52nd IEEE Conference on Decision and Control
December 10-13, 2013. Florence, Italy

Proximal Newton Methods for Convex Composite Optimization

Panagiotis Patrinos and Alberto Bemporad

Abstract— **This paper proposes two proximal Newton methods for convex nonsmooth optimization problems in composite form. The algorithms are based on a new continuously differentiable exact penalty function, namely the *Composite Moreau Envelope*. 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 solution of a linear system of usually small dimension.**

I. Introduction

The focus of this work is on efficient Newton-like algorithms for convex optimization problems in composite form, i.e., the goal is to minimize an extended-real-valued function $F = f + g$, where f is convex twice continuously differentiable and g is convex and it can be nonsmooth and extended-real-valued. Problems of this form have found many applications ranging from optimization-based control such as Model Predictive Control (MPC), to machine learning, signal processing, linear inverse problems, and image analysis. For example when f is quadratic and g is the indicator of a polyhedral set then the problem of minimizing F reduces a quadratic program (QP), which has found numerous applications in embedded MPC, whereas if $g = \|\cdot\|_1$ then the problem becomes an ℓ_1-regularized optimization problem which has found many applications in sparse approximation techniques.

Perhaps the most well known algorithm for convex composite optimization is the forward-backward or proximal gradient algorithm [1], a generalization of the classical gradient and gradient projection methods to problems involving a nonsmooth term. Accelerated versions of the proximal gradient algorithm based on the work of Nesterov [2]–[4] have also gained popularity. All the aforementioned algorithms are based on computing at every iteration a solution of the following linearized version of the problem around the current iterate x

$$\min_u \{f(x) + \nabla f(x)'(u-x) + g(u) + \tfrac{1}{2\gamma}\|u-x\|^2\}. \quad (1)$$

Although these algorithms share favorable global convergence rate estimates of order $O(\epsilon^{-1})$ or $O(\epsilon^{-1/2})$ (ϵ is solution accuracy), they are first-order methods and as such, they are usually effective on computing solutions of low or medium accuracy only. An evident remedy is to include second-oder information by replacing the term $\frac{1}{2\gamma}\|u-x\|^2$ with $\frac{1}{2}(u-x)'Q(u-x)$, 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 [5], [6]. However, a severe limitation of the approach is that, unless Q has 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 route by showing that the value function of problem (1) (viewed as function of x), which we call the *Composite Moreau Envelope (CME)*, 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 [7] and as an unconstrained continuously differentiable optimization problem via the D-gap function [8], see [9, Ch. 10] for a survey and [10], [11] for applications to constrained optimization and MPC.

Next, we show that one can design Newton-like methods to minimize the CME by using tools from nonsmooth analysis. Unlike the approaches of [5], [6], 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 asymptotic convergence rates in the nondegenerate cases. We show that one can achieve this goal by embedding Newton-like iterations on the penalty function, directly into the proximal gradient method. This is possible by relating directions of descent for the penalty function with those for the original nonsmooth function.

The methods proposed in this paper are also particularly suitable for real-time optimization applications, such as embedded MPC, where the basic requirements are not only speed but also software simplicity and complexity certification. A step in this direction was taken in [12]–[14], where tight bounds on the number of iterations were obtained using fast gradient methods.

The proofs of the results are omitted due to space limitations, but are available upon request.

II. Convex Composite Optimization

Consider the following nonsmooth optimization problem

$$F_\star \triangleq \inf_{x\in\mathbb{R}^n} F(x) \triangleq f(x) + g(x), \quad (2)$$

The authors are with IMT Institute for Advanced Studies Lucca, Piazza San Ponziano 6, 55100 Lucca, Italy. {panagiotis.patrinos, alberto.bemporad}@imtlucca.it.

