FORWARD-BACKWARD TRUNCATED NEWTON METHODS FOR CONVEX COMPOSITE OPTIMIZATION[1]

PANAGIOTIS PATRINOS, LORENZO STELLA, AND ALBERTO BEMPORAD

Abstract. This paper proposes two proximal Newton-CG methods for convex nonsmooth optimization problems in composite form. The algorithms are based on a a reformulation of the original nonsmooth problem as the unconstrained minimization of a continuously differentiable function, namely the *forward-backward envelope (FBE)*. The first algorithm is based on a standard line search strategy, whereas the second one combines the global efficiency estimates of the corresponding first-order methods, while achieving fast asymptotic convergence rates. Furthermore, they are computationally attractive since each Newton iteration requires the approximate solution of a linear system of usually small dimension.

1. Introduction

The focus of this work is on efficient Newton-like algorithms for convex optimization problems in composite form, *i.e.*,

$$\text{minimize } F(x) = f(x) + g(x), \tag{1.1}$$

where $f \in \mathcal{S}^{2,1}_{\mu_f,L_f}(\mathbb{R}^n)$[2] and $g \in \mathcal{S}^0(\mathbb{R}^n)$[3] has a cheaply computable proximal mapping [2]. Problems of the form (1.1) are abundant in many scientific areas such as control, signal processing, system identification, machine learning and image analysis, to name a few. For example, when g is the indicator of a convex set then (1.1) becomes a constrained optimization problem, while for $f(x) = \|Ax - b\|_2^2$ and $g(x) = \lambda\|x\|_1$ it becomes the ℓ_1-regularized least-squares problem which is the main building block of compressed sensing. When g is equal to the nuclear norm, then problem (1.1) can model low-rank matrix recovery problems. Finally, conic optimization problems such as LPs, SOCPs and SPDs can be brought into the form of (1.1), see [3].

Perhaps the most well known algorithm for problems in the form (1.1) is the forward-backward splitting (FBS) or proximal gradient method [4, 5], a generalization of the classical gradient and gradient projection methods to problems involving a nonsmooth term. Accelerated versions of FBS, based on the work of Nesterov [6–8], have also gained popularity. Although these algorithms share favorable global convergence rate estimates of order $O(\epsilon^{-1})$ or $O(\epsilon^{-1/2})$ (where ϵ is the solution accuracy), they are first-order methods and therefore usually effective at computing solutions of low or medium accuracy only. An evident remedy is to

[1]A preliminary version of this paper [1] was presented at the 52nd IEEE Conference on Decision and Control, Florence, Italy, December 11, 2013.

[2]$\mathcal{S}^{2,1}_{\mu,L}(\mathbb{R}^n)$: class of twice continuously differentiable, strongly convex functions with modulus of strong convexity $\mu \geq 0$, whose gradient is Lipschitz continuous with constant $L \geq 0$.

[3]$\mathcal{S}^0(\mathbb{R}^n)$: class of proper, lower semicontinuous, convex functions from $\mathbb{R}^n$ to $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$.

include second-order information by replacing the Euclidean norm in the proximal mapping with the Q-norm, where Q is the Hessian of f at x or some approximation of it, mimicking Newton or quasi-Newton methods for unconstrained problems. This route is followed in the recent work of [9, 10]. However, a severe limitation of the approach is that, unless Q has a special structure, the linearized subproblem is very hard to solve. For example, if F models a QP, the corresponding subproblem is as hard as the original problem.

In this paper we follow a different approach by defining a function, which we call *forward-backward envelope (FBE)*, that has favorable properties and can serve as a real-valued, smooth, exact penalty function for the original problem. Our approach combines and extends ideas stemming from the literature on merit functions for *variational inequalities* (VIs) and *complementarity problems* (CPs), specifically the reformulation of a VI as a constrained continuously differentiable optimization problem via the regularized gap function [11] and as an unconstrained continuously differentiable optimization problem via the D-gap function [12] (see [13, Ch. 10] for a survey and [14], [15] for applications to constrained optimization and model predictive control of dynamical systems).

Next, we show that one can design Newton-like methods to minimize the FBE by using tools from nonsmooth analysis. Unlike the approaches of [9, 10], where the corresponding subproblems are expensive to solve, our algorithms require only the solution of a usually small linear system to compute the Newton direction. However, this work focuses on devising algorithms that have good complexity guarantees provided by a global (non-asymptotic) convergence rate while achieving Q-superlinear or Q-quadratic[4] asymptotic convergence rates in the nondegenerate cases. We show that one can achieve this goal by interleaving Newton-like iterations on the FBE and FBS iterations. This is possible by relating directions of descent for the considered penalty function with those for the original nonsmooth function.

The main contributions of the paper can be summarized as follows. We show how Problem (1.1) can be reformulated as the unconstrained minimization of a real-valued, continuously differentiable function, the FBE, providing a framework that allows to extend classical algorithms for smooth unconstrained optimization to nonsmooth or constrained problems in composite form (1.1). Moreover, based on this framework, we devise efficient proximal Newton algorithms with Q-superlinear or Q-quadratic asymptotic convergence rate to solve (1.1), with global complexity bounds. The conjugate gradient (CG) method is employed to compute efficiently an approximate Newton direction at every iteration. Therefore our algorithms are able to handle large-scale problems since they require only the calculation of matrix-vector products and there is no need to form explicitly the generalized Hessian matrix.

The outline of the paper is as follows. In Section 2 we introduce the FBE, a continuously differentiable penalty function for (1.1), and discuss some of its properties. In Section 3 we discuss the generalized differentiability properties of the gradient of the FBE and introduce a linear Newton approximation (LNA) for it, which plays a role similar to that of the Hessian in the classical Newton method. Section 4

[4] A sequence $\{x^k\}_{k\in\mathbb{N}}$ converges to $x_\star$ with Q-superlinear rate if $\frac{\|x^{k+1}-x_\star\|}{\|x_k-x_\star\|} \to 0$. It converges to $x_\star$ with Q-quadratic rate if there exists a $\bar{k} > 0$ such that $\frac{\|x^{k+1}-x_\star\|}{\|x_k-x_\star\|^2} \leq M$, for some $M > 0$ and all $k \geq \bar{k}$.

is the core of the paper, presenting two algorithms for solving Problem (1.1) and discussing their local and global convergence properties. In Section 5 we consider some examples of g and discuss the generalized Jacobian of their proximal operator, on which the LNA is based. Finally, in Section 6, we consider some practical problems and show how the proposed methods perform in solving them.

2. Forward-backward envelope

In the following we indicate by $X_\star$ and $F_\star$, respectively, the set of solutions of problem (1.1) and its optimal objective value. Forward-backward splitting for solving (1.1) relies on computing, at every iteration, the following update

$$x^{k+1} = \mathrm{prox}_{\gamma g}(x^k - \gamma \nabla f(x^k)), \tag{2.1}$$

where the *proximal mapping* [2] of g is defined by

$$\mathrm{prox}_{\gamma g}(x) \triangleq \operatorname*{argmin}_u \left\{ g(u) + \tfrac{1}{2\gamma}\|u - x\|^2 \right\}. \tag{2.2}$$

The value function of the optimization problem (2.2) defining the proximal mapping is called the *Moreau envelope* and is denoted by g^γ, *i.e.*,

$$g^\gamma(x) \triangleq \inf_u \left\{ g(u) + \tfrac{1}{2\gamma}\|u - x\|^2 \right\}. \tag{2.3}$$

Properties of the Moreau envelope and the proximal mapping are well documented in the literature [5,16–18]. For example, the proximal mapping is single-valued, continuous and nonexpansive (Lipschitz continuous with Lipschitz 1) and the envelope function g^γ is convex, continuously differentiable, with γ^{-1}-Lipschitz continuous gradient

$$\nabla g^\gamma(x) = \gamma^{-1}(x - \mathrm{prox}_{\gamma g}(x)). \tag{2.4}$$

We will next proceed to the reformulation of (1.1) as the minimization of an unconstrained continuously differentiable function. It is well known [16] that an optimality condition for (1.1) is

$$x = \mathrm{prox}_{\gamma g}(x - \gamma \nabla f(x)). \tag{2.5}$$

Since $f \in \mathcal{S}^{2,1}_{\mu_f, L_f}(\mathrm{I\!R}^n)$, we have that $\|\nabla^2 f(x)\| \leq L_f$ [19, Lem. 1.2.2], therefore $I - \gamma \nabla^2 f(x)$ is symmetric and positive definite whenever $\gamma \in (0, 1/L_f)$. Premultiplying both sides of (2.5) by $\gamma^{-1}(I - \gamma\nabla^2 f(x))$, $\gamma \in (0, 1/L_f)$, one obtains the equivalent condition

$$\gamma^{-1}(I - \gamma \nabla^2 f(x))(x - \mathrm{prox}_{\gamma g}(x - \gamma \nabla f(x))) = 0. \tag{2.6}$$

The left-hand side of equation (2.6) is the gradient of the function that we call *forward-backward envelope*, indicated by F_γ. Using (2.4) to integrate (2.6), one obtains the following definition.

Definition 2.1. *Let $F(x) = f(x) + g(x)$, where $f \in \mathcal{S}^{2,1}_{\mu_f, L_f}(\mathrm{I\!R}^n)$, $g \in \mathcal{S}^0(\mathrm{I\!R}^n)$. The forward-backward envelope of F is given by*

$$F_\gamma(x) \triangleq f(x) - \tfrac{\gamma}{2}||\nabla f(x)||_2^2 + g^\gamma(x - \gamma \nabla f(x)). \tag{2.7}$$

Alternatively, one can express F_γ as the value function of the minimization problem that yields forward-backward splitting. In fact

$$F_\gamma(x) = \min_{u \in \mathrm{I\!R}^n} \left\{ f(x) + \nabla f(x)'(u - x) + g(u) + \tfrac{1}{2\gamma}\|u - x\|^2 \right\} \tag{2.8a}$$

$$= f(x) + g(P_\gamma(x)) - \gamma \nabla f(x)' G_\gamma(x) + \tfrac{\gamma}{2}\|G_\gamma(x)\|^2, \tag{2.8b}$$

where

$$P_\gamma(x) \triangleq \mathrm{prox}_{\gamma g}(x - \gamma\nabla f(x)),$$
$$G_\gamma(x) \triangleq \gamma^{-1}(x - P_\gamma(x)).$$

One distinctive feature of F_γ is the fact that it is real-valued despite the fact that F can be extended-real-valued. In addition, F_γ enjoys favorable properties, summarized in the next theorem.

Theorem 2.2. *The following properties of F_γ hold:*

(i) *F_γ is continuously differentiable with*

$$\nabla F_\gamma(x) = \left(I - \gamma\nabla^2 f(x)\right) G_\gamma(x). \tag{2.9}$$

If $\gamma \in (0, 1/L_f)$ then the set of stationary points of F_γ equals $X_\star$.

(ii) *For any $x \in \mathbb{R}^n$, $\gamma > 0$*

$$F_\gamma(x) \le F(x) - \tfrac{\gamma}{2}\|G_\gamma(x)\|^2. \tag{2.10}$$

(iii) *For any $x \in \mathbb{R}^n$, $\gamma > 0$*

$$F(P_\gamma(x)) \le F_\gamma(x) - \tfrac{\gamma}{2}\left(1 - \gamma L_f\right)\|G_\gamma(x)\|^2. \tag{2.11}$$

In particular, if $\gamma \in (0, 1/L_f]$ then

$$F(P_\gamma(x)) \le F_\gamma(x). \tag{2.12}$$

(iv) *If $\gamma \in (0, 1/L_f)$ then $X_\star = \operatorname{argmin} F_\gamma$.*

Proof. Part (i) has already been proven. Regarding (ii), from the optimality condition for the problem defining the proximal mapping we have

$$G_\gamma(x) - \nabla f(x) \in \partial g(P_\gamma(x)),$$

i.e., $G_\gamma(x) - \nabla f(x)$ is a subgradient of g at $P_\gamma(x)$. From the subgradient inequality

$$\begin{aligned} g(x) &\ge g(P_\gamma(x)) + (G_\gamma(x) - \nabla f(x))'(x - P_\gamma(x)) \\ &= g(P_\gamma(x)) - \gamma\nabla f(x)' G_\gamma(x) + \gamma\|G_\gamma(x)\|^2 \end{aligned}$$

Adding $f(x)$ to both sides proves the claim. For part (iii), we have

$$\begin{aligned} F_\gamma(x) &= f(x) + \nabla f(x)'(P_\gamma(x) - x) + g(P_\gamma(x)) + \tfrac{\gamma}{2}\|G_\gamma(x)\|^2 \\ &\ge f(P_\gamma(x)) + g(P_\gamma(x)) - \tfrac{L_f}{2}\|P_\gamma(x) - x\|^2 + \tfrac{\gamma}{2}\|G_\gamma(x)\|^2. \end{aligned}$$

where the inequality follows by Lipschitz continuity of ∇f and the descent lemma, see e.g. [20, Prop. A.24]. For part (iv), putting $x_\star \in X_\star$ in (2.10) and (2.11) and using $x_\star = P_\gamma(x_\star)$ we obtain $F(x_\star) = F_\gamma(x_\star)$. Now, for any $x \in \mathbb{R}^n$ we have $F_\gamma(x_\star) = F(x_\star) \le F(P_\gamma(x)) \le F_\gamma(x)$, where the first inequality follows by optimality of $x_\star$ for F, while the second inequality follows by (2.11). This shows that every $x_\star \in X_\star$ is also a (global) minimizer of F_γ. The proof finishes by recalling that the set of minimizers of F_γ are a subset of the set of its stationary points, which by (i) is equal to $X_\star$. □

Parts (i) and (iv) of Theorem (2.2) show that if $\gamma \in (0, 1/L_f)$, the nonsmooth problem (1.1) is completely equivalent to the unconstrained minimization of the continuously differentiable function F_γ, in the sense that the sets of minimizers and optimal values are equal. In other words we have

$$\operatorname{argmin} F = \operatorname{argmin} F_\gamma, \qquad \inf F = \inf F_\gamma.$$

Part (ii) shows that an ϵ-optimal solution x of F is automatically ϵ-optimal for F_γ, while part (iii) implies that from an ϵ-optimal for F_γ we can directly obtain an ϵ-optimal solution for F if γ is chosen sufficiently small, *i.e.*,

$$\begin{aligned} F(x) - F_\star \leq \epsilon &\implies F_\gamma(x) - F_\star \leq \epsilon, \\ F_\gamma(x) - F_\star \leq \epsilon &\implies F(P_\gamma(x)) - F_\star \leq \epsilon. \end{aligned}$$

Notice that part (iv) of Theorem 2.2 states that if $\gamma \in (0, 1/L_f)$, then not only do the stationary points of F_γ agree with $X_\star$ (cf. Theorem 2.2(i)), but also that its set of minimizers agrees with $X_\star$, *i.e.*, although F_γ may not be convex, the set of stationary points turns out to be equal to the set of its minimizers. However, in the particular but important case where f is convex quadratic, the FBE is convex with Lipschitz continuous gradient, as the following theorem shows.

Theorem 2.3. *If $f(x) = \frac{1}{2}x'Qx + q'x$ and $\gamma \in (0, 1/L_f)$, then $F_\gamma \in \mathcal{S}^{1,1}_{\mu_{F_\gamma}, L_{F_\gamma}}(\mathbb{R}^n)$, where*

$$L_{F_\gamma} = 2(1 - \gamma\mu_f)/\gamma, \tag{2.13a}$$

$$\mu_{F_\gamma} = \min\{(1 - \gamma\mu_f)\mu_f, (1 - \gamma L_f)L_f\} \tag{2.13b}$$

and $\mu_f = \lambda_{\min}(Q) \geq 0$, $L_f = \lambda_{\max}(Q)$.

Proof. Let

$$\begin{aligned} \psi_1(x) &\triangleq f(x) - (\gamma/2)\|\nabla f(x)\|^2 = (1/2)x'Q(I - \gamma Q)x - \gamma q'Qx - \gamma q'q, \\ \psi_2(x) &\triangleq g^\gamma(x - \gamma\nabla f(x)). \end{aligned}$$

Due to Lemma A.1 (in the Appendix), ψ_1 is strongly convex with modulus μ_{F_γ}. Function $\psi_2(x)$ is convex, as the composition of the convex function g^γ with the linear mapping $x - \gamma\nabla f(x)$. Therefore, $F_\gamma(x) = \psi_1(x) + \psi_2(x)$ is strongly convex with convexity parameter μ_{F_γ}. On the other hand, for every $x_1, x_2 \in \mathbb{R}^n$

$$\begin{aligned} \|\nabla F_\gamma(x_1) - \nabla F_\gamma(x_2)\| &\leq \|I - \gamma Q\|\|G_\gamma(x_1) - G_\gamma(x_2)\| \\ &\leq 2(1 - \gamma\mu_f)/\gamma\|x_1 - x_2\| \end{aligned}$$

where the second inequality is due to Lemma A.2 in the Appendix. □

Notice that if $\mu_f > 0$ and we choose $\gamma = 1/(L_f + \mu_f)$, then $L_{F_\gamma} = 2L_f$ and $\mu_{F_\gamma} = L_f\mu_f/(L_f + \mu_f)$, so $L_{F_\gamma}/\mu_{F_\gamma} = 2(L_f/\mu_f + 1)$. In other words the condition number of F_γ is roughly double compared to that of f.

2.1. **Interpretations.** It is apparent from (2.1) and (2.5) that FBS is a Picard iteration for computing a fixed point of the nonexpansive mapping P_γ. It is well known that fixed-point iterations may exhibit slow asymptotic convergence. On the other hand, Newton methods achieve much faster asymptotic convergence rates. However, in order to devise globally convergent Newton-like methods one needs a merit function on which to perform a line search, in order to determine a step size that guarantees sufficient decrease and damps the Newton steps when far from the solution. This is exactly the role that FBE plays in this paper.

