garber and Elad Hazan. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. arXiv preprint arXiv:1301.4666, 2013. [9] Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for normregularized smooth convex optimization. Mathematical Programming, pages 1?38, 2013. [10] E. Hazan and S. Kale. Projection-free online learning. In ICML, 2012. [11] 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. [12] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In ICML, pages 427? 435, 2013. [13] Anatoli Juditsky and Arkadi Nemirovski. Solving variational inequalities with monotone operators on domains given by linear minimization oracles. arXiv preprint arXiv:1312.107, 2013. [14] Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle. arXiv, 2013. [15] Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. arXiv, 2014. [16] Cun Mu, Yuqian Zhang, John Wright, and Donald Goldfarb. Scalable robust matrix recovery: Frankwolfe meets proximal methods. arXiv preprint arXiv:1403.7588, 2014. [17] 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. [18] 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. [19] Yurii Nesterov. Smooth minimization of non-smooth functions. Mathematical programming, 103(1):127? 152, 2005. [20] Yurii Nesterov. Smoothing technique and its applications in semidefinite optimization. Math. Program., 110(2):245?259, 2007. [21] Yurii Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. Technical report, Universit?e catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2015. [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 J?er?ome Malick. A smoothing approach for composite conditional gradient with nonsmooth loss. In Conf?erence 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.

9