where $f \in \mathcal{S}^{2,1}_{\mu_f,L_f}(\mathbb{R}^n)$ and $g \in \mathcal{S}^0(\mathbb{R}^n)$ [1]. We assume that the set of minimizers $X_\star = \operatorname{argmin} F$ is nonempty. It is well known [15] that $x_\star \in X_\star$ if and only if

$$x_\star = \operatorname{prox}_{\gamma g}(x_\star - \gamma \nabla f(x_\star)),$$

where $\gamma > 0$ and $\operatorname{prox}_{\gamma g}$ is the *proximal mapping* of g [16] defined by

$$\operatorname{prox}_{\gamma g}(x) \triangleq \underset{u}{\operatorname{argmin}} \left\{ g(u) + \tfrac{1}{2\gamma}\|u - x\|^2 \right\}. \tag{3}$$

The value function $g^\gamma : \mathbb{R}^n \to \mathbb{R}$ of the optimization problem (3) is called the *Moreau envelope*, i.e.,

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

Properties of the Moreau envelope and the proximal mapping are well documented in the literature [1], [15], [17], [18]. For example, one has $g^\gamma \le g$, $\operatorname{argmin}_x g^\gamma(x) = \operatorname{argmin}_x g(x)$, $\inf_x g^\gamma(x) = \inf_x g(x)$. Furthermore, the proximal mapping is single-valued, continuous and nonexpansive (Lipschitz continuous with Lipschitz constant 1) and the envelope function g^γ is convex, continuously differentiable, with γ^{-1}-Lipschitz continuous gradient given by $\nabla g^\gamma(x) = \gamma^{-1}(x - \operatorname{prox}_{\gamma g}(x))$. In many cases the Moreau envelope and the proximal mapping can be computed explicitly [18], [1]. However, computing $(f + g)^\gamma$ is usually as hard as solving Problem (2).

III. Composite Moreau Envelope

Next, we introduce the Composite Moreau Envelope which is a continuously differentiable penalty function for (2). For $\gamma > 0$, consider $F_\gamma : \mathbb{R}^n \to \mathbb{R}$ defined by

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

An alternative way to express F_γ is

$$F_\gamma(x) = \min_{u \in \mathbb{R}^n} \left\{ f(x) + \nabla f(x)'(u - x) + g(u) + \tfrac{1}{2\gamma}\|u - x\|^2 \right\}.$$

Let

$$y_\gamma(x) \triangleq \operatorname{prox}_{\gamma g}(x - \gamma \nabla f(x)),$$
$$z_\gamma(x) \triangleq \gamma^{-1}(x - y_\gamma(x)).$$

We note immediately that $x_\star \in X_\star \iff z_\gamma(x_\star) = 0$, $\gamma > 0$. 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 first-order differentiability properties.

Proposition 1: F_γ is continuously differentiable with

$$\nabla F_\gamma(x) = \left(I - \gamma \nabla^2 f(x)\right) z_\gamma(x).$$

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

Another important property of F_γ is that it minorizes F.

Proposition 2: For any $x \in \mathbb{R}^n$, $\gamma > 0$

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

On the other hand, at any $x \in \mathbb{R}^n$, F_γ can be lower bounded by F evaluated at $y_\gamma(x)$.

Proposition 3: For any $x \in \mathbb{R}^n$, $\gamma > 0$

$$F(y_\gamma(x)) \le F_\gamma(x) - \tfrac{\gamma}{2}\left(1 - \gamma L_f\right)\|z_\gamma(x)\|^2.$$

In particular, if $\gamma \in (0, L_f^{-1}]$ then $F(y_\gamma(x)) \le F_\gamma(x)$.

The two preceding propositions lead us to the following corollary. It states that if $\gamma \in (0, L_f^{-1})$ then not only do the stationary points of F_γ agree with $X_\star$ (cf. Prop. 1) but also that its optimal set agrees with $X_\star$. Although F_γ may not be convex, the set of stationary points turns out to be equal to the set of its minimizers.

Corollary 4: If $\gamma \in (0, L_f^{-1})$ then $X_\star = \operatorname{argmin} F_\gamma$.