Another interesting observation is that the FBE provides a link between gradient methods and FBS, just like the Moreau envelope (2.3) does for the proximal point algorithm [21]. To see this, consider the problem

$$\text{minimize } g(x) \tag{2.14}$$

where $g \in \mathcal{S}^0(\mathbb{R}^n)$. The proximal point algorithm for solving (2.14) is

$$x^{k+1} = \operatorname{prox}_{\gamma g}(x^k). \tag{2.15}$$

It is well known that the proximal point algorithm can be interpreted as a gradient method for minimizing the Moreau envelope of g, cf. (2.3). Indeed, due to (2.4), iteration (2.15) can be expressed as

$$x^{k+1} = x^k - \gamma \nabla g^\gamma(x^k).$$

This simple idea provides a link between nonsmooth and smooth optimization and has led to the discovery of a variety of algorithms for problem (2.14), such as semismooth Newton methods [22], variable-metric [23] and quasi-Newton methods [24], and trust-region methods [25], to name a few. However, when dealing with composite problems, even if $\operatorname{prox}_{\gamma g}$ and g^γ are cheaply computable, computing proximal mapping and Moreau envelope of $(f+g)$ is usually as hard as solving (1.1) itself. On the other hand, forward-backward splitting takes advantage of the structure of the problem by operating separately on the two summands, cf. (2.1). The question that naturally arises is the following:

> *Is there a continuously differentiable function that provides an interpretation of FBS as a gradient method, just like the Moreau envelope does for the proximal point algorithm and problem (2.14)?*

The forward-backward envelope provides an affirmative answer. Specifically, FBS can be interpreted as the following (variable metric) gradient method on the FBE:

$$x^{k+1} = x^k - \gamma(I - \gamma\nabla^2 f(x^k))^{-1}\nabla F_\gamma(x^k).$$

Furthermore, the following properties holding for g^γ

$$g^\gamma \le g, \quad \inf g^\gamma = \inf g, \quad \operatorname{argmin} g^\gamma = \operatorname{argmin} g.$$

correspond to Theorem 2.2(iii) and Theorem 2.2(iv) for the FBE. The relationship between Moreau envelope and forward-backward envelope is then apparent. This opens the possibility of extending FBS and devising new algorithms for problem (1.1) by simply reconsidering and appropriately adjusting methods for unconstrained minimization of continuously differentiable functions, the most well studied problem in optimization. In this work we exploit one of the numerous alternatives, by devising Newton-like algorithms that are able to achieve fast asymptotic convergence rates. The next section deals with the other obstacle that needs to be overcome, *i.e.*, constructing a second-order expansion for the $\mathcal{C}^1$ (but not $\mathcal{C}^2$) function F_γ around any optimal solution, that behaves similarly to the Hessian for $\mathcal{C}^2$ functions and allows us to devise algorithms with fast local convergence.

3. Second-order Analysis of F_γ

As it was shown in Section 2, F_γ is continuously differentiable over $\mathbb{R}^n$. However F_γ fails to be $\mathcal{C}^2$ in most cases: since g is nonsmooth, its Moreau envelope g^γ is hardly ever $\mathcal{C}^2$. For example, if g is real-valued then g^γ is $\mathcal{C}^2$ and $\operatorname{prox}_{\gamma g}$ is $\mathcal{C}^1$ if and only if g is $\mathcal{C}^2$ [26]. Therefore, we hardly ever have the luxury of assuming continuous differentiability of ∇F_γ and we must resort into generalized notions of differentiability stemming from nonsmooth analysis. Specifically, our analysis is largely based upon generalized differentiability properties of $\operatorname{prox}_{\gamma g}$ which we study next.

3.1. Generalized Jacobians of proximal mappings.

Since $\operatorname{prox}_{\gamma g}$ is globally Lipschitz continuous, by Rademacher's theorem [17, Th. 9.60] it is almost everywhere differentiable. Recall that Rademacher's theorem asserts that if a mapping $G : \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz continuous on $\mathbb{R}^n$, then it is almost everywhere differentiable, *i.e.*, the set $\mathbb{R}^n \setminus C_G$ has measure zero, where C_G is the subset of points in $\mathbb{R}^n$ for which G is differentiable. Hence, although the Jacobian of $\operatorname{prox}_{\gamma g}$ in the classical sense might not exist everywhere, generalized differentiability notions, such as the B-subdifferential and the generalized Jacobian of Clarke, can be employed to provide a local first-order approximation of $\operatorname{prox}_{\gamma g}$.

Definition 3.1. *Let $G : \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz continuous at $x \in \mathbb{R}^n$. The B-subdifferential (or limiting Jacobian) of G at x is*

$$\partial_B G(x) \triangleq \left\{ H \in \mathbb{R}^{m \times n} \mid \exists \, \{x^k\} \subset C_G \text{ with } x^k \to x, \nabla G(x^k) \to H \right\},$$

whereas the (Clarke) generalized Jacobian of G at x is

$$\partial_C G(x) \triangleq \operatorname{conv}(\partial_B G(x)).$$

If $G : \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz on $\mathbb{R}^n$ then $\partial_C G(x)$ is a nonempty, convex and compact subset of m by n matrices, and as a set-valued mapping it is outer-semicontinuous at every $x \in \mathbb{R}^n$. The next theorem shows that the elements of the generalized Jacobian of the proximal mapping are symmetric and positive semidefinite. Furthermore, it provides a bound on the magnitude of their eigenvalues.

Theorem 3.2. *Suppose that $g \in \mathcal{S}^0(\mathbb{R}^n)$ and $x \in \mathbb{R}^n$. Every $P \in \partial_C(\operatorname{prox}_{\gamma g})(x)$ is a symmetric positive semidefinite matrix that satisfies $\|P\| \leq 1$.*

Proof. Since g is convex, its Moreau envelope is a convex function as well, therefore every element of $\partial_C(\nabla g^\gamma)(x)$ is a symmetric positive semidefinite matrix (see e.g. [13, Sec. 8.3.3]). Due to (2.4), we have that $\operatorname{prox}_{\gamma g}(x) = x - \gamma \nabla g^\gamma(x)$, therefore

$$\partial_C(\operatorname{prox}_{\gamma g})(x) = I - \gamma \partial_C(\nabla g^\gamma)(x). \tag{3.1}$$

The last relation holds with equality (as opposed to inclusion in the general case) due to the fact that one of the summands is continuously differentiable. Now from (3.1) we easily infer that every element of $\partial_C(\operatorname{prox}_{\gamma g})(x)$ is a symmetric matrix. Since $\nabla g^\gamma(x)$ is Lipschitz continuous with Lipschitz constant γ^{-1}, using [27, Prop. 2.6.2(d)], we infer that every $H \in \partial_C(\nabla g^\gamma)(x)$ satisfies $\|H\| \leq \gamma^{-1}$. Now, according to (3.1), it holds

$$P \in \partial_C(\operatorname{prox}_{\gamma g})(x) \iff P = I - \gamma H, \quad H \in \partial_C(\nabla g^\gamma)(x).$$

Therefore,

$$d' P d = \|d\|^2 - \gamma d' H d \geq \|d\|^2 - \gamma \gamma^{-1} \|d\|^2 = 0, \quad \forall P \in \partial_C(\operatorname{prox}_{\gamma g})(x).$$

On the other hand, since $\operatorname{prox}_{\gamma g}$ is Lipschitz continuous with Lipschitz constant 1, using [27, Prop. 2.6.2(d)] we obtain that $\|P\| \leq 1$, for all $P \in \partial_C(\operatorname{prox}_{\gamma g})(x)$. □

An interesting property of $\partial_C \operatorname{prox}_{\gamma g}$, documented in the following proposition, is useful whenever g is (block) separable, *i.e.*, $g(x) = \sum_{i=1}^N g_i(x_i)$, $x_i \in \mathbb{R}^{n_i}$, $\sum_{i=1}^N n_i = n$. In such cases every $P \in \partial_C(\operatorname{prox}_{\gamma g})(x)$ is a (block) diagonal matrix. This has favorable computational implications especially for large-scale problems. For example, if g is the ℓ_1 norm or the indicator function of a box, then the elements of $\partial_C \operatorname{prox}_{\gamma g}(x)$ (or $\partial_B \operatorname{prox}_{\gamma g}(x)$) are diagonal matrices with diagonal elements in $[0, 1]$ (or in $\{0, 1\}$).

Proposition 3.3 (separability)**.** *If $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is (block) separable then every element of $\partial_B(\mathrm{prox}_{\gamma g})(x)$ and $\partial_C(\mathrm{prox}_{\gamma g})(x)$ is (block) diagonal.*

Proof. Since g is block separable, its proximal mapping has the form

$$\mathrm{prox}_{\gamma g}(x) = (\mathrm{prox}_{\gamma g_1}(x_1), \ldots, \mathrm{prox}_{\gamma g_N}(x_N)).$$

The result follows directly by Definition 3.1. □

The following proposition provides a connection between the generalized Jacobian of the proximal mapping for a convex function and that of its conjugate, stemming from the celebrated Moreau's decomposition [16, Th. 14.3].

Proposition 3.4 (Moreau's decomposition)**.** *Suppose that $g \in \mathcal{S}^0(\mathbb{R}^n)$. Then*

$$\partial_B(\mathrm{prox}_{\gamma g^\star})(x) = \{P = I - Q \mid Q \in \partial_B(\mathrm{prox}_{g/\gamma})(x/\gamma)\},$$
$$\partial_C(\mathrm{prox}_{\gamma g^\star})(x) = \{P = I - Q \mid Q \in \partial_C(\mathrm{prox}_{g/\gamma})(x/\gamma)\}.$$

Proof. Using Moreau's decomposition we have

$$\mathrm{prox}_{\gamma g^\star}(x) = x - \gamma\, \mathrm{prox}_{g/\gamma}(x/\gamma).$$

The first result follows directly by Definition 3.1, since $\mathrm{prox}_{\gamma g^\star}$ is expressed as the difference of two functions, one of which is continuously differentiable. The second result follows from the fact that, with a little abuse of notation,

$$\mathrm{conv}\{I - Q \mid Q \in \partial_B(\mathrm{prox}_{g/\gamma})(x/\gamma)\} = I - \mathrm{conv}(\partial_B(\mathrm{prox}_{g/\gamma})(x/\gamma)).$$

□

Semismooth mappings [28] are precisely Lipschitz continuous mappings for which the generalized Jacobian (and consequenlty the B-subdifferential) furnishes a first-order approximation.

Definition 3.5. *Let $G : \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz continuous at x. We say that G is* semismooth *at $\bar{x}$ if*

$$\|G(x) + H(\bar{x} - x) - G(\bar{x})\| = o(\|x - \bar{x}\|) \text{ as } x \to \bar{x},\ \forall H \in \partial_C G(x)$$

whereas G is said to be strongly semismooth *if $o(\|x - \bar{x}\|)$ can be replaced with $O(\|x - \bar{x}\|^2)$.*

We remark that the original definition of semismoothness given by [29] requires G to be directionally differentiable at x. The definition given here is the one employed by [30]. Another worth spent remark is that $\partial_C G(x)$ can be replaced with the smaller set $\partial_B G(x)$ in Definition 3.5.

Fortunately, the class of semismooth mappings is rich enough to include proximal mappings of most of the functions arising in interesting applications. For example *piecewise smooth (PC^1) mappings* are semismooth everywhere. Recall that a continuous mapping $G : \mathbb{R}^n \to \mathbb{R}^m$ is PC^1 if there exists a finite collection of smooth mappings $G_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, N$ such that

$$G(x) \in \{G_1(x), \ldots, G_N(x)\}, \quad \forall x \in \mathbb{R}^n.$$

The definition of PC^1 mappings given here is less general than the one of, e.g., [31, Ch. 4] but it suffices for our purposes. For every $x \in \mathbb{R}^n$ we introduce the set of essentially active indices

$$I_G^e(x) = \{i \in [N] \mid x \in \mathrm{cl}(\mathrm{int}\{x \mid G(x) = G_i(x)\})\}^5.$$

[5] $[N] \triangleq \{1, \ldots, N\}$ for any positive integer N.

In other words, $I_G^e(x)$ contains only indices of the pieces G_i for which there exists a full-dimensional set on which G agrees with G_i. In accordance to Definition 3.1, the generalized Jacobian of G at x is the convex hull of the Jacobians of the essentially active pieces, *i.e.*, [31, Prop. 4.3.1]

$$\partial_C G(x) = \operatorname{conv}\{\nabla G_i(x) \mid i \in I_G^e(x)\}. \tag{3.2}$$

As it will be clear in Section 5, in many interesting cases $\operatorname{prox}_{\gamma g}$ is PC^1 and thus semismooth. Furthermore, through (3.2) an element of $\partial_C \operatorname{prox}_{\gamma g}(x)$ can be easily computed once $\operatorname{prox}_{\gamma g}(x)$ has been computed.

A special but important class of convex functions whose proximal mapping is PC^1 are piecewise quadratic (PWQ) functions. A convex function $g \in \mathcal{S}^0(\mathbb{R}^n)$ is called PWQ if $\operatorname{dom} g$ can be represented as the union of finitely many polyhedral sets, relative to each of which $g(x)$ is given by an expression of the form $(1/2)x'Qx + q'x + c$ ($Q \in \mathbb{R}^{n\times n}$ must necessarily be symmetric positive semidefinite) [17, Def. 10.20]. The class of PWQ functions is quite general since it includes e.g. polyhedral norms, indicators and support functions of polyhedral sets, and it is closed under addition, composition with affine mappings, conjugation, inf-convolution and inf-projection [17, Prop. 10.22, Proposition 11.32]. It turns out that the proximal mapping of a PWQ function is *piecewise affine* (PWA) [17, 12.30] ($\mathbb{R}^n$ is partitioned in polyhedral sets relative to each of which $\operatorname{prox}_{\gamma g}$ is an affine mapping), hence strongly semismooth [13, Prop. 7.4.7]. Another example of a proximal mapping that it is strongly semismooth is the projection operator over symmetric cones [32]. We refer the reader to [33–36] for conditions that guarantee semismoothness of the proximal mapping for more general convex functions.

3.2. **Approximate generalized Hessian for F_γ.** Having established properties of generalized Jacobians for proximal mappings, we are now in position to construct a generalized Hessian for F_γ that will allow the development of Newton-like methods with fast asymptotic convergence rates. The obvious route to follow is to assume that ∇F_γ is semismooth and employ $\partial_C(\nabla F_\gamma)$ as a generalized Hessian for F_γ. However, semismoothness would require extra assumptions on f. Furthermore, the form of $\partial_C(\nabla F_\gamma)$ is quite complicated involving third-order partial derivatives of f. On the other hand, what is really needed to devise Newton-like algorithms with fast local convergence rates is a *linear Newton approximation (LNA)*, cf. Definition 3.6, at some stationary point of F_γ, which by Theorem 2.2(iv) is also a minimizer of F, provided that $\gamma \in (0, 1/L_f)$. The approach we follow is largely based on [37], [13, Prop. 10.4.4]. The following definition is taken from [13, Def. 7.5.13].

Definition 3.6. *Let $G : \mathbb{R}^n \to \mathbb{R}^m$ be continuous on $\mathbb{R}^n$. We say that G admits a linear Newton approximation at a vector $\bar{x} \in \mathbb{R}^n$ if there exists a set-valued mapping $\mathscr{G} : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n\times m}$ that has nonempty compact images, is upper semicontinuous at $\bar{x}$ and for any $H \in \mathscr{G}(x)$*

$$\|G(x) + H(\bar{x} - x) - G(\bar{x})\| = o(\|x - \bar{x}\|) \quad \text{as } x \to \bar{x}.$$

If instead

$$\|G(x) + H(\bar{x} - x) - G(\bar{x})\| = O(\|x - \bar{x}\|^2) \quad \text{as } x \to \bar{x},$$

then we say that G admits a strong linear Newton approximation at $\bar{x}$.

Arguably the most notable example of a LNA for semismooth mappings is the generalized Jacobian, cf. Definition 3.1. However, semismooth mappings can admit

LNAs different from the generalized Jacobian. More importantly, mappings that are not semismooth may also admit a LNA. It turns out that we can define a LNA for ∇F_γ at any stationary point, whose elements have a simpler form than those of $\partial_C(\nabla F_\gamma)$, without assuming semismoothness of ∇F_γ. We call it *approximate generalized Hessian* and it is given by

$$\hat{\partial}^2 F_\gamma(x) \triangleq \{\gamma^{-1}(I-\gamma\nabla^2 f(x))(I-P(I-\gamma\nabla^2 f(x))) \mid P \in \partial_C(\mathrm{prox}_{\gamma g})(x-\gamma\nabla f(x))\}.$$

The key idea in the definition of $\hat{\partial}^2 F_\gamma$, reminiscent to the Gauss-Newton method for nonlinear least-squares problems, is to omit terms vanishing at $x_\star$ that contain third-order derivatives of f. The following proposition shows that $\hat{\partial}^2 F_\gamma$ is indeed a LNA of ∇F_γ at any $x_\star \in X_\star$.

Proposition 3.7. *Let $T(x) = x - \gamma\nabla f(x)$, $\gamma \in (0, 1/L_f)$ and $x_\star \in X_\star$. Then*
- (i) *if $\mathrm{prox}_{\gamma g}$ is semismooth at $T(x_\star)$, then $\hat{\partial}^2 F_\gamma$ is a LNA for ∇F_γ at $x_\star$,*
- (ii) *if $\mathrm{prox}_{\gamma g}$ is strongly semismooth at $T(x_\star)$, and $\nabla^2 f$ is locally Lipschitz around $x_\star$, then $\hat{\partial}^2 F_\gamma$ is a strong LNA for ∇F_γ at $x_\star$.*

Proof. See Appendix. □

The next proposition shows that every element of $\hat{\partial}^2 F_\gamma(x)$ is a symmetric positive semidefinite matrix, whose eigenvalues are lower and upper bounded uniformly over all $x \in \mathbb{R}^n$.

Proposition 3.8. *Any $H \in \hat{\partial}^2 F_\gamma(x)$ is symmetric positive semidefinite and satisfies*

$$\xi_1\|d\|^2 \leq d'Hd \leq \xi_2\|d\|^2, \ \forall d \in \mathbb{R}^n, \tag{3.3}$$

where $\xi_1 \triangleq \min\{(1-\gamma\mu_f)\mu_f, (1-\gamma L_f)L_f\}$, $\xi_2 \triangleq \gamma^{-1}(1-\gamma\mu_f)$.

Proof. See Appendix. □

The next lemma shows uniqueness of the solution of (1.1) under a nonsingularity assumption on the elements of $\hat{\partial}^2 F_\gamma(x_\star)$. Its proof is similar to [13, Lem. 7.2.10], however ∇F_γ is not required to be locally Lipschitz around $x_\star$.

Lemma 3.9. *Let $x_\star \in X_\star$. Suppose that $\gamma \in (0, 1/L_f)$, $\mathrm{prox}_{\gamma g}$ is semismooth at $x_\star - \nabla f(x_\star)$ and every element of $\hat{\partial}^2 F_\gamma(x_\star)$ is nonsingular. Then $x_\star$ is the unique solution of (1.1). In fact, there exist positive constants δ and c such that*

$$\|x - x_\star\| \leq c\|G_\gamma(x)\|, \text{ for all } x \text{ with } \|x - x_\star\| \leq \delta.$$

Proof. See Appendix. □

4. Forward-Backward Newton-CG Methods

Having established the equivalence between minimizing F and F_γ, as well as a LNA for ∇F_γ, it is now very easy to design globally convergent Newton-like algorithms with fast asymptotic convergence rates, for computing a $x_\star \in X_\star$. Algorithm 1 is a standard line-search method for minimizing F_γ, where a conjugate gradient method is employed to solve (approximately) the corresponding regularized Newton system. Therefore our algorithm does not require to form an element of the generalized Hessian of F_γ explicitly. It only requires the computation of the corresponding matrix-vector product and is thus suitable for large-scale problems. Similarly, there is no need to form explicitly the Hessian of f, in order to compute the directional derivative $\nabla F_\gamma(x^k)'d^k$ needed in the backtracking procedure for computing the stepsize (4.4); only matrix-vector products with $\nabla^2 f(x)$ are required. Under nonsingularity of the elements of $\hat{\partial}^2 F_\gamma(x_\star)$, eventually the stepsize becomes equal to 1 and Algorithm 1 reduces to a regularized version of the (undamped) linear Newton method [13, Alg. 7.5.14] for solving $\nabla F_\gamma(x) = 0$.

Algorithm 1: Forward-Backward Newton-CG Method (FBN-CG I)

Input: $\gamma \in (0, 1/L_f)$, $\sigma \in (0, 1/2)$, $\bar{\eta} \in (0,1)$, $\zeta \in (0,1)$, $\rho \in (0,1]$, $x^0 \in \mathbb{R}^n$, $k = 0$

1 Select a $H^k \in \hat{\partial}^2 F_\gamma(x^k)$. Apply CG to

$$(H^k + \delta_k I)d^k = -\nabla F_\gamma(x^k) \tag{4.1}$$

to compute a $d^k \in \mathbb{R}^n$ that satisfies

$$\|(H^k + \delta_k I)d^k + \nabla F_\gamma(x^k)\| \leq \eta_k \|\nabla F_\gamma(x^k)\|, \tag{4.2}$$

where

$$\delta_k = \zeta \|\nabla F_\gamma(x^k)\|, \tag{4.3a}$$

$$\eta_k = \min\{\bar{\eta}, \|\nabla F_\gamma(x^k)\|^\rho\}. \tag{4.3b}$$

2 Compute $\tau_k = \max\{2^{-i} \mid i = 0, 1, 2, \ldots\}$ such that

$$F_\gamma(x^k + \tau_k d^k) \leq F_\gamma(x^k) + \sigma \tau_k \nabla F_\gamma(x^k)'d^k. \tag{4.4}$$

3 $x^{k+1} \leftarrow x^k + \tau_k d^k$

4 $k \leftarrow k + 1$ and go to Step 1.

The next theorem delineates the basic convergence properties of Algorithm 1.

Theorem 4.1. *Every accumulation point of the sequence $\{x^k\}$ generated by Algorithm 1 belongs to $X_\star$.*

Proof. We will first show that the sequence $\{d^k\}$ is *gradient related to* $\{x^k\}$ [20, Sec. 1.2]. That is, for any subsequence $\{x^k\}_{k\in\mathcal{N}}$ that converges to a nonstationary point of F_γ, *i.e.*,

$$\lim_{k\to\infty, k\in\mathcal{N}} \|\nabla F_\gamma(x^k)\| = \kappa \neq 0, \tag{4.5}$$

the corresponding subsequence $\{d^k\}_{k\in\mathcal{N}}$ is bounded and satisfies

$$\limsup_{k\to\infty, k\in\mathcal{N}} \nabla F_\gamma(x^k)'d^k < 0. \tag{4.6}$$

Without loss of generality we can restrict to subsequences for which $\nabla F_\gamma(x^k) \neq 0$, for all $k \in \mathcal{N}$. Suppose that $\{x^k\}_{k\in\mathcal{N}}$ is one such subsequence. Due to (4.3a), we have $\delta_k > 0$ for all $k \in \mathcal{N}$. Matrix H^k is positive semidefinite due to Proposition 3.8, therefore $H^k + \delta_k I$ is nonsingular for all $k \in \mathcal{N}$ and

$$\|(H^k + \delta_k I)^{-1}\| \leq \delta_k^{-1} = \frac{1}{\zeta \|\nabla F_\gamma(x^k)\|}.$$

Now, direction d^k satisfies

$$d^k = (H^k + \delta_k I)^{-1}(r^k - \nabla F_\gamma(x^k)),$$

where $r^k = (H^k + \delta_k I)d^k + \nabla F_\gamma(x^k)$. Therefore

$$\begin{aligned}\|d^k\| &\leq \|(H^k + \delta_k I)^{-1}\|(\|r^k\| + \|\nabla F_\gamma(x^k)\|) \\ &\leq \frac{1}{\zeta\|\nabla F_\gamma(x^k)\|}(\eta_k\|\nabla F_\gamma(x^k)\| + \|\nabla F_\gamma(x^k)\|) \leq (1+\bar{\eta})/\zeta,\end{aligned} \tag{4.7}$$

proving that $\{d^k\}_{k\in\mathcal{N}}$ is bounded. According to [38, Lemma A.2], when CG is applied to (4.1) we have that

$$\nabla F_\gamma(x^k)'d^k \leq -\frac{1}{\|H^k + \delta_k I\|}\|\nabla F_\gamma(x^k)\|^2. \tag{4.8}$$

Using (4.3a) and Proposition 3.8, we have that

$$\|H^k + \delta_k I\| \leq \gamma^{-1} + \zeta\|\nabla F_\gamma(x^k)\|,$$

therefore

$$\nabla F_\gamma(x^k)'d^k \leq -\frac{\|\nabla F_\gamma(x^k)\|^2}{\gamma^{-1} + \zeta\|\nabla F_\gamma(x^k)\|}, \quad \forall k \in \mathcal{N}, \tag{4.9}$$

As $k(\in \mathcal{N}) \to \infty$, the right hand side of (4.9) converges to $-\kappa^2/(\gamma^{-1} + \zeta\kappa)$, which is either a finite negative number (if κ is finite) or $-\infty$. In any case, this together with (4.9) confirm that (4.6) is valid as well, proving that $\{d^k\}$ is gradient related to $\{x^k\}$. All the assumptions of [20, Prop. 1.2.1] hold, therefore every accumulation point of $\{x^k\}$ converges to a stationary point of F_γ, which by Theorem 2.2(iv) is also a minimizer of F. □

The next theorem shows that under a nonsingularity assumption on $\hat{\partial}^2 F_\gamma(x_\star)$, the asymptotic rate of convergence of the sequence generated by Algorithm 1 is at least superlinear.

Theorem 4.2. *Suppose that $x_\star$ is an accumulation point of the sequence $\{x^k\}$ generated by Algorithm 1. If $\operatorname{prox}_{\gamma g}$ is semismooth at $x_\star - \gamma\nabla f(x_\star)$ and every element of $\hat{\partial}^2 F_\gamma(x_\star)$ is nonsingular, then the entire sequence converges to $x_\star$ and the convergence rate is Q-superlinear. Furthermore, if $\operatorname{prox}_{\gamma g}$ is strongly semismooth at $x_\star - \gamma\nabla f(x_\star)$ and $\nabla^2 f$ is locally Lipschitz continuous around $x_\star$ then $\{x^k\}$ converges to $x_\star$ with Q-order at least ρ.*

Proof. Theorem 4.1 asserts that $x_\star$ must be a stationary point for F_γ. Due to Proposition 3.7, $\hat{\partial}^2 F_\gamma$ is a LNA of ∇F_γ at $x_\star$. Due to Lemma 3.9, $x_\star$ is the globally unique minimizer of F. Therefore, by Theorem 4.1 every subsequence must converge to this unique accumulation point, implying that the entire sequence converges to $x_\star$. Furthermore, for any k

$$\begin{aligned}\|\nabla F_\gamma(x^k)\| &\leq \|I - \gamma\nabla^2 f(x^k)\|\|G_\gamma(x^k)\| \\ &\leq \|G_\gamma(x^k) - G_\gamma(x_\star)\| \leq 2\gamma^{-1}\|x^k - x_\star\|,\end{aligned} \tag{4.10}$$

where the second inequality follows from $G_\gamma(x_\star) = 0$ and Lemma A.2 (in the Appendix).

We know that d^k satisfies $(H^k+\delta_k I)d^k+\nabla F_\gamma(x^k)=r^k$. Therefore, for sufficiently large k, we have

$$\begin{aligned}
\|x^k+d^k-x_\star\| &= \|x^k+(H^k+\delta_k I)^{-1}(r^k-\nabla F_\gamma(x^k))-x_\star\|\\
&= \|(H^k+\delta_k I)^{-1}(H^k(x^k-x_\star)-\nabla F_\gamma(x^k)+\delta_k(x^k-x_\star)+r^k)\|\\
&\leq \|(H^k+\delta_k I)^{-1}\|\Big(\|H^k(x^k-x_\star)+\nabla F_\gamma(x_\star)-\nabla F_\gamma(x^k)\|\\
&\qquad\qquad\qquad + \delta_k\|x^k-x_\star\|+\|r^k\|\Big)\\
&\leq \kappa\Big(\|H^k(x^k-x_\star)+\nabla F_\gamma(x_\star)-\nabla F_\gamma(x^k)\|\\
&\qquad + 2\zeta\gamma^{-1}\|x^k-x_\star\|^2+\eta\gamma^{-1}\|x-x_\star\|^{1+\rho}\Big)
\end{aligned}\tag{4.11}$$

where the last inequality follows by (4.3a), (4.3b), (4.10). Therefore, since $\hat\partial^2 F_\gamma$ is a LNA of ∇F_γ at $x_\star$, we have

$$\|x^k+d^k-x_\star\| = o(\|x^k-x_\star\|),\tag{4.12}$$

while if it is a strong LNA we have

$$\|x^k+d^k-x_\star\| = O(\|x^k-x_\star\|^{1+\rho}).\tag{4.13}$$

In other words, $\{d^k\}$ is *superlinearly convergent with respect to* $\{x^k\}$ [13, Sec. 7.5]. Eventually, we have

$$\begin{aligned}
\nabla F_\gamma(x^k)'d^k+{d^k}'(H^k+\delta_k I)d^k &\leq \eta_k\|\nabla F_\gamma(x^k)\|\|d^k\| \leq \|\nabla F_\gamma(x^k)\|^{\rho+1}\|d^k\|\\
&\leq 2\gamma^{-(\rho+1)}\|x^k-x_\star\|^{\rho+1}\|d^k\|\\
&= O(\|d^k\|^{\rho+2}),
\end{aligned}\tag{4.14}$$

where the first inequality follows by (4.2), the second by (4.3b), the third inequality follows by (4.10) and the equality follows from the fact that $\{d^k\}$ is superlinearly convergent with respect to $\{x^k\}$, which implies $\|x^k-x_\star\|=O(\|d^k\|)$ [13, Lem. 7.5.7].

Since $\hat\partial^2 F_\gamma$ is a LNA of ∇F_γ at $x_\star$, it has nonempty compact images and is upper semicontinuous at $x_\star$. This, together with the fact that $\{x^k\}$ converges to $x_\star$ and the nonsingularity assumption on the elements of $\hat\partial^2 F_\gamma(x_\star)$ imply through [13, Lem. 7.5.2] that for sufficiently large k, H^k is nonsingular and there exists a $\kappa>0$ such that

$$\max\{\|H^k\|,\|H^k\|^{-1}\}\leq\kappa.$$

Therefore, eventually we have $\lambda_{\min}(H^k+\delta_k I)\geq\lambda_{\min}(H^k)\geq\kappa$. The last inequality together with (4.14) imply that there exists a $\theta>0$ such that eventually

$$\nabla F_\gamma(x^k)'d^k\leq-\theta\|d^k\|^2.\tag{4.15}$$

Following the same line of proof as in [13, Prop. 7.4.10], it can be shown that

$$\lim_{\substack{\|d\|\to 0\\ H\in\hat\partial^2 F_\gamma(x_\star+d)}}\frac{F_\gamma(x_\star+d)-F_\gamma(x_\star)-\nabla F_\gamma(x_\star)'d-\frac12 d'Hd}{\|d\|^2}=0.\tag{4.16}$$

We remark here that [13, Prop. 7.4.10] assumes semismoothness of ∇F_γ at $x_\star$ and proves (4.16) with $\partial_C(\nabla F_\gamma)$ in place of $\hat\partial^2 F_\gamma$, but exactly the same arguments apply for any LNA of ∇F_γ at $x_\star$ even without the semismoothness assumption.

Using (4.15), (4.16) and exactly the same arguments as in the proof of [13, Prop. 8.3.18(d)] or [39, Th. 3.2] we have that eventually

$$F_\gamma(x^k + d^k) \leq F_\gamma(x^k) + \sigma \nabla F_\gamma(x^k)' d^k, \tag{4.17}$$

which means that there exists a positive integer $\bar{k}$ such that $\tau_k = 1$, for all $k \geq \bar{k}$. Therefore, for all $k \geq \bar{k}$

$$x^{k+1} = x^k + d^k.$$

This together with (4.12), (4.13) proves the corresponding convergence rates for $\{x^k\}$. ☐

When f is strongly convex quadratic, Theorem 2.3 guarantees that F_γ is strongly convex and we can give a complexity estimate for Algorithm 1. In particular, the global convergence rate for the function values and the iterates is linear.

Theorem 4.3. *Suppose that f is quadratic and $\mu_f > 0$. If $\zeta = 0$ then*

$$F(P_\gamma(x^k)) - F_\star \leq r_{F_\gamma}(F_\gamma(x^0) - F_\star)), \tag{4.18a}$$

$$\|x^k - x_\star\|^2 \leq \frac{L_{F_\gamma}}{\mu_{F_\gamma}} r_{F_\gamma}^k \|x^0 - x_\star\|^2 \tag{4.18b}$$

where $r_{F_\gamma} = 1 - 2\left(\frac{\mu_{F_\gamma}}{L_{F_\gamma}}\right)^3 \frac{\sigma(1-\sigma)}{1+\eta}$.

Proof. See Appendix. ☐

Algorithm 1 exhibits fast asymptotic convergence rates provided that the elements of $\hat{\partial}^2 F_\gamma(x_\star)$ are nonsingular, but not much can be said about its global convergence rate, unless f is convex quadratic. Even in this favorable case the corresponding complexity estimates are very loose due to the variable metric used by the algorithm, cf. Theorem 4.3.

Another reason for the failure to derive meaningful complexity estimates is the fact that Algorithm 1 "forgets" about the convex structure of F, since it tries to minimize directly F_γ which can be nonconvex and its gradient may not be globally Lipschitz continuous. Specifically, Algorithm 1 may fail to be a descent method for F (although it satisfies that property for F_γ). Furthermore the iterates x^k produced by Algorithm 1 may lie outside $\operatorname{dom} g$ (but $P_\gamma(x^k) \in \operatorname{dom} g$, see Theorem 2.2(iii)). In this section, we show how Algorithm 1 can be modified so as to be able to derive global complexity estimates, similar to the ones for the proximal gradient method, and at the same time retain fast asymptotic convergence rates. The key idea is to inject a forward-backward step after the Newton step (cf. Alg. 2) and analyze the consequences of this choice on F, directly. This guarantees that the sequence of function values for both F and F_γ are monotone nonincreasing.

We show below that the sequence of iterates $\{x^k\}_{k\in\mathbb{N}}$ produced by Algorithm 2 enjoys the same favorable properties in terms of convergence and local convergence rates, as the one of Algorithm 1.

Theorem 4.4. *Every accumulation point of the sequence $\{x^k\}$ generated by Algorithm 2 belongs to $X_\star$.*

Proof. If $\mathcal{K} = \emptyset$ then Algorithm 2 is equivalent to FBS and the result has been already proved in [40, Th. 1.2]. Let us then assume $\mathcal{K} \neq \emptyset$ and distinguish between

Algorithm 2: Forward-Backward Newton-CG Method II (FBN-CG II)

Input: $\gamma \in (0, 1/L_f)$, $\sigma \in (0, 1/2)$, $\mathcal{K} \subseteq \mathbb{N}$, $k = 0$, $s_0 = 0$, $x^0 \in \operatorname{dom} g$

1 **if** $k \in \mathcal{K}$ *or* $s_k = 1$ **then**
2   Execute steps 1 and 2 of Algorithm 1 to compute direction d^k and step τ_k
3     $\hat{x}^k \leftarrow x^k + \tau_k d^k$
4   **if** $\tau_k = 1$ **then** $s_{k+1} \leftarrow 1$ **else** $s_{k+1} \leftarrow 0$
5 **else**
6     $\hat{x}^k \leftarrow x^k$, $s_{k+1} \leftarrow 0$
7 **end**
8 $x^{k+1} \leftarrow \operatorname{prox}_{\gamma g}(\hat{x}^k - \gamma \nabla f(\hat{x}^k))$
9 $k \leftarrow k + 1$ and go to Step 1.

two cases. First, we deal with the case where $k \notin \mathcal{K}$ and $s_k = 0$. Putting $x = \bar{x} = x^k$ in (A.3) we obtain

$$F(x^{k+1}) - F(x^k) \leq -\tfrac{\gamma}{2}\|G_\gamma(x^k)\|^2. \tag{4.19}$$

For the case where $k \in \mathcal{K}$ or $s_k = 1$, unless $\nabla F_\gamma(x^k) = 0$ (which means that x^k is a minimizer of F), we have $F_\gamma(\hat{x}^k) < F_\gamma(x^k)$ due to (4.4). Using parts (ii) and (iii) of Theorem 2.2 we obtain

$$\begin{aligned} F(x^{k+1}) &= F(P_\gamma(\hat{x}^k)) \leq F_\gamma(\hat{x}^k) \\ &\leq F_\gamma(x^k) \leq F(x^k) - \tfrac{\gamma}{2}\|G_\gamma(x^k)\|^2 \end{aligned}$$

and again we arrive at (4.19).