IV. Second-order Analysis of F_γ

As it was shown in Section III, F_γ is real-valued and continuously differentiable. However, second order differentiability of g^γ is impossible unless g is twice continuously differentiable [19]. In this section, we construct a *linear Newton approximation* of ∇F_γ which can be considered as a generalized Hessian for F_γ and will allow the development of Newton-like methods with fast asymptotic convergence rates. For completeness, we describe the tools of nonsmooth analysis needed to proceed with the construction.

Definition 1 ([9, Def. 7.5.13]): Let $G : \mathbb{R}^n \to \mathbb{R}^n$ be locally Lipschitz 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 multifunction $\mathcal{T} : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n \times n}$ that has nonempty compact images, is upper semicontinuous at $\bar{x}$ and for any $H \in \mathcal{T}(x)$

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

Instead, if

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

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

Arguably the most notable example of a linear Newton approximation is *Clarke's generalized Jacobian*, $\partial_C G(x) \triangleq \operatorname{conv}(\partial_B G(x))$, where the B-subdifferential is

$$\partial_B G(x) \triangleq \left\{ H \in \mathbb{R}^{m \times n} \;\middle|\; \begin{array}{c} \exists \{x_k\} \subset C_G \text{ with} \\ x_k \to x, \nabla G(x_k) \to H \end{array} \right\}$$

and C_G is the subset of $\mathbb{R}^n$ consisting of the points where G is differentiable (if G is locally Lipschitz continuous, according to Rademarcher's theorem $\mathbb{R}^n \setminus C_G$ is of zero measure). In particular, if G is (strongly) *semismooth* at $\bar{x}$, i.e., G is directionally differentiable at $\bar{x}$ and $\|Hd - G'(x; d)\| = o(\|d\|)$ for all $d \to 0$ and all $H \in \partial G(x + d)$, then $\partial_C G$, $\partial_B G$ are linear Newton approximations of G at $\bar{x}$. If G is strongly semismooth at $\bar{x}$, i.e., G is directionally differentiable at $\bar{x}$ and $\|Hd - G'(x; d)\| = O(\|d\|^2)$ for all $d \to 0$ and all $H \in \partial G(x + d)$, then $\partial_C G$, $\partial_B G$ are strong linear Newton approximations of G at $\bar{x}$.

However, semismooth mappings can have linear Newton approximations other than the generalized Jacobian. More

[1] $\mathcal{S}^{2,1}_{\mu,L}(\mathbb{R}^n)$: class of twice continuously differentiable, strongly convex functions with convexity parameter $\mu \ge 0$, whose gradient is Lipschitz continuous with constant $L \ge 0$. $\mathcal{S}^0(\mathbb{R}^n)$: class of proper, lower semicontinuous convex functions from $\mathbb{R}^n$ to $\overline{\mathbb{R}}$.

importantly, mappings that are not semismooth can admit linear Newton approximations as well.

A. Generalized Jacobian of proximal mappings

The next theorem gives the basic properties of the generalized Jacobian of the proximal mapping.

Theorem 5: 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 $\|P\| \leq 1$.

Remark 1: In many cases where $\operatorname{prox}_{\gamma g}$ is explicitly computable, it can be shown that it is (strongly) semismooth, hence it admits $\partial_C \operatorname{prox}_{\gamma g}$, $\partial_B \operatorname{prox}_{\gamma g}$ as linear Newton approximations. For example, when $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is piecewise quadratic (e.g., indicator or support function of polyhedral sets, ℓ_1 and ℓ_∞ norms, etc.) then $\operatorname{prox}_{\gamma g}$ is piecewise affine, hence strongly semismooth. The projection operator over symmetric cones is strongly semismooth [20]. General conditions that guarantee semismoothness of the proximal mapping can be found in [21], [22].

Remark 2 (block separable cost): An interesting property of $\partial_C \operatorname{prox}_{\gamma g}$, that follows directly from its definition, is that if g is (block) separable, then every $P \in \partial \operatorname{prox}_{\gamma g}(x)$ is a (block-) diagonal matrix. This has favorable computational implications especially for large-scale problems. For example, if g is separable (ℓ_1-norm, indicator of ℓ_∞-ball or of a box), i.e., $g(x) = \sum_{i=1}^n g_i(x_i)$ 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 $\{0, 1\}$).

B. Linear Newton approximation of ∇F_γ

In order to be able to devise Newton-like algorithms with fast asymptotic convergence rates for minimizing F_γ, we need to construct a linear Newton approximation for ∇F_γ. One way to do that is to impose extra regularity assumptions on f so that ∇F_γ is semismooth. In that case one can employ $\partial_C(\nabla F_\gamma)$ as a linear Newton approximation. However, the computation of an element of $\partial_C(\nabla F_\gamma)$ can become too complicated, since it will involve third-order 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 at some stationary point of F_γ, which by Corollary 4 is also a minimizer of F, provided that $\gamma \in (0, L_f^{-1})$. It turns out that we can define a linear Newton approximation at a stationary point, whose elements have a simpler form than those of $\partial_C(\nabla F_\gamma)$, without assuming semismoothness of ∇F_γ. The approach we follow is largely based on [23], [9, Prop. 10.4.4]. First, we will need the following lemma.

Lemma 6: Suppose that $\mathcal{P}_\gamma : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n\times n}$ is a linear Newton approximation for $\operatorname{prox}_{\gamma g}$ at $x - \gamma\nabla f(x) \in \mathbb{R}^n$, $x \in \mathbb{R}^n$. Then

$$\mathcal{Z}_\gamma(x)=\{\gamma^{-1}(I-P(I-\gamma\nabla^2 f(x)))|P \in \mathcal{P}_\gamma(x - \gamma\nabla f(x))\}$$

is a linear Newton approximation for z_γ at x. Furthermore, if $\mathcal{P}_\gamma$ is a strong linear Newton approximation for $\operatorname{prox}_{\gamma g}$ at $x - \gamma\nabla f(x)$ and $\nabla^2 f$ is Lipschitz continuous at x, then $\mathcal{Z}_\gamma$ is a strong linear Newton approximation for z_γ at x.

Theorem 7: Suppose that $\mathcal{P}_\gamma : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n\times n}$ is a linear Newton approximation for $\operatorname{prox}_{\gamma g}$ at $x_\star - \gamma\nabla f(x_\star) \in \mathbb{R}^n$, $x_\star \in X_\star$. Then

$$\mathcal{T}_\gamma(x)=\{(I-\gamma\nabla^2 f(x))Z|Z \in \mathcal{Z}_\gamma(x)\}$$

is a linear Newton approximation for ∇F_γ at $x_\star$. Furthermore, if $\mathcal{P}_\gamma$ is a strong linear Newton approximation for $\operatorname{prox}_{\gamma g}$ at $x_\star - \gamma\nabla f(x_\star)$ and $\nabla^2 f$ is Lipschitz continuous at $x_\star$, then $\mathcal{T}_\gamma$ is a strong linear Newton approximation for ∇F_γ at $x_\star$.

The next proposition shows that every element of $\mathcal{T}_\gamma$ is a symmetric positive semidefinite matrix, therefore the corresponding *Newton system* (cf. Eq. (5)) can be solved (approximately) by (modified) Cholesky factorization or conjugate gradient methods.

Proposition 8: Let $x \in \mathbb{R}^n$ and assume that $\mathcal{P}_\gamma(x - \gamma\nabla f(x)) \subseteq \partial_C \operatorname{prox}_{\gamma g}(x - \gamma\nabla f(x))$. Any $H \in \mathcal{T}_\gamma(x)$ is symmetric positive semidefinite and satisfies

$$c_1\|d\|^2 \leq d'Hd \leq c_2\|d\|^2, \ \forall d \in \mathbb{R}^n, \tag{4}$$

where $c_1 \triangleq \min\{(1-\gamma\mu_f)\mu_f, (1-\gamma L_f)L_f\}$, $c_2 \triangleq \gamma^{-1}(1-\gamma\mu_f)$.