Summing up, Eq. (4.19) is satisfied for every $k \in \mathbb{N}$. Since $\{F(x^k)\}$ is monotonically nonincreasing, it converges to a finite value (since we have assumed that F is proper), therefore $\{F(x^k) - F(x^{k+1})\}$ converges to zero. This implies through (4.19) that $\{\|G_\gamma(x^k)\|^2\}$ converges to zero. Since $\|G_\gamma(\cdot)\|^2$ is a continuous nonnegative function which becomes zero if and only if $x \in X_\star$, it follows that every accumulation point of $\{x^k\}$ belongs to $X_\star$. ☐

Theorem 4.5. *Suppose $\mathcal{K}$ is infinite. Under the assumptions of Theorem 4.2 the same results apply also to the sequence of iterates produced by Algorithm 2.*

Proof. Following exactly the same steps as in the proof of Theorem 4.2 we can show that $\{d^k\}$ is superlinearly convergent with respect to $\{x^k\}$. Indeed, the derivation is independent of the algorithmic scheme and it is only related to how the direction d^k is generated. This means that unit stepsize is eventually accepted, *i.e.*, , there exists a positive integer $\bar{k}$ such that $s^k = 1$ for all $k \geq \bar{k}$. Therefore, eventually the iterates are given by

$$x^{k+1} = P_\gamma(x^k + d^k), \qquad k \geq \bar{k}.$$

Due to nonexpansiveness of P_γ we have

$$\|x^{k+1} - x_\star\| = \|P_\gamma(x^k + d^k) - P_\gamma(x_\star)\| \leq \|x^k + d^k - x_\star\|.$$

The proof finishes by invoking (4.11). ☐

As the next theorem shows, Algorithm 2 not only enjoys fast asymptotic convergence rate properties but also comes with the following global complexity estimate.

Theorem 4.6. *Let $\{x^k\}$ be a sequence generated by Algorithm 2. Assume that the level sets of F are bounded,* i.e., *$\|x - x_\star\| \leq R$ for some $x_\star \in X_\star$ and all $x \in \mathbb{R}^n$ with $F(x) \leq F(x^0)$. If $F(x^0) - F_\star \geq R^2/\gamma$ then*

$$F(x^1) - F_\star \leq \frac{R^2}{2\gamma}. \tag{4.20}$$

Otherwise, for any $k \in \mathbb{N}$ we have

$$F(x^k) - F_\star \leq \frac{2R^2}{\gamma(k+2)}. \tag{4.21}$$

Proof. See Appendix. □

When f is strongly convex the global rate of convergence is linear. The next theorem gives the corresponding complexity estimates.

Theorem 4.7. *If $f \in \mathcal{S}^{1,1}_{\mu_f, L_f}(\mathbb{R}^n)$, $\mu_f > 0$, then*

$$F\left(x^k\right) - F_\star \leq (1 + \gamma\mu_f)^{-k}(F(x^0) - F_\star), \tag{4.22a}$$

$$\|x^{k+1} - x_\star\|^2 \leq \frac{1 - \gamma\mu_f}{\gamma\mu_f(1 + \gamma\mu_f)^k}\|x^0 - x_\star\|^2. \tag{4.22b}$$

Proof. See Appendix. □

Remark 4.8. We should remark that Theorems 4.6 and 4.7 remain valid even if L_f (and thus γ) is unknown and instead a backtracking line search procedure similar to those described in [6, 7], is performed to determine a suitable value for γ.

5. Examples

In this section we discuss the generalized Jacobian of the proximal mapping of many relevant nonsmooth functions. Some of the considered examples will be particularly useful in Section 6 to test the effectiveness of Algorithms 1 and 2 on specific problems.

5.1. Indicator functions.

Constrained convex problems can be cast in the composite form (1.1) by encoding the feasible set D with the appropriate indicator function δ_D. Whenever Π_D, the projection onto D, is efficiently computable, then algorithms like the forward-backward splitting (2.1) can be conveniently considered. In the following we analyze the generalized Jacobian of some of such projections.

5.1.1. *Affine sets.* If $D = \{x \mid Ax = b\}$, $A \in \mathbb{R}^{m \times n}$, then $\Pi_D(x) = x - A^\dagger(Ax - b)$, where $A^\dagger$ is the Moore-Penrose pseudoinverse of A. For example if $m < n$ and A has full row rank, then $A^\dagger = A'(AA')^{-1}$. Obviously Π_D is an affine mapping, thus everywhere differentiable with

$$\partial_C(\Pi_D)(x) = \partial_B(\Pi_D)(x) = \{\nabla\Pi_D(x)\} = \{I - A^\dagger A\}. \tag{5.1}$$

5.1.2. *Polyhedral sets.* In this case $D = \{x \mid Ax = b,\ Cx \le d\}$, with $A \in \mathbb{R}^{m_1 \times n}$ and $C \in \mathbb{R}^{m_2 \times n}$. It is well known that Π_D is piecewise affine. In particular let

$$\mathscr{I}_D = \left\{ I \subseteq [m_2] \;\middle|\; \begin{array}{c} \text{there exists a vector } x \in \mathbb{R}^n \text{ with } Ax = b, \\ C_{i\cdot} x = d_i,\ i \in I,\ C_{j\cdot} x < d_j,\ j \in [m_2] \setminus I \end{array} \right\}$$

For each $I \in \mathscr{I}_D$ let

$$\begin{aligned} F_I &= \{x \in D \mid C_{i\cdot} x = d_i,\ i \in I\}, \\ S_I &= \operatorname{aff} F_I = \{x \in \mathbb{R}^n \mid Ax = b,\ C_{i\cdot} x = d_i,\ i \in I\}, \\ N_I &= \operatorname{cone}\left\{ \begin{bmatrix} A' & C'_{I\cdot} \end{bmatrix} \right\}, \\ C_I &= F_I + N_I. \end{aligned}$$

We then have $\Pi_D(x) \in \{\Pi_{S_I}(x) \mid I \in \mathscr{I}_D\}$, *i.e.*, Π_D is a piecewise affine function. The affine pieces of Π_D are the projections on the corresponding affine subspaces S_I, see Section 5.1.1. In fact for each $x \in C_I$ we have $\Pi_D(x) = \Pi_{S_I}(x)$, each C_I is full dimensional and $\mathbb{R}^n = \bigcup_{I \in \mathscr{I}_D} C_I$. For each $I \in \mathscr{I}_D$ let $P_I = \nabla \Pi_{S_I}$ and for each $x \in \mathbb{R}^n$ let $J(x) = \{I \in \mathscr{I}_D \mid x \in C_I\}$. Then

$$\partial_C(\Pi_D)(x) = \operatorname{conv} \partial_B(\Pi_D)(x) = \operatorname{conv}\{P_I \mid I \in J(x)\}.$$

Therefore, in order to determine an element P of $\partial_B(\Pi_D)(x)$ it suffices to compute $\bar{x} = \Pi_D(x)$ and take $P = I - B^\dagger B$, where

$$B = \begin{bmatrix} A \\ C_{I(x)\cdot} \end{bmatrix},$$

and $I(x) = \{i \in [n] \mid A_{i\cdot} \bar{x} = b_i\}$.

5.1.3. *Halfspaces.* We denote $(x)_+ = \max\{0, x\}$. If $D = \{x \mid a'x \le b\}$ then

$$\Pi_D(x) = x - \left(\frac{(a'x - b)_+}{\|a\|_2^2} \right) a$$

and

$$\partial_C(\Pi_D)(x) = \begin{cases} \{I - (1/\|a\|^2) aa'\}, & \text{if } a'x > b, \\ \{I\}, & \text{if } a'x < b, \\ \operatorname{conv}\{I, I - (1/\|a\|^2) aa'\}, & \text{if } a'x = b. \end{cases}$$

5.1.4. *Boxes.* Consider the box $D = \{x \mid \ell \le x \le u\}$, with $\ell_i \le u_i$. We have

$$\Pi_D(x) = \min\{\max\{x, \ell\}, u\}.$$

The corresponding indicator function δ_D is clearly separable, therefore (Prop. 3.3) every element $P \in \partial_B(\Pi_D)(x)$ is diagonal with

$$P_{ii} = \begin{cases} 1, & \text{if } \ell < x < u, \\ 0, & \text{if } x < \ell \text{ or } x > u, \\ \{0, 1\}, & \text{if } x = \ell \text{ or } x = u. \end{cases}$$

5.1.5. *Unit simplex.* When $D = \{x \mid x \geq 0,\ \sum_{i=1}^n x_i = 1\}$, one can easily see, by writing down the optimality conditions for the corresponding projection problem, that

$$\Pi_D(x) = (x - \lambda\mathbf{1})_+,$$

where λ solves $\mathbf{1}'(x - \lambda\mathbf{1})_+ = 1$. Since the unit simplex is a polyhedral set, we are dealing with a special case of Section 5.1.2, where $A = \mathbf{1}_n'$, $b = 1$, $C = -I_n$ and $d = 0$. Therefore, to calculate an element of the generalized Jacobian of the projection, we first compute $\Pi_D(x)$ and then determine the set of active indices $J = \{i \in [n] \mid (\Pi_D(x))_i = 0\}$. Let $n_J = |J|$ and $J_c = [n] \setminus J$. An element P of $\partial_B(\Pi_D)(x)$ is given by

$$P_{ij} = \begin{cases} 0, & \text{if } i, j \in J \\ -1/(n - n_J), & \text{if } i \neq j, i, j \in J_c, \\ 1 - 1/(n - n_J), & \text{if } i = j, i, j \in J_c. \end{cases}$$

Notice that P is block-diagonal after a permutation of rows and columns. The nonzero part $P_{J_cJ_c}$ is Toeplitz, so we can compute matrix vector products in $O(n_{J_c} \log n_{J_c})$ instead of $O(n_{J_c}^2)$ operations. Computing an element of the generalized Jacobian of the projection on $D = \{x \mid a'x = b,\ \ell \leq x \leq u\}$ can be treated in a similar fashion.

5.1.6. *Euclidean unit ball.* If $g = \delta_{B_2}$, where B_2 is the Euclidean unit ball then

$$\Pi_{B_2}(x) = \begin{cases} x/\|x\|_2, & \text{if } \|x\|_2 > 1, \\ x, & \text{otherwise} \end{cases}$$

and

$$\partial_C(\Pi_{B_2})(x) = \begin{cases} \{(1/\|x\|_2)(I - ww')\}, & \text{if } \|x\|_2 > 1, \\ \{I\}, & \text{if } \|x\|_2 < 1, \\ \operatorname{conv}\{(1/\|x\|_2)(I - ww'), I\}, & \text{otherwise,} \end{cases}$$

where $w = x/\|x\|_2^2$. Equality follows from the fact that $\Pi_{B_2} : \mathbb{R}^n \to \mathbb{R}^n$ is a piecewise smooth function.

5.1.7. *Second-order cone.* Given a point $x = (x_0, \bar{x}) \in \mathbb{R} \times \mathbb{R}^n$, each element of $V \in \partial_B(\Pi_K)(z)$ has the following representation [41, Lem. 2.6]:

$$V = 0 \text{ or } V = I_{n+1} \text{ or } V = \begin{bmatrix} 1 & \bar{w}' \\ \bar{w} & H \end{bmatrix},$$

for some vector $\bar{w} \in \mathbb{R}^n$ with $\|\bar{w}\|_2 = 1$ and some matrix $H \in \mathbb{R}^{n\times n}$ of the form

$$H = (1 + \alpha)I_n - \alpha\bar{w}\bar{w}', \quad |\alpha| \leq 1. \tag{5.2}$$

More precisely:

(i) if $x_0 \neq \pm\|\bar{x}\|_2$, then $\bar{w} = \bar{x}/\|\bar{x}\|,\ \alpha = x_0/\|\bar{x}\|$,
(ii) if $\bar{x} \neq 0$ and $x_0 = +\|\bar{x}\|_2$, then $\bar{w} = \bar{x}/\|\bar{x}\|,\ \alpha = +1$,
(iii) if $\bar{x} \neq 0$ and $x_0 = -\|\bar{x}\|_2$, then $\bar{w} = \bar{x}/\|\bar{x}\|,\ \alpha = -1$,
(iv) if $\bar{x} = 0$ and $x_0 = 0$, then either $V = 0$ or $V = I_{n+1}$ or it has H as in (5.2) for any $\bar{w}$ with $\|\bar{w}\| = 1$ and α with $|\alpha| \leq 1$.

5.2. **Vector norms.**

5.2.1. *Euclidean norm.* If $g(x) = \|x\|_2$ then the proximal mapping is given by

$$\mathrm{prox}_{\gamma g}(x) = \begin{cases} (1-\gamma/\|x\|_2)x, & \text{if } \|x\|_2 \geq \gamma, \\ 0, & \text{otherwise.} \end{cases}$$

Since $\mathrm{prox}_{\gamma g}$ is a PC^1 mapping, its B-subdifferential can be computed by simply computing the Jacobians of its smooth pieces. Specifically we have

$$\partial_B(\mathrm{prox}_{\gamma g})(x) = \begin{cases} \{I - \gamma/\|x\|_2\,(I - ww')\}, & \text{if } \|x\|_2 > \gamma, \\ \{0\}, & \text{if } \|x\|_2 < \gamma, \\ \{I - \gamma/\|x\|_2\,(I - ww')\,, 0\}, & \text{otherwise.} \end{cases}$$

where $w = x/\|x\|_2$.

5.2.2. ℓ_1 *norm.* The proximal mapping of $g(x) = \|x\|_1$ is the well known soft-thresholding operator

$$(\mathrm{prox}_{\gamma g}(x))_i = (\mathrm{sign}(x_i)(|x_i| - \gamma)_+)_i, \quad i \in [n].$$

Function g is separable, therefore according to Proposition 3.3 every element of $\partial_B(\mathrm{prox}_{\gamma g})$ is a diagonal matrix. The explicit form of the elements of $\partial_B(\mathrm{prox}_{\gamma g})$ is as follows. Let $\alpha = \{i \mid |x_i| > \gamma\}$, $\beta = \{i \mid |x_i| = \gamma\}$, $\delta = \{i \mid |x_i| < \gamma\}$. Then $P \in \partial_B(\mathrm{prox}_{\gamma g})(x)$ if and only if P is diagonal with elements

$$P_{ii} = \begin{cases} 1, & \text{if } i \in \alpha, \\ \in \{0,1\}, & \text{if } i \in \beta, \\ 0, & \text{if } i \in \delta. \end{cases}$$

We could also arrive to the same conclusion by applying Proposition 3.4 to the function of Section 5.1.4 with $u = -\ell = \mathbf{1}_n$, since the ℓ_1 norm is the conjugate of the indicator of the ℓ_∞ -norm ball.

5.2.3. *Sum of norms.* If $g(x) = \sum_{s \in \mathcal{S}} \|x_s\|_2$, where $\mathcal{S}$ is a partition of $[n]$, then

$$(\mathrm{prox}_{\gamma g}(x))_s = \left(1 - \frac{\gamma}{\|x_s\|_2}\right)_+ x_s,$$

for all $s \in \mathcal{S}$. Any $P \in \partial_B(\mathrm{prox}_{\gamma g})(x)$ is block diagonal with the s-th block equal to $I - \gamma/\|x_s\|_2 \left(I - (1/\|x_s\|_2^2)x_s x_s'\right)$, if $\|x_s\|_2 > \gamma$, I if $\|x_s\|_2 < \gamma$ and any of these two matrices if $\|x_s\|_2 = \gamma$.

5.3. **Support function.** Since $\sigma_C(x) = \sup_{y \in C} x'y$ is the conjugate of the indicator δ_C, one can use Proposition 3.4 to find that

$$\partial_B(\mathrm{prox}_{\gamma g})(x) = \{P = I - Q : Q \in \partial_B(\Pi_C)(x/\gamma)\}.$$

Depending on the specific set C (see Section 5.1) one obtains the appropriate subdifferential. A particular example is the following.