The mapping $\mathcal{T}_\gamma$ can be used to compute Newton-like directions for F_γ. At any $x \in \mathbb{R}^n$, we can pick a matrix $H \in \mathcal{T}_\gamma(x)$ and solve the following linear system, which we call *Newton system*, to determine a Newton direction:

$$Hd = -\nabla F_\gamma(x). \tag{5}$$

Due to the structure of the elements of $\mathcal{T}_\gamma(x)$, this simplifies to

$$(I - P(I - \gamma\nabla^2 f(x)))d = -(x - y_\gamma(x)), \tag{6}$$

where $P \in \mathcal{P}_\gamma(x - \gamma\nabla f(x))$.

V. Proximal Newton Method

Having established the equivalence between minimizing F and the CME F_γ, as well as a linear Newton approximation for ∇F_γ, it is now very easy to design globally convergent Newton-like algorithms with strong asymptotic convergence rates, for computing a $x_\star \in X_\star$. Algorithm 1 is a standard line search method for minimizing F_γ. Under a nondegeneracy assumption on $\mathcal{T}_\gamma$, eventually the stepsize becomes equal to 1 and Algorithm 1 reduces to the (undamped) linear Newton method [9, Alg. 7.5.14] for solving $\nabla F_\gamma(x) = 0$.

The next theorem summarizes the convergence properties of Algorithm 1, as well as its asymptotic convergence rate. The proof of the first two parts is quite standard in unconstrained optimization of continuously differentiable functions [24, Prop. 1.2.1], [9, Th. 10.4.9(a), (c)]. The proof of the third part is similar to [9, Th. 10.4.9(d)].

Theorem 9: Let $\{x^\nu\}$ be an infinite sequence generated by Algorithm 1. Assume that there exist $\delta_2 \geq \delta_1 > 0$ such that $\delta_1\|d\|^2 \leq d'D^\nu d \leq \delta_2\|d\|^2$ for all $d \in \mathbb{R}^n$. (i)

1) Every accumulation point of $\{x^\nu\}$ belongs to $X_\star$.
2) If $\{x^\nu\}$ has an isolated accumulation point, then the whole sequence $\{x^\nu\}$ converges to that point.

Algorithm 1: Proximal Newton Method (PNM)

```
Input: $\gamma \in (0, L_f^{-1})$, $\sigma \in (0, 1/2)$, $\rho > 0$, $p > 2$, $\nu = 0$,
       $x^0 \in \mathbb{R}^n$.
1 Select $H \in \mathcal{T}_\gamma(x^\nu)$ and solve
      $H^\nu d = -\nabla F_\gamma(x^\nu).$
  If system is not solvable or if
      $\nabla F_\gamma(x^\nu)' d^\nu \leq -\rho \|d^\nu\|^p$
  is not satisfied, $d^\nu \leftarrow -D^\nu \nabla F_\gamma(x^\nu)$, $D^\nu$ positive
  definite
2 Find smallest $i_\nu \in \mathbb{N}$ such that $\tau_\nu = 2^{-i_\nu}$ satisfies
  $F_\gamma(x^\nu + \tau_\nu d^\nu) \leq F_\gamma(x^\nu) + \sigma\tau_\nu \nabla F_\gamma(x^\nu)' d^\nu$
3 $x^{\nu+1} \leftarrow x^\nu + \tau_\nu d^\nu$
4 $\nu \leftarrow \nu + 1$ and go to Step 1.
```

Algorithm 2: Proximal Gradient–Newton Method (PGNM)

```
Input: $\gamma \in (0, L_f^{-1})$, $\sigma \in (0, 1/2)$, $\rho > 0$, $p > 2$,
       $\Upsilon \subseteq \mathbb{N}$, $\nu = 0$, $s_0 = 0$, $x^0 \in \operatorname{dom} g$
1 if $\nu \in \Upsilon$ or $s_\nu = 1$ then
2     Select $H^\nu \in \mathcal{T}_\gamma(x^\nu)$ and solve
          $H^\nu d = -\nabla F_\gamma(x^\nu).$
      If system is not solvable or if
          $\nabla F_\gamma(x^\nu)' d^\nu \leq -\rho \|d^\nu\|^p$
      is not satisfied, $d^\nu \leftarrow -D^\nu \nabla F_\gamma(x^\nu)$, $D^\nu$ positive
      definite
3     Find smallest $i_\nu \in \mathbb{N}$ such that $\tau_\nu = 2^{-i_\nu}$ satisfies
      $F_\gamma(x^\nu + \tau_\nu d^\nu) \leq F_\gamma(x^\nu) + \sigma\tau_\nu \nabla F_\gamma(x^\nu)' d^\nu$
4     $x^{\nu+1} \leftarrow y_\gamma(x^\nu + \tau_\nu d^\nu)$
5     if $i_\nu = 0$ then $s_{\nu+1} \leftarrow 1$ else $s_{\nu+1} \leftarrow 0$
6 else
7     $x^{\nu+1} \leftarrow y_\gamma(x^\nu)$, $s_{\nu+1} \leftarrow 0$
8 end
9 $\nu \leftarrow \nu + 1$ and go to Step 2.
```

3) Suppose that $x_\star$ is a limit point of $\{x^\nu\}$. If $\mathcal{P}_\gamma$ is a linear Newton approximation of $\operatorname{prox}_{\gamma g}$ at $x_\star - \gamma\nabla f(x_\star)$ and all elements of $\mathcal{T}_\gamma(x_\star)$ are nonsingular, then the whole sequence converges to $x_\star$ and the convergence rate is Q-superlinear; furthermore if $\nabla^2 f$ is Lipschitz continuous in a neighborhood of $x_\star$, the convergence rate is Q-quadratic.

Remark 3: An obvious choice for D^ν in Step 1 is the identity matrix, which gives $d^\nu = -\nabla F_\gamma(x^\nu)$. Another interesting choice is $D^\nu = \gamma(I - \gamma\nabla^2 f(x^\nu))^{-1}$, which gives $d^\nu = -\gamma z_\gamma(x)$ and $x^{\nu+1} = x^\nu - \tau_\nu(x^\nu - y_\gamma(x^\nu))$. In that case, we can select $\tau_\nu = 1$ to obtain $x^{\nu+1} = y_\gamma(x^\nu)$, i.e., the proximal gradient step. It can be seen that this is a direction of descent for F_γ. In fact, using Props. 2 and 3 we obtain $F_\gamma(x^{\nu+1}) \leq F(x^{\nu+1}) \leq F_\gamma(x^\nu) - \frac{\gamma}{2}(1 - \gamma L_f)\|z_\gamma(x^\nu)\|^2$. It can be easily shown that the conclusions of Theorem 9 are valid also with this choices of D^ν, τ_ν.

VI. Proximal Gradient-Newton Method

Algorithm 1 exhibits fast asymptotic convergence rates provided that the elements of $\mathcal{T}_\gamma(x_\star)$ are nonsingular, but not much can be said about its global convergence rate. This is mainly due to 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 Lipschitz continuous. Another reason for this is that the iterates x^ν produced by Algorithm 1 may be outside $\operatorname{dom} g$ (but $y_\gamma(x^\nu) \in \operatorname{dom} g$, see Prop. 3). 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 proximal gradient step after the Newton step (cf. Alg. 2) and analyze the consequences of this choice on F, directly.

Remark 4: If $\Upsilon = \emptyset$ in Algorithm 2, then it becomes the proximal gradient method [2], [3], [15], i.e., $x^{\nu+1} = \operatorname{prox}_{\gamma g}(x^\nu - \gamma\nabla f(x^\nu))$.

It can be shown that the sequence of iterates $\{x^\nu\}$ produced by Algorithm 2 enjoy the same favorable properties in terms of convergence and local convergence rates, as the one of Algorithm 1.