5.4. **Pointwise maximum.** Function $g(x) = \max\{x_1, \dots, x_n\}$ is conjugate to the indicator of the unit simplex already analyzed in Section 5.1.5. Applying Proposition 3.4 we obtain

$$\partial_B(\mathrm{prox}_{\gamma g})(x) = \{P = I - Q \mid Q \in \partial_B(\Pi_D)(x/\gamma)\}$$

Then $\Pi_D(x/\gamma) = (x/\gamma - \lambda\mathbf{1})_+$ where λ solves $\mathbf{1}'(x/\gamma - \lambda\mathbf{1})_+ = 1$. Let $J = \{i \in [n] \mid (\Pi_D(x/\gamma))_i = 0\}$, $n_J = |J|$ and $J_c = [n] \setminus J$. It follows that an element of $\partial_B(\mathrm{prox}_{\gamma g})(x)$ is block-diagonal (after a reordering of variables) with

$$P_{ij} = \begin{cases} 1, & \text{if } i, j \in J \\ 1 + 1/(n - n_J), & \text{if } i \neq j, i, j \in J_c, \\ 1/(n - n_J), & \text{if } i = j, i, j \in J_c. \end{cases}$$

5.5. **Spectral functions.** For any symmetric n by n matrix X, the eigenvalue function $\lambda : \mathbb{S}^n \to \mathbb{R}^n$ returns the vector of its eigenvalues in nonincreasing order. Now consider function $G : \mathbb{S}^n \to \bar{\mathbb{R}}$

$$G(X) = h(\lambda(X)), \quad X \in \mathbb{S}^n, \tag{5.3}$$

where $h : \mathbb{R}^n \to \bar{\mathbb{R}}$ is proper, closed, convex and symmetric, *i.e.*, invariant under coordinate permutations. Functions of this form are called *spectral functions* [42]. Being a spectral function, G inherits most of the properties of h [43, 44]. In particular, its proximal mapping is simply [45, Sec. 6.7]

$$\mathrm{prox}_{\gamma G}(X) = Q \operatorname{diag}(\mathrm{prox}_{\gamma h}(\lambda(X)))Q',$$

where $X = Q \operatorname{diag}(\lambda(X))Q'$ is the spectral decomposition of X (Q is an orthogonal matrix). Next, we further assume that

$$h(x) = g(x_1) + \cdots + g(x_N), \tag{5.4}$$

where $g : \mathbb{R} \to \bar{\mathbb{R}}$. Since h is also separable we have that

$$\mathrm{prox}_{\gamma h}(x) = (\mathrm{prox}_{\gamma g}(x_1), \dots, \mathrm{prox}_{\gamma g}(x_N)),$$

therefore the proximal mapping of G can be expressed as

$$\mathrm{prox}_{\gamma G}(X) = Q \operatorname{diag}(\mathrm{prox}_{\gamma g}(\lambda_1(X)), \dots, \mathrm{prox}_{\gamma g}(\lambda_n(X)))Q'. \tag{5.5}$$

Functions of this form are called *symmetric matrix-valued functions* [46, Chap. V], [47, Sec. 6.2]. Now we can use the theory of nonsmooth symmetric matrix-valued functions developed in [48] to analyze differentiability properties of $\mathrm{prox}_{\gamma G}$. In particular $\mathrm{prox}_{\gamma G}$ is (strongly) semismooth at X if and only if $\mathrm{prox}_{\gamma g}$ is (strongly) semismooth at the eigenvalues of X [48, Prop. 4.10]. Moreover, for any $X \in \mathbb{S}^n$ and $P \in \partial_B(\mathrm{prox}_{\gamma G})(X)$ we have [48, Lem. 4.7]

$$P(S) = Q(\Omega \circ (Q'SQ))Q', \ \forall S \in \mathbb{S}^n, \tag{5.6}$$

where $\circ$ denotes the Hadamard product and the matrix $\Omega \in \mathbb{R}^{n \times n}$ is defined by

$$\Omega_{ij} = \begin{cases} \frac{\mathrm{prox}_{\gamma g}(\lambda_i) - \mathrm{prox}_{\gamma g}(\lambda_j)}{\lambda_i - \lambda_j}, & \text{if } \lambda_i \neq \lambda_j, \\ \in \partial(\mathrm{prox}_{\gamma g})(\lambda_i), & \text{if } \lambda_i = \lambda_j. \end{cases} \tag{5.7}$$

5.5.1. *Indicator of the positive semidefinite cone.* The indicator of $\mathbb{S}^n_+$ can be expressed as in (5.3) with h given by (5.4) and $g = \delta_{\mathbb{R}_+}$. Then $\operatorname{prox}_{\gamma g}(x) = \Pi_{\mathbb{R}_+}(x) = (x)_+$ and according to (5.5) we have

$$\Pi_{\mathbb{S}^n_+}(X) = Q\operatorname{diag}((\lambda_1)_+, \dots, (\lambda_n)_+)Q'.$$

Let $\alpha = \{i \mid \lambda_i > 0\}$ and $\bar{\alpha} = [n] \setminus \alpha$. An element of $\partial_B \Pi_{\mathbb{S}^n_+}(X)$ is given by (5.6) with

$$\Omega = \begin{bmatrix} \Omega_{\alpha\alpha} & k_{\alpha\bar{\alpha}} \\ k'_{\alpha\bar{\alpha}} & 0 \end{bmatrix},$$

where $\Omega_{\alpha\alpha}$ is a matrix of ones and $k_{ij} = \frac{\lambda_i}{\lambda_i - \lambda_j}$, $i \in \alpha$, $j \in \bar{\alpha}$. In fact we have $P(S) = H + H'$ [49, Sec. 4] where

$$H = Q_\alpha\left(\tfrac{1}{2}(UQ_\alpha)Q'_\alpha + (k_{\alpha\bar{\alpha}} \circ (UQ_{\bar{\alpha}}))Q'_{\bar{\alpha}}\right)$$

and $U = Q'_\alpha S$. Therefore we can form $P(S)$ in at most $8|\alpha|n^2$ flops. When $|\alpha| > |\bar{\alpha}|$, we can alternatively express $P(S)$ as $S - Q'((E - \Omega) \circ (Q'SQ))Q'$, where E is a matrix of all ones and compute it in $8|\bar{\alpha}|n^2$ flops.

5.6. **Orthogonally invariant functions.** A function $G : \mathbb{R}^{m\times n} \to \bar{\mathbb{R}}$ is called *orthogonally invariant* if

$$G(UXV') = G(X),$$

for all $X \in \mathbb{R}^{m\times n}$ and all orthogonal matrices $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$. When the elements of X are allowed to be complex numbers then functions of this form are called *unitarily invariant* [50]. A function $h : \mathbb{R}^q \to \bar{\mathbb{R}}$ is *absolutely symmetric* if $h(Qx) = h(x)$ for all $x \in \mathbb{R}^p$ and any generalized permutation matrix Q, *i.e.*, a matrix $Q \in \mathbb{R}^{q\times q}$ that has exactly one nonzero entry in each row and each column, that entry being ± 1 [50]. There is a one-to-one correspondence between orthogonally invariant functions on $\mathbb{R}^{m\times n}$ and absolutely symmetric functions on $\mathbb{R}^q$. Specifically if G is orthogonally invariant then

$$G(X) = h(\sigma(X)),$$

for the absolutely symmetric function $h(x) = G(\operatorname{diag}(x))$. Here for $X \in \mathbb{R}^{m\times n}$, the spectral function $\sigma : \mathbb{R}^{m\times n} \to \mathbb{R}^q$, $q = \min\{m, n\}$ returns the vector of its singular values in nonincreasing order. Conversely, if h is absolutely symmetric then $G(X) = h(\sigma(X))$ is orthogonally invariant. Therefore, convex-analytic and generalized differentiability properties of orthogonally invariant functions can be easily derived from those of the corresponding absolutely symmetric functions [50]. For example, assuming for simplicity that $m \le n$, the proximal mapping of G is given by (see e.g. [45, Sec. 6.7])

$$\operatorname{prox}_{\gamma G}(X) = U\operatorname{diag}(\operatorname{prox}_{\gamma h}(\sigma(X)))V'_1,$$

where $X = U\begin{bmatrix}\operatorname{diag}(\sigma(X)), & 0\end{bmatrix}\begin{bmatrix}V_1, & V_2\end{bmatrix}'$ is the singular value decomposition of X. If we further assume that h is separable as in (5.4) then

$$\operatorname{prox}_{\gamma G}(X) = U\Sigma_g(X)V'_1, \tag{5.8}$$

where $\Sigma_g(X) = \operatorname{diag}(\operatorname{prox}_{\gamma g}(\sigma_1(X)), \dots, \operatorname{prox}_{\gamma g}(\sigma_n(X)))$. Functions of this form are called *nonsymmetric matrix-valued functions.* We also assume that g is a non-negative function such that $g(0) = 0$. This implies that $\operatorname{prox}_{\gamma g}(0) = 0$ and guarantees that the nonsymmetric matrix-valued function (5.8) is well-defined [51, Prop. 2.1.1]. Now we can use the results of [51, Ch. 2] to draw conclusions about generalized differentiability properties of $\operatorname{prox}_{\gamma G}$. For example, through [51, Th. 2.27] we have that $\operatorname{prox}_{\gamma G}$ is continuously differentiable at X if and only if $\operatorname{prox}_{\gamma g}$ is continuously differentiable at the singular values of X. Furthermore, $\operatorname{prox}_{\gamma G}$ is (strongly) semismooth at X if $\operatorname{prox}_{\gamma g}$ is (strongly) semismooth at the singular values of X [51, Th. 2.3.11].

For any $X \in \mathbb{R}^{m\times n}$ the generalized Jacobian $\partial_B(\operatorname{prox}_{\gamma G})(X)$ is well defined and nonempty and any $P \in \partial_B(\operatorname{prox}_{\gamma G})(X)$ acts on $H \in \mathbb{R}^{m\times n}$ as [51, Prop. 2.3.7]

$$P(H) = U\left[\left(\Omega_1 \circ \left(\tfrac{H_1+H_1'}{2}\right) + \Omega_2 \circ \left(\tfrac{H_1-H_1'}{2}\right)\right), \quad (\Omega_3 \circ H_2)\right]\begin{bmatrix}V_1, & V_2\end{bmatrix}' \tag{5.9}$$

where $H_1 = U'HV_1 \in \mathbb{R}^{m\times m}$, $H_2 = U'HV_2 \in \mathbb{R}^{m\times(n-m)}$ and $\Omega_1 \in \mathbb{R}^{m\times m}$, $\Omega_2 \in \mathbb{R}^{m\times m}$, $\Omega_3 \in \mathbb{R}^{m\times(n-m)}$ are given by

$$(\Omega_1)_{ij} = \begin{cases} \frac{\operatorname{prox}_{\gamma g}(\sigma_i)-\operatorname{prox}_{\gamma g}(\sigma_j)}{\sigma_i-\sigma_j}, & \text{if } \sigma_i \neq \sigma_j, \\ \in \partial \operatorname{prox}_{\gamma g}(\sigma_i), & \text{if } \sigma_i = \sigma_j, \end{cases}$$
$$(\Omega_2)_{ij} = \begin{cases} \frac{\operatorname{prox}_{\gamma g}(\sigma_i)-\operatorname{prox}_{\gamma g}(-\sigma_j)}{\sigma_i+\sigma_j}, & \text{if } \sigma_i \neq -\sigma_j, \\ \in \partial \operatorname{prox}_{\gamma g}(0), & \text{if } \sigma_i = \sigma_j = 0, \end{cases}$$
$$(\Omega_3)_{ij} = \begin{cases} \frac{\operatorname{prox}_{\gamma g}(\sigma_i)}{\sigma_i}, & \text{if } \sigma_i \neq 0, \\ \in \partial \operatorname{prox}_{\gamma g}(0), & \text{if } \sigma_i = 0. \end{cases}$$

5.6.1. *Nuclear norm.* For an m by n matrix X the nuclear norm, $G(X) = \|X\|_*$, is the sum of its singular values, *i.e.*, $G(X) = \sum_{i=1}^m \sigma_i(X)$ (we are again assuming, for simplicity, that $m \leq n$). The nuclear norm serves as a convex surrogate for the rank of a matrix. It has found many applications in systems and control theory, including system identification and model reduction [52–56]. Other fields of application include *matrix completion problems* arising in machine learning [57, 58] and computer vision [59, 60], and *nonnegative matrix factorization problems* arising in data mining [61].

The nuclear norm can be expressed as $G(X) = h(\sigma(X))$, where $h(x) = \|x\|_1$. Apparently, h is absolutely symmetric and separable. Specifically, it takes the form (5.4) with $g = |\cdot|$, for which $0 \in \operatorname{dom} g$ and $0 \in \partial g(0)$. The proximal mapping of the absolute value is the soft-thresholding operator. In fact, since the case of interest here is $x \geq 0$ (because $\sigma_i(X) \geq 0$), we have $\operatorname{prox}_{\gamma g}(x) = (x-\gamma)_+$. Consequently, the proximal mapping of $\|X\|_*$ is given by (5.8) with

$$\Sigma_g(X) = \operatorname{diag}((\sigma_1(X)-\gamma)_+, \ldots, (\sigma_m(X)-\gamma)_+).$$

For $x \in \mathbb{R}_+$ we have that

$$\partial(\operatorname{prox}_{\gamma g})(x) = \begin{cases} 0, & \text{if } 0 \leq x < \gamma, \\ [0,1], & \text{if } x = \gamma, \\ 1, & \text{if } x > \gamma. \end{cases} \tag{5.10}$$

Let $\alpha = \{i \mid \sigma_i(X) > \gamma\}$, $\beta = \{i \mid \sigma_i(X) = \gamma\}$ and $\delta = \{i \mid \sigma_i(X) < \gamma\}$. Taking into account (5.10), an element P of the B-subdifferential $\partial_B(\operatorname{prox}_{\gamma G})(X)$ satisfies (5.9) with

$$\Omega_1 = \begin{bmatrix} \omega^1_{\alpha\alpha} & \omega^1_{\alpha\beta} & \omega^1_{\alpha\delta} \\ (\omega^1_{\alpha\beta})' & \omega^1_{\beta\beta} & 0 \\ (\omega^1_{\alpha\delta})' & 0 & 0 \end{bmatrix}, \qquad \begin{array}{ll} \omega^1_{ij} = 1, & i \in \alpha, j \in \alpha \cup \beta, \\ \omega^1_{ij} = \frac{\sigma_i(X)-\gamma}{\sigma_i(X)-\sigma_j(X)}, & i \in \alpha, j \in \delta, \\ \omega^1_{ij} = \omega^1_{ji} = [0,1], & i,j \in \beta \end{array}$$

$$\Omega_2 = \begin{bmatrix} \omega^2_{\alpha\alpha} & \omega^2_{\alpha\beta} & \omega^2_{\alpha\delta} \\ (\omega^2_{\alpha\beta})' & 0 & 0 \\ (\omega^2_{\alpha\delta})' & 0 & 0 \end{bmatrix}, \qquad \omega^2_{ij} = \frac{(\sigma_i(X)-\gamma)_+ + (\sigma_j(X)-\gamma)_+}{\sigma_i(X)+\sigma_j(X)}, \quad i \in \alpha, j \in [m],$$

$$\Omega_3 = \begin{bmatrix} \omega^3_{\alpha[n-m]} \\ 0 \end{bmatrix}, \qquad \omega^3_{ij} = \frac{\sigma_i(X)-\gamma}{\sigma_i(X)}, \quad i \in \alpha, j \in [n-m].$$

6. Simulations

This section is devoted to the application of Algorithms 1 and 2 to some practical problems. Based on the results obtained in Section 5, we discuss the Newton system for each of the examples, and compare the proposed approach against other algorithms on the basis of numerical results obtained with Matlab.

6.1. Box constrained QPs.

A quadratic program with box constraints can be reformulated in the form (1.1) by adding to the cost the indicator of the feasible set, namely $\delta_{[l,u]}$. Then

$$f(x) = \frac{1}{2}x'Qx + q'x, \quad g(x) = \delta_{[l,u]}(x).$$

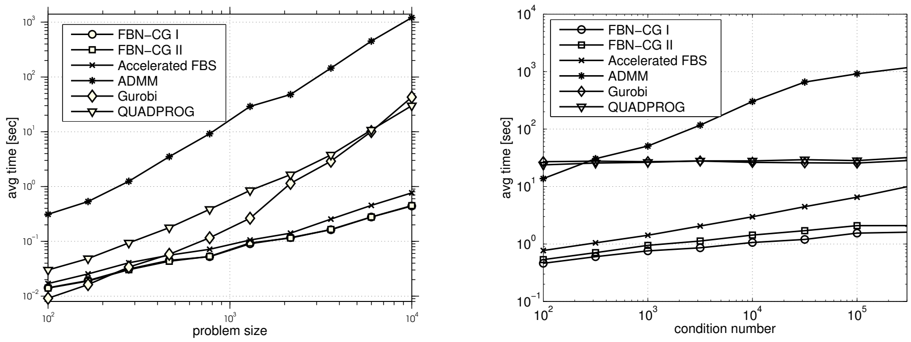
The B-subdifferential, in this case, is composed of diagonal matrices, with diagonal elements in $\{0,1\}$, cf. Section 5.1.4. More precisely, in Algorithm 1, we can split variable indices in the two sets

$$\alpha = \{i \mid l_i < [x - \gamma\nabla f(x)]_i < u_i\},$$
$$\bar{\alpha} = \{1,\dots,n\} \setminus \alpha,$$

and choose $P = \mathrm{diag}(p_1,\dots,p_n)$, with $p_i = 1$ if $i \in \alpha$ and $p_i = 0$ otherwise. Then the Newton system (4.1) reduces the triangular form

$$\begin{bmatrix} I_{|\bar{\alpha}|} & \\ \gamma Q_{\alpha\bar{\alpha}} & \gamma Q_{\alpha\alpha} \end{bmatrix} d^k = P_\gamma(x^k) - x^k.$$

This can be solved by forward substitution, where only the $|\alpha|$-by-$|\alpha|$ block is solved via CG. We tested the proposed algorithms against the commercial QP solver Gurobi, Matlab's built-in "quadprog" solver, the accelerated forward-backward splitting [6] (with constant stepsize) and the alternating directions method of multipliers (ADMM) [62]. The latter was both implemented using a direct solver, which requires the initial computation of the Cholesky factor of Q, and the conjugate gradient method. Random problems were generated with chosen size, density and condition number, as explained in [63]. Figures 1-2 show the results obtained: the proposed algorithms are generally faster then the others, and also appear to scale good with respect to problem size and condition number.



(A) Problem size (B) Condition number

FIGURE 1. Box constrained QPs. Average running times over a sample of 20 random instances, with increasing problem size and condition number.

6.2. **General QPs.** If we consider the more general quadratic programming problem with constraint $l \leq Ax \leq u$, $A \in \mathbb{R}^{m \times n}$, then the projection onto the feasible set is not explicitly computable like in the previous example. Formulating the Fenchel dual, and letting w be the dual variable, one can tackle the composite problem with

$$f(w) = \frac{1}{2}(A'w + q)'Q^{-1}(A'w + q), \qquad g(w) = \sigma_{[l,u]}(w).$$

Also in this case $\mathrm{prox}_{\gamma g}(w) = w - \Pi_{[\gamma l, \gamma u]}(w)$ has its B-subdifferential composed of binary diagonal matrices, cf. Section 5.3:

$$\begin{aligned} \bar{\alpha} &= \{i \mid \gamma l_i \leq [x - \gamma \nabla f(x)]_i \leq \gamma u_i\}, \\ \alpha &= \{1, \dots, n\} \setminus \alpha. \end{aligned}$$

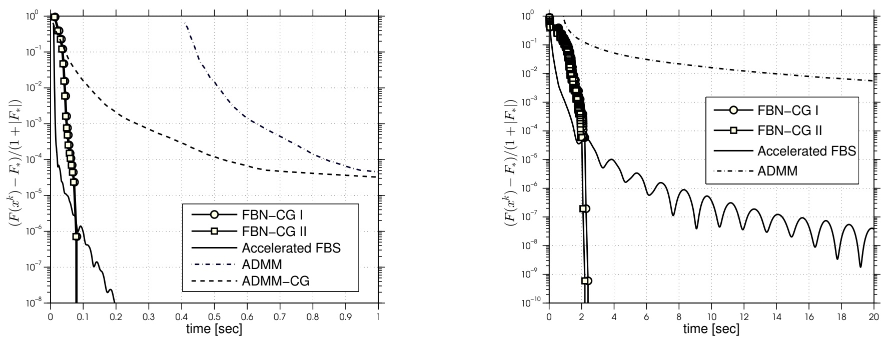
Choosing $P = \mathrm{diag}(p_1, \dots, p_n)$, with $p_i = 1$ if $i \in \alpha$ and $p_i = 0$ otherwise, just like in the previous case system (4.1) is block-triangular:

$$\begin{bmatrix} I_{|\bar{\alpha}|} & \\ \gamma A_\alpha Q^{-1} A'_{\bar{\alpha}} & \gamma A_\alpha Q^{-1} A'_\alpha \end{bmatrix} d = P_\gamma(w) - w.$$

Here subscripts denote *row* subsets. The latter is solved by forward substitution, and the $|\alpha|$-by-$|\alpha|$ block is solved via CG. Note that all the products with Q^{-1} are merely formal, and require a previous computation of the Cholesky factor of Q. Figure 2 compares Algorithm 1 and 2 to the accelerated version of FBS [6] and to ADMM [62], in terms of objective value decrease.