Theorem 10: Theorem 9 holds for the sequence of iterates produced by Algorithm 2.

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 11: Let $\{x^\nu\}$ 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 $F(x^0) - F(x_\star) \geq \gamma^{-1}R^2$ then $F(x^1) - F(x_\star) \leq \frac{1}{2}\gamma^{-1}R^2$. Otherwise, for any $\nu \in \mathbb{N}$ we have

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

When $f \in \mathcal{S}^{1,1}_{\mu_f,L_f}(\mathbb{R}^n)$, $\mu_f > 0$ the global rate of convergence is linear. The next theorem gives the corresponding estimates.

Theorem 12: If $f \in \mathcal{S}^{1,1}_{\mu_f,L_f}(\mathbb{R}^n)$, $\mu_f > 0$, then

$$F(x^\nu) - F_\star \leq (1 + \gamma\mu_f)^{-\nu}(F(x^0) - F_\star),$$
$$\|x^{\nu+1} - x_\star\|^2 \leq \tfrac{1-\gamma\mu_f}{\gamma\mu_f}(1 + \gamma\mu_f)^{-\nu}\|x^0 - x_\star\|^2.$$

VII. Simulations

A. Box constrained QPs

We compare the performance of Algorithms 1 and 2 aginst the fast gradient method (FGM) [25, Eq. 2.2.19] and the interior point solver GUROBI 5.0 [26], on box-constrained strongly convex QPs, i.e.,

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x'Qx + q'x \\ \text{subject to} & x \in C, \end{array}$$

where $C = \{x \in \mathbb{R}^n | \ell \leq x \leq v\}$. With $g(x) = \delta_C(x)$, we have that $\operatorname{prox}_{\gamma g}(x) = \max\{\min\{x, \ell\}, v\}$ is separable, therefore any $P \in \partial_C g(x)$ is a diagonal matrix with elements in $[0,1]$, cf. Remark 2, simplifying considerably the Newton system (6). Specifically, let

$$\delta \triangleq \{i \in \mathbb{N}_{[1,n]} | \ell_i < x_i - \gamma \nabla_i f(x) < v_i\},$$
$$\underline{\beta} \triangleq \{i \in \mathbb{N}_{[1,n]} | x_i - \gamma \nabla_i f(x) \leq \ell_i\},$$
$$\overline{\beta} \triangleq \{i \in \mathbb{N}_{[1,n]} | x_i - \gamma \nabla_i f(x) \geq v_i\},$$

and $\beta = \underline{\beta} \cup \overline{\beta}$. Then $P = \mathrm{diag}(p_1, \ldots, p_n)$ where $p_i = 1$, for $i \in \delta$ and $p_i = 0$ otherwise, belongs to $\mathcal{P}_\gamma(x - \gamma \nabla f(x))$. Therefore, the Newton system (6) simplifies to $d_{\underline{\beta}} = \ell_{\underline{\beta}} - x_{\underline{\beta}}$, $d_{\overline{\beta}} = v_{\overline{\beta}} - x_{\overline{\beta}}$ and

$$Q_{\delta\delta} d_\delta = -\nabla_\delta f(x) - Q_{\delta\beta} d_\beta. \tag{7}$$

Notice that we need to solve a linear system whose dimension is only $|\delta| \times |\delta|$. This is particularly favorable for constrained optimal control problems as those arising in MPC applications, since the set of active constraints, δ, is usually very small. Since in the particular example we deal with medium scale problems, at every Newton iteration of Algorithms 1 and 2 we solve (7) by computing the Cholesky factor of $Q_{\delta\delta}$. A more efficient implementation would involve modifying the Cholesky factor at every iteration, by doing a series of rank-one updates for the indices of constraints that enter or leave the active set δ.