6.3. **ℓ_1-regularized least squares.** This is a classical problem arising in many fields like statistics, machine learning, signal and image processing. The purpose is to find a sparse solution to an underdetermined linear system. We have

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2, \qquad g(x) = \lambda \|x\|_1,$$



(A) Box constrained QP, $n = 1500$ and $\kappa = 10^4$.

(B) General QP, $n = 1000$, $m = 2000$ and $\kappa = 10^3$.

FIGURE 2. QPs. Comparison of the methods applied to a box constrained QP (primal) and to a general QP (dual).

where $A \in \mathbb{R}^{m\times n}$ with $m < n$. The ℓ_1-regularization term is known to promote sparsity in the solution vector x^*. As we mentioned in Section 5.2.2, the proximal mapping of the ℓ_1 norm is the soft-thresholding operator, whose generalized Jacobian is diagonal. Specifically, if

$$\alpha = \{i \mid |[x-\gamma\nabla f(x)]_i| > \gamma\lambda\},$$
$$\bar\alpha = \{1,\dots,n\}\setminus\alpha,$$

then $P = \operatorname{diag}(p_1,\dots,p_n)$, with $p_i = 1$ if $i \in \alpha$ and $p_i = 0$ otherwise, is an element of $\partial_B(\operatorname{prox}_{\gamma g})(x-\gamma\nabla f(x))$. The simplified system (4.1) reduces then to

$$\begin{bmatrix} I_{|\bar\alpha|} & \\ \gamma A_\alpha' A_{\bar\alpha} & \gamma A_\alpha' A_\alpha \end{bmatrix} d = P_\gamma(x) - x. \tag{6.1}$$

Here subscripts denote *column* subsets. The dimension of the problem to solve at each iteration is then $|\alpha|$: the smaller this set is, the cheaper the computation of the Newton direction is. Noting that at, any given x, larger values of λ allow for smaller size of α, and that decreasing λ decreases the objective value, we can set up a simple continuation scheme in order to keep the size of α small: starting from a relatively large value of $\lambda = \lambda_{\max} > \lambda_0$, we decrease it every time a certain criterion is met until $\lambda = \lambda_0$, using the solution of one step as to warm-start the next one. Specifically, we set $\lambda_{\max} = \|A'b\|_\infty$, which is the threshold above which the null solution is optimal. For an in-depth analysis of such continuation techniques on this type of problems, see [64]. We compared our method to SpaRSA [65], YALL1 [66] and l1_ls [67]. The algorithms were tested against the datasets available at `wwwopt.mathematik.tu-darmstadt.de/spear` [68]. These include datasets with different sizes and dynamic ranges of the solution. In each test we obtained a reference solution by running the method extensively, with a very small tolerance as stopping criterion. Then we set all the algorithms to stop as soon as the primal objective value reached a threshold at a relative distance $\epsilon_r = 10^{-8}$ from the reference solution. Figure 3 reports the performance profiles [69] of the algorithms considered on the aforementioned problem set. A point (r, f) on a line indicates

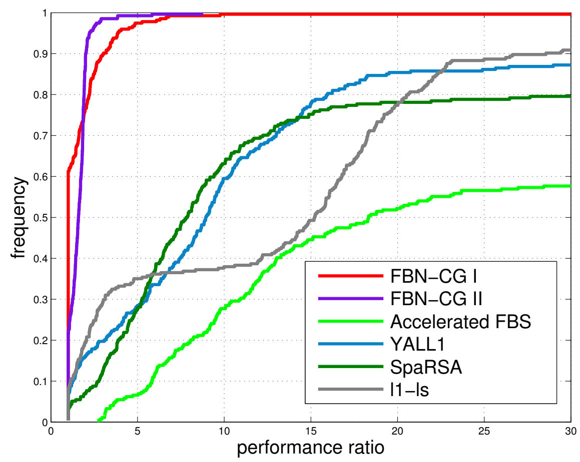


FIGURE 3. ℓ_1-regularized least squares. Performance profiles of the algorithms on the SPEAR test set with $\lambda_0 = 10^{-3}\lambda_{\max}$. The FBN-CG methods considered perform continuation on λ.

that the correspondent algorithm had a performance ratio[6] at most r in a fraction f of problems. It appears that the forward-backward Newton-CG method is very stable compared to the other algorithms considered. The benefits of the continuation scheme are evident from Figure 4, where the size of the linear system solved by FBN-CG at every iteration is shown.

6.4. **ℓ_1-regularized logistic regression.** This is another example of sparse fitting problem, although here the solution is used to perform binary classification. The composite objective function consists of

$$f(x) = \sum_{i=1}^{m} \log(1 + e^{-a_i' x}), \qquad g(x) = \lambda \|x_{[n-1]}\|_1,$$

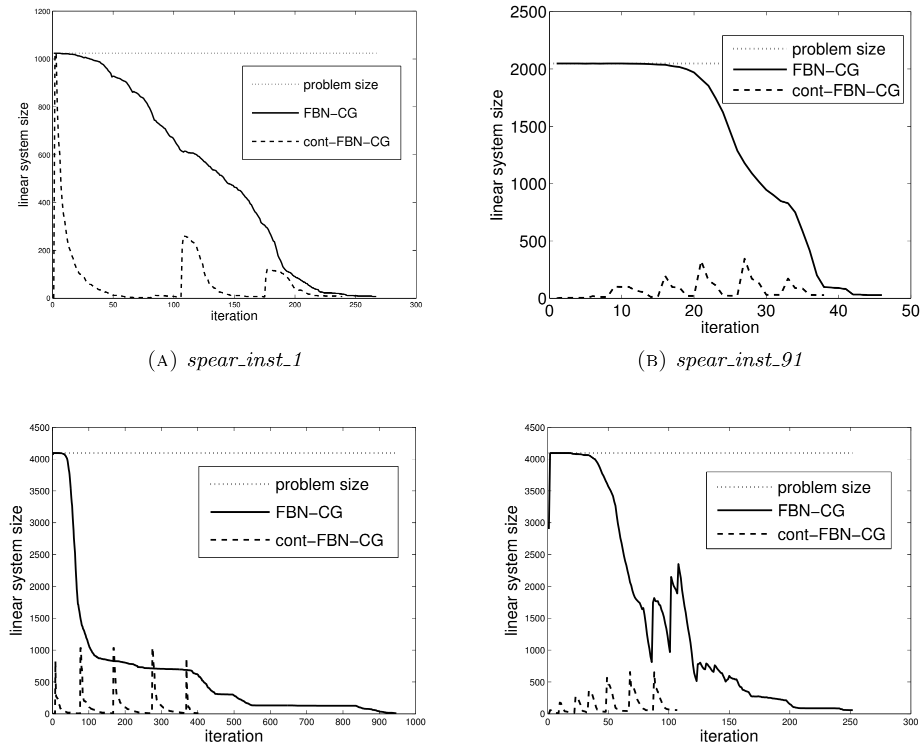
and again the ℓ_1-regularization enforces sparsity in the solution. We have

$$(\text{prox}_{\gamma g}(x))_i = \begin{cases} (\text{sign}(x_i)(|x_i| - \lambda\gamma)_+)_i, & i = 1, \ldots, n+1, \\ x_i & i = n. \end{cases}$$

Let $A \in \mathbb{R}^{m \times n}$ be the feature matrix with rows a_i having the trailing feature (the *bias* term) equals to one. If we set $\sigma(x) = (1 + e^{-Ax})^{-1}$ and let $\Sigma(x) = \text{diag}(\sigma(x) \circ (1 - \sigma(x)))$, then the Newton system (4.1) is

$$\begin{bmatrix} I_{|\bar{\alpha}|} & \\ \gamma A_\alpha' \Sigma(x) A_{\bar{\alpha}} & \gamma A_\alpha' \Sigma(x) A_\alpha \end{bmatrix} d = P_\gamma(x) - x,$$

[6]An algorithm has a performance ratio r, with respect to a problem, if its running time is r times the running time of the top performing algorithm among the ones considered.



(A) *spear_inst_1* (B) *spear_inst_91*

(C) *spear_inst_131* (D) *spear_inst_151*

FIGURE 4. ℓ_1-regularized least squares. Size of the linear system solved, by FBN-CG with and without warm-starting, compared to the full problem size.

where this time

$$\alpha = \{i \mid |[x - \gamma\nabla f(x)]_i| > \gamma\lambda\} \cup \{n\},$$
$$\bar{\alpha} = \{1, \ldots, n\} \setminus \alpha.$$

We compared FBN-CG to the accelerated FBS [6]. A continuation technique, similar to what described for the previous example, is employed in order to keep $|\alpha|$ small. As in the previous example, an approximate solution to the problem was first computed by means of extensive runs of one of the methods, and then the algorithms were set to stop once at a relative distance of $\epsilon_r = 10^{-8}$ from it. Table 1 shows how the methods scale with the number of features n, for sparse random datasets with $m = n/10$ and ≈ 50 nonzero features per row. The datasets were generated according to what described in [67, Sec. 4.2]. It is apparent how FBN-CG improves with respect to the accelerated version of forward-backward splitting.

6.5. **Matrix completion.** We consider the problem of recovering the entries of a matrix, which is known to have small rank, from a sample of them. One may refer to [70] for a detailed theoretical analysis of the problem. Since we are now dealing

	FBN-CG I		FBN-CG II		Accel. FBS	
n	time	iter.	time	iter.	time	iter.
100	0.04	51.1	0.05	57.3	0.06	292.4
215	0.05	52.8	0.06	61.0	0.11	462.1
464	0.06	54.4	0.09	69.4	0.18	647.2
1000	0.08	62.2	0.12	74.4	0.33	962.3
2154	0.27	98.8	0.35	108.2	0.82	1553.2
4641	0.95	151.1	0.94	142.2	3.58	2451.3
10000	2.40	217.7	2.54	207.0	9.36	3553.6

TABLE 1. ℓ_1-regularized logistic regression. Average running time (in seconds) and average number of iterations, for random datasets with $m = n/10$ and increasing n, $\lambda = 1$.

with matrix variables, we conveniently adopt the notation of *vector representation* of the matrix X, denoted by $\mathrm{vec}(X)$, *i.e.*, the mn-dimensional vector obtained by stacking the columns of X. The problem is formulated in a composite form as

$$f(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|^2, \qquad g(X) = \lambda\|X\|_*.$$

The linear mapping $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^k$ is represented as a k-by-mn matrix A acting on $\mathrm{vec}(X)$. The problem is nothing more than a least squares problem with a nuclear norm regularization term, having $\nabla f(X) = A'(A\,\mathrm{vec}(X) - b)$ and $\nabla^2 f(X) = A'A$. For a matrix completion task, matrix A is a binary matrix that selects k elements from X. Hence $\nabla^2 f(X)$ is actually diagonal, with k nonzero elements:

$$A'A = \mathrm{diag}(h_1, \ldots, h_{mn}), \quad h_i = \begin{cases} 1 & i \text{ is selected by } A, \\ 0 & \text{otherwise.} \end{cases}$$

The proximal mapping associated with $g = \|\cdot\|_*$ is the soft-thresholding applied to the singular values of the matrix argument. Its B-subdifferential elements act on m-by-n matrices as expressed in (5.9): if we consider, again, vector representations the linear mapping P is explicitly expressed by some symmetric and positive semidefinite matrix $Q \in \mathbb{R}^{mn\times mn}$ with eigenvalues in the interval $[0, 1]$. Hence we can express (4.1) as follows:

$$(G - GQG + \delta I)\,\mathrm{vec}(D) = -G\,\mathrm{vec}(X - P_\gamma(X)), \tag{6.2}$$

where

$$G = I - \gamma\nabla^2 f(x) = I_{mn} - \gamma A'A = \mathrm{diag}(g_1, \ldots, g_{mn}),$$

has diagonal elements $1 - \gamma$ and 1. Note however that we don't need to form the system (6.2) order compute residuals and carry out CG iterations, as matrix Q is indeed very large and dense. Instead, one can observe that pre-multiplication of $\mathrm{vec}(D)$ by a diagonal matrix G is equivalent to the Hadamard product $\widehat{G} \circ D$, where

$$\widehat{G} = \begin{bmatrix} g_1 & g_{m+1} & \cdots & g_{(n-1)m+1} \\ \vdots & \vdots & & \vdots \\ g_m & g_{2m} & \cdots & g_{nm} \end{bmatrix}.$$

Furthermore, with arguments similar to the ones in [49], the computational effort needed to evaluate P can be drastically reduced due to the sparsity pattern of

	m $(=n)$	density	iterations	SVDs	error
FBN-CG I	100	0.56	67.3	86.2	6.89e-04
	200	0.35	76.8	100.3	3.56e-04
	500	0.20	83.8	96.8	1.92e-04
FBN-CG II	100	0.56	54.1	126.1	6.89e-04
	200	0.35	65.6	153.3	3.56e-04
	500	0.20	71.0	151.2	1.92e-04
APGL	100	0.56	92.4	92.4	5.94e-04
	200	0.35	94.9	94.9	3.56e-04
	500	0.20	67.3	67.3	1.92e-04
LADM	100	0.56	183.2	183.2	4.58e-03
	200	0.35	494.2	494.2	7.57e-03
	500	0.20	1000.0	1000.0	2.70e-02

TABLE 2. Matrix completion. Average performance on 10 randomly generated instances M with $\operatorname{rank}(M) = 10$, $\lambda = 10^{-2}$. The density column refers to the fraction of observed coefficients. APGL and LADM require one SVD per iteration. The error reported is $\|X-M\|_F/\|M\|_F$, the relative distance X, the computed solution, and the original matrix M.

matrices $\Omega_1, \Omega_2, \Omega_3$ in (5.9). Hence it is convenient to compute residuals according to the following rewriting of (6.2):

$$\widehat{G} \circ D - \widehat{G} \circ P(\widehat{G} \circ D) + \delta D = -\widehat{G} \circ (X - P_\gamma(X)). \tag{6.3}$$

Even in this case, as in the previous examples, we can warm start our methods by approximately solving it for $\lambda_{\max} \geq \lambda > \lambda_0$ and updating λ in a continuation scheme until the final stage in which $\lambda = \lambda_0$.

We considered the accelerated proximal gradient with line search (APGL) [71] and the linearized alternating direction method (LADM) [72] in performing our tests. Both the methods also implement continuation on their parameters. Table 2 shows the average performance, in terms of number of iterations and SVD computations, on random matrices generated according to [71]. FBN-CG always succeeds at finding a low-error solution within a moderate number of iterations and SVD computations, which is not the case for LADM. Regarding APGL, it is worth noticing that it takes advantage from different acceleration techniques for this specific problem, which we have not considered for our algorithm. The drawback of our method is that at every iteration, the computation of (5.9) requires a full SVD as opposed to a decomposition in reduced form. Whether this can be avoided, and how this would affect the overall method, requires further investigation.

7. Conclusions and Future Work

In this paper we presented a framework, based on the continuously differentiable function (2.7) which we called *forward-backward envelope (FBE)*, to address a wide class of nonsmooth convex optimization problems in composite form. Problems of this form arise in many fields such as control, signal and image processing, system identification and machine learning. Using tools from nonsmooth analysis we derived two algorithms, namely FBN-CG I and II, that are Newton-like methods minimizing the FBE, for which we proved fast asymptotic convergence rates. Furthermore, Theorems 4.3, 4.6 and 4.7 provide global complexity estimates, making the algorithms also appealing for real-time applications. The considered approach makes it possible to exploit the sparsity patterns of many problems in the vicinity of the solution, so that the resulting Newton system is usually of small dimension for many significant problems. This also implies that the algorithms can favorably take advantage of warm-starting techniques. Our computational experience supports the theoretical results, and shows how in some scenarios our method challenges other well known approaches.

The framework we introduced opens up the possibility of extending many existing and well known algorithms, originally introduced for smooth unconstrained optimization, to the nonsmooth or constrained case. This is the case for example of Newton methods based on a trust-region approach, as well as quasi-Newton methods. Future work includes embedding the Newton iterations in accelerated versions of the forward-backward splitting, in order to obtain better global convergence rates. Finally, the extension of the framework to the nonconvex case (*i.e.*, to the case in which the smooth term f in (1.1) is nonconvex) can also be considered in order to address a wider range of applications.

References

[1] P. Patrinos and A. Bemporad, "Proximal Newton methods for convex composite optimization," in *IEEE Conference on Decision and Control*, 2013, pp. 2358–2363.

[2] J.-J. Moreau, "Proximité et dualité dans un espace Hilbertien," *Bull. Soc. Math. France*, vol. 93, pp. 273–299, 1965.

[3] G. Lan, Z. Lu, and R. Monteiro, "Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming," *Mathematical Programming*, vol. 126, no. 1, pp. 1–29, 2011.

[4] P.-L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.

[5] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212, 2011.

[6] Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, vol. 140, no. 1, pp. 125–161, 2013.

[7] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[8] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization," Department of Mathematics, University of Washington, Tech. Rep., 2008.

[9] S. Becker and M. J. Fadili, "A quasi-Newton proximal splitting method," in *Advances in Neural Information Processing Systems 25*, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012, vol. 1, pp. 2618–2626.

[10] J. Lee, Y. Sun, and M. Saunders, "Proximal Newton-type methods for convex optimization," in *Advances in Neural Information Processing Systems 25*, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012, vol. 1, pp. 827–835.

[11] M. Fukushima, "Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems," *Mathematical programming*, vol. 53, no. 1, pp. 99–110, 1992.

[12] N. Yamashita, K. Taji, and M. Fukushima, "Unconstrained optimization reformulations of variational inequality problems," *Journal of Optimization Theory and Applications*, vol. 92, no. 3, pp. 439–456, 1997.

[13] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003, vol. II.

[14] W. Li and J. Peng, "Exact penalty functions for constrained minimization problems via regularized gap function for variational inequalities," *Journal of Global Optimization*, vol. 37, pp. 85–94, 2007.

[15] P. Patrinos, P. Sopasakis, and H. Sarimveis, "A global piecewise smooth Newton method for fast large-scale model predictive control," *Automatica*, vol. 47, pp. 2016–2022, 2011.
[16] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces.* Springer, 2011.
[17] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis.* Springer, 2011, vol. 317.
[18] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.
[19] Y. Nesterov, *Introductory lectures on convex optimization: A basic course.* Springer, 2003, vol. 87.
[20] D. Bertsekas, *Nonlinear programming.* Athena Scientific, 1999.
[21] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM Journal on Control and Optimization*, vol. 14, no. 5, pp. 877–898, 1976.
[22] M. Fukushima and L. Qi, "A globally and superlinearly convergent algorithm for nonsmooth convex minimization," *SIAM Journal on Optimization*, vol. 6, no. 4, pp. 1106–1120, 1996.
[23] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, "A family of variable metric proximal methods," *Mathematical Programming*, vol. 68, no. 1-3, pp. 15–47, 1995.
[24] R. Mifflin, D. Sun, and L. Qi, "Quasi-Newton bundle-type methods for nondifferentiable convex optimization," *SIAM Journal on Optimization*, vol. 8, no. 2, pp. 583–603, 1998.
[25] N. Sagara and M. Fukushima, "A trust region method for nonsmooth convex optimization," *Management*, vol. 1, no. 2, pp. 171–180, 2005.
[26] C. Lemaréchal and C. Sagastizábal, "Practical aspects of the Moreau–Yosida regularization: Theoretical preliminaries," *SIAM Journal on Optimization*, vol. 7, no. 2, pp. 367–385, 1997.
[27] F. Clarke, *Optimization and nonsmooth analysis.* New York: Wiley, 1983.
[28] L. Qi and J. Sun, "A nonsmooth version of Newton's method," *Mathematical programming*, vol. 58, no. 1-3, pp. 353–367, 1993.
[29] R. Mifflin, "Semismooth and semiconvex functions in constrained optimization," *SIAM Journal on Control and Optimization*, vol. 15, no. 6, pp. 959–972, 1977.
[30] M. S. Gowda, "Inverse and implicit function theorems for H-differentiable and semismooth functions," *Optimization Methods and Software*, vol. 19, no. 5, pp. 443–461, 2004.
[31] S. Scholtes, *Introduction to piecewise differentiable equations.* Springer, 2012.
[32] D. Sun and J. Sun, "Semismooth matrix-valued functions," *Mathematics of Operations Research*, vol. 27, no. 1, pp. 150–169, 2002.
[33] R. Mifflin, L. Qi, and D. Sun, "Properties of the Moreau-Yosida regularization of a piecewise C^2 convex function," *Mathematical programming*, vol. 84, no. 2, pp. 269–281, 1999.
[34] F. Meng, G. Zhao, M. Goh, and R. De Souza, "Lagrangian-dual functions and Moreau-Yosida regularization," *SIAM Journal on Optimization*, vol. 19, no. 1, pp. 39–61, 2008.
[35] F. Meng, "Moreau–Yosida regularization of Lagrangian-dual functions for a class of convex optimization problems," *Journal of Global Optimization*, vol. 44, no. 3, pp. 375–394, 2009.
[36] F. Meng, D. Sun, and G. Zhao, "Semismoothness of solutions to generalized equations and the Moreau-Yosida regularization," *Mathematical programming*, vol. 104, no. 2, pp. 561–581, 2005.
[37] D. Sun, M. Fukushima, and L. Qi, "A computable generalized Hessian of the D-gap function and Newton-type methods for variational inequality problems," in *Complementarity and Variational Problems: State of the Art*, M. Ferris and J. Pang, Eds. SIAM Publications, 1997, pp. 452–473.
[38] R. S. Dembo and T. Steihaug, "Truncated-Newton algorithms for large-scale unconstrained optimization," *Mathematical Programming*, vol. 26, no. 2, pp. 190–212, 1983.
[39] F. Facchinei, "Minimization of SC^1 functions and the Maratos effect," *Operations Research Letters*, vol. 17, no. 3, pp. 131–137, 1995.
[40] A. Beck and M. Teboulle, "Gradient-based algorithms with applications to signal recovery problems," in *Convex Optimization in Signal Processing and Communications*, D. Palomar and Y. Eldar, Eds. Cambridge University Press, 2010, pp. 42–88.
[41] C. Kanzow, I. Ferenczi, and M. Fukushima, "On the local convergence of semismooth Newton methods for linear and nonlinear second-order cone programs without strict complementarity," *SIAM Journal on Optimization*, vol. 20, no. 1, pp. 297–320, 2009.
[42] A. S. Lewis, "Convex analysis on the Hermitian matrices," *SIAM Journal on Optimization*, vol. 6, no. 1, pp. 164–177, 1996.

[43] ——, “Derivatives of spectral functions,” *Mathematics of Operations Research*, vol. 21, no. 3, pp. 576–588, 1996.
[44] A. S. Lewis and H. S. Sendov, “Twice differentiable spectral functions,” *SIAM Journal on Matrix Analysis and Applications*, vol. 23, no. 2, pp. 368–386, 2001.
[45] N. Parikh and S. Boyd, “Proximal algorithms,” *Foundations and Trends in Optimization*, pp. 1–96, 2013.
[46] R. Bhatia, *Matrix analysis*. Springer, 1997, vol. 169.
[47] R. A. Horn, *Topics in matrix analysis*. Cambridge university press, 1991.
[48] X. Chen, H. Qi, and P. Tseng, “Analysis of nonsmooth symmetric-matrix-valued functions with applications to semidefinite complementarity problems,” *SIAM Journal on Optimization*, vol. 13, no. 4, pp. 960–985, 2003.
[49] X.-Y. Zhao, D. Sun, and K.-C. Toh, “A Newton-CG augmented Lagrangian method for semidefinite programming,” *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1737–1765, 2010.
[50] A. S. Lewis, “The convex analysis of unitarily invariant matrix functions,” *Journal of Convex Analysis*, vol. 2, no. 1, pp. 173–183, 1995.
[51] Z. Yang, “A study on nonsymmetric matrix-valued functions,” Master’s thesis, Department of Mathematics, National University of Singapore, 2009.
[52] M. Fazel, H. Hindi, and S. P. Boyd, “A rank minimization heuristic with application to minimum order system approximation,” in *American Control Conference. Proceedings of the 2001*, vol. 6. IEEE, 2001, pp. 4734–4739.
[53] M. Fazel, “Matrix rank minimization with applications,” Ph.D. dissertation, Stanford University, 2002.
[54] M. Fazel, H. Hindi, and S. Boyd, “Rank minimization and applications in system theory,” in *American Control Conference. Proceedings of the 2004*, vol. 4. IEEE, 2004, pp. 3273–3278.
[55] Z. Liu and L. Vandenberghe, “Interior-point method for nuclear norm approximation with application to system identification,” *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1235–1256, 2009.
[56] B. Recht, M. Fazel, and P. A. Parrilo, “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.
[57] N. Srebro, “Learning with matrix factorizations,” Ph.D. dissertation, Massachusetts Institute of Technology, 2004.
[58] J. D. Rennie and N. Srebro, “Fast maximum margin matrix factorization for collaborative prediction,” in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 713–719.
[59] C. Tomasi and T. Kanade, “Shape and motion from image streams under orthography: a factorization method,” *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.
[60] T. Morita and T. Kanade, “A sequential factorization method for recovering shape and motion from image streams,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 8, pp. 858–867, 1997.
[61] L. Eldén, *Matrix methods in data mining and pattern recognition*. SIAM, 2007.
[62] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.
[63] C. C. Gonzaga, E. W. Karas, and D. R. Rossetto, “An optimal algorithm for constrained differentiable convex optimization,” *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 1939–1955, 2013.
[64] L. Xiao and T. Zhang, “A proximal-gradient homotopy method for the sparse least-squares problem,” *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1062–1091, 2013.
[65] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, “Sparse reconstruction by separable approximation,” *Signal Processing, IEEE Transactions on*, vol. 57, no. 7, pp. 2479–2493, 2009.
[66] J. Yang and Y. Zhang, “Alternating direction algorithms for ℓ_1-problems in compressive sensing,” *SIAM Journal on Scientific Computing*, vol. 33, no. 1, pp. 250–278, 2011.
[67] K. Koh, S.-J. Kim, and S. P. Boyd, “An interior-point method for large-scale ℓ_1-regularized logistic regression.” *Journal of Machine learning research*, vol. 8, no. 8, pp. 1519–1555, 2007.
[68] D. Lorenz, “Constructing test instances for basis pursuit denoising,” *IEEE Transactions on Signal Processing*, vol. 61, no. 5, pp. 1210–1214, 2013.

[69] E. D. Dolan and J. J. Moré, "Benchmarking optimization software with performance profiles," *Mathematical Programming*, vol. 91, no. 2, pp. 201–213, Jan. 2002.
[70] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.
[71] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific Journal of Optimization*, vol. 6, no. 3, pp. 615–640, 2010.
[72] J. Yang and X. Yuan, "Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization," *Mathematics of Computation*, vol. 82, no. 281, pp. 301–329, 2013.

Appendix A.

We provide results that are used throughout the paper and all the omitted proofs. The following result is useful in bounding the eigenvalues of the linear Newton approximation of F_γ, and is required by Theorem 2.3 and Proposition 3.8.

Lemma A.1. *If $Q \in \mathbb{S}^n_+$ and $\mu_f = \lambda_{\min}(Q)$, $L_f = \lambda_{\max}(Q)$ then*

$$\lambda_{\min}(Q(I-\gamma Q)) = \begin{cases} \mu_f(1-\gamma\mu_f), & \text{if } 0 < \gamma \leq 1/(L_f+\mu_f), \\ L_f(1-\gamma L_f), & \text{if } 1/(L_f+\mu_f) \leq \gamma < 1/L_f. \end{cases}$$

Proof. Since Q is symmetric positive semidefinite, there exists an invertible matrix $S \in \mathbb{R}^{n\times n}$ such that $Q = SJS^{-1}$, where $J = \operatorname{diag}(\lambda_1(Q), \dots, \lambda_n(Q))$. Therefore,

$$\begin{aligned} Q(I-\gamma Q) &= SJS^{-1}(I - \gamma SJS^{-1}) \\ &= SJS^{-1}S(I-\gamma J)S^{-1} \\ &= SJ(I-\gamma J)S^{-1}, \end{aligned}$$

and the eigenvalues of $Q(I-\gamma Q)$ are exactly

$$\lambda_1(Q)(1-\gamma\lambda_1(Q)), \dots, \lambda_n(Q)(1-\gamma\lambda_n(Q)).$$

Next, consider the minimization problem $\min_{\lambda\in[\mu_f,L_f]} \phi(\lambda) \triangleq \lambda(1-\gamma\lambda)$. Since γ is positive, ϕ is concave and the minimum is attained either at μ_f or L_f. The proof finishes by noticing that

$$\mu_f(1-\gamma\mu_f) \leq L_f(1-\gamma L_f) \Leftrightarrow \gamma \in (0, 1/(L_f+\mu_f)).$$

□

The next result gives condition for the Lipschitz-continuity of P_γ and G_γ, and is needed by Theorem 2.3 to obtain the Lipschitz constant of ∇F_γ in the case where f is quadratic, and by Theorem 4.2 and 4.5 in order to assess the local convergence properties of Algorithm 1 and 2.

Lemma A.2. *Suppose that $\gamma < 1/L_f$. Then $P_\gamma : \mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive, i.e.,*

$$\|P_\gamma(x) - P_\gamma(y)\| \leq \|x-y\|, \tag{A.1}$$

and $G_\gamma : \mathbb{R}^n \to \mathbb{R}^n$ is $(2/\gamma)$-Lipschitz continuous, i.e.,

$$\|G_\gamma(x) - G_\gamma(y)\| \leq 2/\gamma\|x-y\|. \tag{A.2}$$

Proof. On one hand we know that $\operatorname{prox}_{\gamma g}$ is firmly nonexpansive [2], therefore $\operatorname{prox}_{\gamma g}$ is a 1/2-averaged operator [16, Rem. 4.24(iii)]. On the other hand, being ∇f the Lipschitz continuous gradient of a convex function, it is $1/L_f$-cocoercive. Therefore, since $\gamma < 1/L_f$, the operator $x \to x - \gamma\nabla f(x)$ is $\gamma L_f/2$-averaged [16, Prop. 4.33]. Since P_γ is the composition of two averaged operators, it is an averaged operator as well [16, Prop. 4.32]. By [16, Rem. 4.24(i)] this implies that P_γ is nonexpansive, proving (A.1). Next, consider any $x, y \in \mathbb{R}^n$:

$$\begin{aligned}\|G_\gamma(x) - G_\gamma(y)\| &\leq 1/\gamma \|P_\gamma(x) - P_\gamma(y) - (x-y)\| \\ &\leq 1/\gamma \left(\|P_\gamma(x) - P_\gamma(y)\| + \|x-y\|\right) \\ &\leq 2/\gamma \|x-y\|.\end{aligned}$$

□

The following proposition is an extension of [7, Lemma 2.3] that handles the case where f can be strongly convex.

Proposition A.3. *For any $\gamma \in (0, 1/L_f]$, $x \in \mathbb{R}^n$, $\bar{x} \in \mathbb{R}^n$*

$$F(x) \geq F(P_\gamma(\bar{x})) + G_\gamma(\bar{x})'(x-\bar{x}) + \tfrac{\gamma}{2}\|G_\gamma(\bar{x})\|^2 + \tfrac{\mu_f}{2}\|x-\bar{x}\|^2.$$

Proof. For any $x \in \mathbb{R}^n$, $\bar{x} \in \mathbb{R}^n$ we have

$$\begin{aligned}F(x) &\geq f(\bar{x}) + \nabla f(\bar{x})'(x-\bar{x}) + \tfrac{\mu_f}{2}\|x-\bar{x}\|^2 \\ &\qquad + g(P_\gamma(\bar{x})) + (G_\gamma(\bar{x}) - \nabla f(\bar{x}))'(x - P_\gamma(\bar{x})) \\ &= f(\bar{x}) + g(P_\gamma(\bar{x})) - \nabla f(\bar{x})'(\bar{x} - P_\gamma(\bar{x})) + G_\gamma(\bar{x})'(x - P_\gamma(\bar{x})) + \tfrac{\mu_f}{2}\|x-\bar{x}\|^2 \\ &= F_\gamma(\bar{x}) - \tfrac{\gamma}{2}\|G_\gamma(\bar{x})\|^2 + G_\gamma(\bar{x})'(\bar{x} - P_\gamma(\bar{x})) + G_\gamma(\bar{x})'(x-\bar{x}) + \tfrac{\mu_f}{2}\|x-\bar{x}\|^2 \\ &= F_\gamma(\bar{x}) - \tfrac{\gamma}{2}\|G_\gamma(\bar{x})\|^2 + \gamma\|G_\gamma(\bar{x})\|^2 + G_\gamma(\bar{x})'(x-\bar{x}) + \tfrac{\mu_f}{2}\|x-\bar{x}\|^2 \\ &\geq F(P_\gamma(\bar{x})) + \tfrac{\gamma}{2}(2-\gamma L_f)\|G_\gamma(\bar{x})\|^2 + G_\gamma(\bar{x})'(x-\bar{x}) + \tfrac{\mu_f}{2}\|x-\bar{x}\|^2.\end{aligned}$$

The first inequality follows by strong convexity of f and $G_\gamma(\bar{x}) - \nabla f(\bar{x}) \in \partial g(P_\gamma(\bar{x}))$, the equality by the definition of F_γ and the final inequality by Theorem 2.2(iii). The result follows by noticing that $\gamma \in (0, 1/L_f]$ implies $2 - \gamma L_f \geq 1$. □

An immediate result of Proposition A.3 is the following.

Corollary A.4. *For any $\gamma \in (0, 1/L_f]$, $x \in \mathbb{R}^n$, it holds*

$$\|G_\gamma(x)\|^2 \geq 2\mu_f(F(P_\gamma(x)) - F_\star).$$

Proof. According to Proposition A.3, if $\gamma \in (0, 1/L_f]$ then for any $x, \bar{x} \in \mathbb{R}^n$ we certainly have

$$F(x) \geq F(P_\gamma(\bar{x})) + G_\gamma(\bar{x})'(x-\bar{x}) + \tfrac{\mu_f}{2}\|x-\bar{x}\|^2. \tag{A.3}$$

Minimizing both sides with respect to x we obtain $F_\star$ for the left hand side and $x = \bar{x} - \mu_f^{-1}G_\gamma(\bar{x})$ for the right hand side. Substituting in (A.3) we obtain

$$F_\star \geq F(P_\gamma(\bar{x})) - \tfrac{1}{2\mu_f}\|G_\gamma(\bar{x})\|^2.$$

□

The next proposition is useful for proving the global linear convergence rate of Algorithm 2, in the case of f strongly convex, cf. Theorem 4.7.

Proposition A.5. *For any $x \in \mathbb{R}^n$, $x_\star \in X_\star$ and $\gamma \in (0, 1/L_f]$*

$$F(P_\gamma(x)) - F_\star \leq \tfrac{1}{2\gamma}(1-\gamma\mu_f)\|x - x_\star\|^2.$$

Proof. By definition of F_γ we have

$$\begin{aligned}F_\gamma(x) &= \min_{z\in\mathbb{R}^n}\left\{f(x)+\nabla f(x)'(z-x)+g(z)+\tfrac{1}{2\gamma}\|z-x\|^2\right\}\\ &\le f(x)+\nabla f(x)'(x_\star-x)+g(x_\star)+\tfrac{1}{2\gamma}\|x_\star-x\|^2\\ &\le f(x_\star)+g(x_\star)-\tfrac{\mu_f}{2}\|x-x_\star\|^2+\tfrac{1}{2\gamma}\|x_\star-x\|^2,\end{aligned}$$

where the second inequality follows from (strong) convexity of f. The proof finishes by invoking Theorem 2.2(iii). □

Hereafter we provide the proofs omitted in Sections 3 and 4.

Proof of Proposition 3.7. Let $T(x) = x-\gamma\nabla f(x)$. Then P_γ can be expressed as the composition of mappings $\operatorname{prox}_{\gamma g}$ and T, *i.e.*, $P_\gamma(x) = \operatorname{prox}_{\gamma g}(T(x))$. Since $\operatorname{prox}_{\gamma g}$ is (strongly) semismooth at $T(x_\star)$ we have that $\partial_C \operatorname{prox}_{\gamma g}$ is a (strong) LNA for $\operatorname{prox}_{\gamma g}$ at $T(x_\star)$. On the other hand, since T is twice continuously differentiable, its Jacobian $\nabla T(x) = I-\gamma\nabla^2 f(x)$ is a LNA of T at $x_\star$. If in addition $\nabla^2 f$ is Lipschitz continuous around $x_\star$ then ∇T is a strong LNA of T at $x_\star$ [13, Prop. 7.2.9]. Invoking [13, Th. 7.5.17] we have that

$$\mathscr{P}_\gamma(x) = \{P(I-\gamma\nabla^2 f(x)) \mid P\in\partial_C(\operatorname{prox}_{\gamma g})(x-\gamma\nabla f(x))\},$$

is a (strong) LNA of P_γ at $x_\star$.

Next consider $G_\gamma(x) = \gamma^{-1}(x-P_\gamma(x))$. Applying [13, Cor. 7.5.18(a)(b)] we have

$$\mathscr{G}_\gamma(x) = \{\gamma^{-1}(I-V) \mid V\in\mathscr{P}_\gamma(x)\},$$

is a (strong) LNA for G_γ at $x_\star$. Reinterpreting $\hat\partial^2 F_\gamma(x)$ with the current notation,

$$\hat\partial^2 F_\gamma(x) = \{(I-\gamma\nabla^2 f(x))Z \mid Z\in\mathscr{G}_\gamma(x)\}.$$

Therefore, for any $H\in\hat\partial^2 F_\gamma(x)$

$$\begin{aligned}\|\nabla F_\gamma(x)+H(x_\star-x)-\nabla F_\gamma(x_\star)\| &= \|(I-\gamma\nabla^2 f(x))(G_\gamma(x)+Z(x-x_\star)-G_\gamma(x_\star))\|\\ &\le \|G_\gamma(x)+Z(x-x_\star)-G_\gamma(x_\star)\|,\end{aligned}$$

where the equality follows by $0 = \nabla F_\gamma(x_\star) = (I-\gamma\nabla f^2(x_\star))G_\gamma(x_\star)$, and the inequality by $\gamma\in(0,1/L_f)$. Since $\mathscr{G}_\gamma$ is a (strong) LNA of G_γ, the last term is $o(\|x-x_\star\|)$ (and $O(\|x-x_\star\|^2)$ in the case where $\nabla^2 f$ is locally Lipschitz continuous). This shows that $\hat\partial F_\gamma$ is a (strong) LNA of ∇F_γ at $x_\star$. □

Proof of Proposition 3.8. Any $H\in\hat\partial^2 F_\gamma(x)$ can be expressed as

$$H = \gamma^{-1}(I-\gamma\nabla^2 f(x)) - \gamma^{-1}(I-\gamma\nabla^2 f(x))P(I-\gamma\nabla^2 f(x))$$

for some $P\in\partial_C(\operatorname{prox}_{\gamma g})(x-\gamma\nabla f(x))$. Obviously, recalling Theorem 3.2, H is a symmetric matrix. We have

$$\begin{aligned}d'Hd &= \gamma^{-1}d'(I-\gamma\nabla^2 f(x))d - \gamma^{-1}d'(I-\gamma\nabla^2 f(x))P(I-\gamma\nabla^2 f(x))d\\ &\ge \gamma^{-1}d'(I-\gamma\nabla^2 f(x))d - \gamma^{-1}\|(I-\gamma\nabla^2 f(x))d\|^2\\ &= d'(I-\gamma\nabla^2 f(x))\nabla^2 f(x)d\\ &\ge \min\{(1-\gamma\mu_f)\mu_f,(1-\gamma L_f)L_f\}\|d\|^2,\end{aligned}$$

where the first inequality follows by Theorem 3.2 and the second by Lemma A.1. On the other hand

$$\begin{aligned} d'Hd &= \gamma^{-1}d'(I-\gamma\nabla^2 f(x))d - \gamma^{-1}d'(I-\gamma\nabla^2 f(x))P(I-\gamma\nabla^2 f(x))d \\ &\le \gamma^{-1}d'(I-\gamma\nabla^2 f(x))d \\ &\le \gamma^{-1}(1-\gamma\mu_f)\|d\|^2, \end{aligned}$$

where the first inequality follows by Theorem 3.2. □

Proof of Lemma 3.9. It suffices to prove that $\|x-x_\star\| \le c\|\nabla F_\gamma(x)\|$, for all x with $\|x-x_\star\| \le \delta$ and some positive c, δ. The result will then follow, since $\|\nabla F_\gamma(x)\| = \|(I-\gamma\nabla^2 f(x))G_\gamma(x)\| \le \|G_\gamma(x)\|$, for $\gamma \in (0, 1/L_f)$. For the sake of contradiction assume that there exists a sequence of vectors $\{x^k\}$ converging to $x_\star$ such that $x^k \ne x_\star$ for every k and

$$\lim_{k\to\infty} \frac{\nabla F_\gamma(x^k)}{\|x^k - x_\star\|} = 0. \tag{A.4}$$

The assumptions of the lemma guarantee through Proposition 3.7 that $\hat\partial^2 F_\gamma$ is a LNA of ∇F_γ at $x_\star$, therefore

$$0 = \lim_{k\to\infty} \frac{\nabla F(x^k) + H^k(x_\star - x^k) - \nabla F_\gamma(x^\star)}{\|x^k - x_\star\|} = \lim_{k\to\infty} \frac{H^k(x_\star - x^k)}{\|x^k - x_\star\|},$$

where the second equality follows from (A.4). This implies that

$$\lim_{k\to\infty} \frac{(x_\star - x^k)'H^k(x_\star - x^k)}{\|x^k - x_\star\|^2} = 0.$$

But since $\hat\partial^2 F_\gamma$ is compact-valued and outer semicontinuous at $x_\star$, and $\{x^k\}$ converges to $x_\star$, the nonsingularity assumption on the elements of $\hat\partial^2 F_\gamma(x_\star)$ implies through [13, Lem. 7.5.2] that for sufficiently large k, the smallest eigenvalue of H^k is minorized by a positive number. Therefore the above limit must be positive, reaching to a contradiction. Uniqueness follows from the fact that the set of zeros of ∇F_γ is equal to the set of optimal solutions of (1.1), through Theorem 2.2(i). □

Proof of Theorem 4.3. Since $\mu_f > 0$ and $\zeta = 0$, using Proposition 3.8, Eq. (4.8) gives

$$\nabla F_\gamma(x^k)'d^k \le -c_1\|\nabla F_\gamma(x^k)\|^2. \tag{A.5}$$

where $c_1 = \frac{\gamma}{(1-\gamma\mu_f)}$ while Eq. (4.7) gives

$$\|d^k\| \le c_2\|\nabla F_\gamma(x^k)\| \tag{A.6}$$

where $c_2 = (\eta+1)/\xi_1$, $\xi_1 \triangleq \min\{(1-\gamma\mu_f)\mu_f, (1-\gamma L_f)L_f\}$. Using Eqs. (4.4), (A.5), step $\tau_k = 2^{-i_k}$ satisfies

$$F_\gamma(x^k + \tau_k d^k) - F_\gamma(x^k) \le -\sigma\tau_k c_1\|\nabla F_\gamma(x^k)\|^2.$$

Due to Theorem 2.3, ∇F_γ is Lipschitz continuous, therefore using the descent Lemma [20, Prop. A.24]

$$\begin{aligned} F_\gamma(x^k + 2^{-i}d^k) - F_\gamma(x^k) &\le 2^{-i}\nabla F_\gamma(x^k)'d^k + \tfrac{L_{F_\gamma}}{2}2^{-2i}\|d^k\|^2 \\ &\le -2^{-i}c_1\|\nabla F_\gamma(x^k)\|^2 + \tfrac{L_{F_\gamma}}{2}c_2^2 2^{-2i}\|\nabla F_\gamma(x^k)\|^2 \\ &\le -2^{-i}c_1(1 - \tfrac{L_{F_\gamma}}{2}\tfrac{c_2^2}{c_1}2^{-i})\|\nabla F_\gamma(x^k)\|^2 \end{aligned} \tag{A.7}$$

where the second inequality follows by (A.6). Let $i_{\min}$ be the first index i for which $1 - \frac{L_{F_\gamma}}{2}\frac{c_2^2}{c_1}2^{-i} \geq \sigma$, *i.e.*,

$$1 - \tfrac{L_{F_\gamma}}{2}\tfrac{c_2^2}{c_1}2^{-i} < \sigma, \quad 0 \leq i < i_{\min} \tag{A.8a}$$

$$1 - \tfrac{L_{F_\gamma}}{2}\tfrac{c_2^2}{c_1}2^{-i_{\min}} \geq \sigma \tag{A.8b}$$

From (4.4), (A.7) and (A.8) we conclude that $i_k \leq i_{\min}$, therefore $\tau_k \geq \hat{\tau}_{\min}$, where $\hat{\tau}_{\min} = 2^{-i_{\min}}$, thus we have

$$F_\gamma(x^k + \tau_k d^k) - F_\gamma(x^k) \leq -\sigma\hat{\tau}_{\min}c_1\|\nabla F_\gamma(x^k)\|^2 \tag{A.9}$$

From Eq. (A.8a) we obtain

$$\sigma > 1 - \tfrac{L_{F_\gamma}}{2}\tfrac{c_2^2}{c_1}2^{-(i_{\min}-1)} = 1 - \tfrac{c_2^2}{c_1}L_{F_\gamma}2^{-i_{\min}} = 1 - \tfrac{c_2^2}{c_1}L_{F_\gamma}\hat{\tau}_{\min}$$

Hence

$$\hat{\tau}_{\min} > \frac{1-\sigma}{L_{F_\gamma}}\frac{c_1}{c_2^2}. \tag{A.10}$$

Subtracting $F_\star$ from both sides of (A.9) and using (A.10)

$$F_\gamma(x^k + \tau_k d^k) - F_\star \leq F_\gamma(x^k) - F_\star - \tfrac{\sigma(1-\sigma)}{L_{F_\gamma}}\tfrac{c_1^2}{c_2^2}\|\nabla F_\gamma(x^k)\|^2. \tag{A.11}$$

Since F_γ is strongly convex (cf. Theorem 2.3) we have [19, Th. 2.1.10]

$$F_\gamma(x^k) - F_\star \leq \frac{1}{2\mu_{F_\gamma}}\|\nabla F_\gamma(x^k)\|^2. \tag{A.12}$$

Combining (A.11) and (A.12) we obtain

$$F_\gamma(x^{k+1}) - F_\star \leq r_{F_\gamma}(F_\gamma(x^k) - F_\star)$$

where $r_{F_\gamma} = 1 - \frac{2\mu_{F_\gamma}\sigma(1-\sigma)}{L_{F_\gamma}}\frac{c_1^2}{c_2^2}$, therefore

$$F_\gamma(x^k) - F_\star \leq r_{F_\gamma}^k(F_\gamma(x^0) - F_\star).$$

Using $F(P_\gamma(x^k)) \leq F_\gamma(x^k)$ (cf. Theorem 2.2(iii)) we arrive at (4.18a). Using [19, Th. 2.1.8]

$$(\mu_{F_\gamma}/2)\|x - x_\star\|^2 \leq F_\gamma(x) - F_\star \leq (L_{F_\gamma}/2)\|x - x_\star\|^2$$

we obtain (4.18b). □

Proof of Theorem 4.6. If $k \notin \mathcal{K}$ and $s_k = 0$, then $F(x^{k+1}) = F(P_\gamma(x^k)) \leq F_\gamma(x^k)$, where the inequality follows from (2.12). If $k \in \mathcal{K}$ or $s_k = 1$, then $F(x^{k+1}) = F(P_\gamma(\hat{x}^k)) \leq F_\gamma(\hat{x}^k) \leq F_\gamma(x^k)$, where the first inequality uses (2.12) while the second uses the fact that d^k is a direction of descent for F_γ. Therefore, we have

$$F(x^{k+1}) \leq F_\gamma(x^k), \quad k \in \mathbb{N}. \tag{A.13}$$

Next, for any $x \in \mathbb{R}^n$

$$F_\gamma(x) \leq \min_{z\in\mathbb{R}^n}\left\{f(z) + g(z) + \tfrac{1}{2\gamma}\|z - x\|^2\right\} = F^\gamma(x), \tag{A.14}$$

where the inequality uses the convexity of f (recall that F^γ is the Moreau envelope of $F = f + g$). Combining (A.13) with (A.14), we obtain $F(x^{k+1}) \leq F^\gamma(x^k)$. The rest of the proof is similar to [6, Th. 4]. In particular we have

$$\begin{aligned} F(x^{k+1}) \leq F^\gamma(x^k) &= \min_{x\in\mathbb{R}^n}\left\{F(x) + \tfrac{1}{2\gamma}\|x - x^k\|^2\right\} \\ &\leq \min_{0\leq\alpha\leq 1}\left\{F(\alpha x_\star + (1-\alpha)x^k) + \tfrac{\alpha^2}{2\gamma}\|x^k - x_\star\|^2\right\} \\ &\leq \min_{0\leq\alpha\leq 1}\left\{F(x^k) - \alpha(F(x^k) - F_\star) + \tfrac{R^2}{2\gamma}\alpha^2\right\}, \end{aligned}$$

where the last inequality follows by convexity of F. If $F(x^0) - F_\star \geq R^2/\gamma$, then the optimal solution of the latter problem for $k = 0$ is $\alpha = 1$ and we obtain (4.20). Otherwise, the optimal solution is $\alpha = \frac{\gamma(F(x^k)-F_\star)}{R^2} \leq \frac{\gamma(F(x^0)-F_\star)}{R^2} \leq 1$ and we obtain

$$F(x^{k+1}) \leq F(x^k) - \frac{\gamma(F(x^k) - F_\star)^2}{2R^2}.$$

Letting $\lambda_k = \frac{1}{F(x^k)-F_\star}$ the latter inequality is expressed as

$$\frac{1}{\lambda_{k+1}} \leq \frac{1}{\lambda_k} - \frac{\gamma}{2R^2\lambda_k^2}.$$

Multiplying both sides by $\lambda_k\lambda_{k+1}$ and rearranging

$$\lambda_{k+1} \geq \lambda_k + \frac{\gamma}{2R^2}\frac{\lambda_{k+1}}{\lambda_k} \geq \lambda_k + \frac{\gamma}{2R^2}$$

where the latter inequality follows from the fact that $\{F(x^k)\}_{k\in\mathbb{N}}$ is nonincreasing (cf. (4.19)). Summing up for $0, \ldots, k-1$ we obtain

$$\lambda_k \geq \lambda_0 + \frac{\gamma}{2R^2}k \geq \frac{\gamma}{2R^2}(k+2)$$

where the last inequality follows by $F(x^0) - F_\star \leq R^2/\gamma$. Rearranging, we arrive at (4.21). □

Proof of Theorem 4.7. If $k \notin \mathcal{K}$ and $s_k = 0$, then $x^{k+1} = P_\gamma(x^k)$ and the decrease condition (4.19) holds. Subtracting $F_\star$ from both sides and using Corollary A.4 we obtain

$$F(x^k) - F_\star \geq (1 + \gamma\mu_f)(F(x^{k+1}) - F_\star). \tag{A.15}$$

If $k \in \mathcal{K}$ or $s_k = 1$, we have $F(x^{k+1}) = F(P_\gamma(\hat{x}^k)) \leq F_\gamma(\hat{x}^k) - \frac{\gamma}{2}\|G_\gamma(\hat{x}^k)\|^2 \leq F_\gamma(x^k) - \frac{\gamma}{2}\|G_\gamma(\hat{x}^k)\|^2 \leq F(x^k) - \frac{\gamma}{2}\|G_\gamma(\hat{x}^k)\|^2$, where the first inequality follows from Theorem 2.2(iii), the second from (4.4) and the descent property of d^k and the third one from Theorem 2.2(ii). Subtacting $F_\star$ from both sides

$$F(x^{k+1}) - F_\star + \tfrac{\gamma}{2}\|G_\gamma(\hat{x}^k)\|^2 \leq F(x^k) - F^\star.$$

Using Corollary A.4, we obtain $\|G_\gamma(\hat{x}^k)\|^2 \geq 2\mu_f(F(P_\gamma(\hat{x}^k) - F_\star) = 2\mu_f(F(x^{k+1}) - F_\star)$. Combining the last two inequalities we again obtain (A.15), which proves (4.22a). Now, from (A.15) we obtain

$$\begin{aligned} F(x^{k+1}) - F_\star &\leq (1+\gamma\mu_f)^{-k}(F(x^1) - F_\star) \\ &= (1+\gamma\mu_f)^{-k}(F(P_\gamma(x^0)) - F_\star) \\ &\leq \frac{1-\gamma\mu_f}{2\gamma(1+\gamma\mu_f)^k}\|x^0 - x_\star\|^2, \end{aligned} \tag{A.16}$$

where the equality comes from the fact that $s_0 = 0$ and the second inequality follows from Proposition A.5. Finally, putting $x = x^{k+1}$, $\bar{x} = x_\star \in X_\star$ in (A.3) and minimizing both sides we obtain

$$F(x^{k+1}) - F_\star \geq \tfrac{\mu_f}{2}\|x^{k+1} - x_\star\|^2. \tag{A.17}$$

Combining (A.16) and (A.17) we arrive at (4.22b). □

(P. Patrinos) IMT Institute for Advanced Studies Lucca
E-mail address: panagiotis.patrinos@imtlucca.it

(L. Stella) IMT Institute for Advanced Studies Lucca
E-mail address: lorenzo.stella@imtlucca.it

(A. Bemporad) IMT Institute for Advanced Studies Lucca
E-mail address: alberto.bemporad@imtlucca.it