Tests were performed on random QP problems, for increasing values of n. The condition number of the Hessians is up to 10^4. Algorithms where terminated when $\|z_\gamma(x^\nu)\|^2/(2\mu_f) \leq 10^{-4}$ which guarantees $F(y_\gamma(x^\nu)) - F_\star \leq 10^{-4}$. For Algorithm 2, the set of iterations that a Newton direction is taken, was chosen as $\Upsilon = \{5, 10, \ldots\}$. Figure 1 shows the running time (averaged on 50 QP's for each n) for each solver. It can be observed that in general, using Newton directions reduces dramatically the number of iterations and CPU time compared to FGM. Evidently, Algorithms PNM and PGNM can reach a very high accuracy (which is almost unreachable by proximal gradient methods) quite fast.

B. ℓ_1-regularized least squares

We test Algorithms 1 and 2 on the following ℓ_1-regularized least squares problem:

$$\text{minimize} \quad \tfrac{1}{2}\|Ax - b\|^2 + \lambda \|x\|_1, \tag{8}$$

where $A \in \mathbb{R}^{m \times n}$, $\lambda > 0$. Problem (8) aims at finding sparse solutions of underdetermined linear systems ($m \ll n$) and has found numerous applications in signal reconstruction techniques, compressive sensing and inverse problems. Our algorithms are compared against FISTA [3], l1_ls [27] and SpaRSA [28]. The proximal mapping of $g(x) = \lambda \|x\|_1$ is

$$\mathrm{prox}_{\gamma g}(x) = \left(\mathrm{sign}(x_i) \max\{|x_i| - \lambda\gamma, 0\}\right)_{i \in \mathbb{N}_{[1,n]}},$$

i.e., the soft-thresholding operator. Since g is separable, according to Remark 2, every element of $P \in \mathcal{P}_\gamma(x) = \partial_C g(x)$ is a diagonal matrix with elements in $[0,1]$, simplifying considerably the Newton system (6). Specifically,

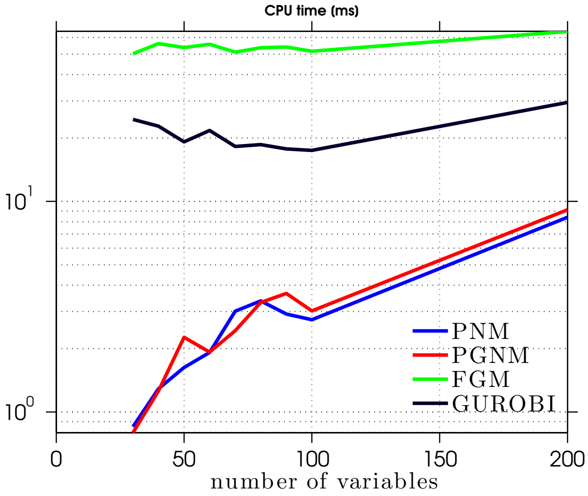

Fig. 1. Comparison on box-constrained strongly convex QPs

let $\delta = \{i \mid |x_i - \gamma \nabla_i f(x)| > \lambda\gamma\}$ and $\beta = \mathbb{N}_{[1,n]} \setminus \delta$. Then $P = \mathrm{diag}(p_1, \ldots, p_n)$ where $p_i = 1$, for $i \in \delta$ and $p_i = 0$ otherwise, belongs to $\mathcal{P}_\gamma(x - \gamma \nabla f(x))$. Therefore, the Newton system (6) simplifies to $d_\beta = -x_\beta$ and

$$A'_{\cdot\delta} A_{\cdot\delta} d_\delta = c \tag{9}$$

where $c = -A'_{\cdot\delta}(A_{\cdot\delta} x_\delta - b) - \zeta_\delta \lambda$, $\zeta = \mathrm{sign}(x - \gamma \nabla f(x))$. Notice that in the vicinity of a solution, δ is usually much smaller than β (after all the goal is to obtain a sparse solution), therefore the dimension of (9) is usually small compared to the number of variables n. In our algorithms, system (9) is solved using the conjugate gradient (CG) method, allowing us to avoid forming explicitly $A'_{\cdot\delta} A_{\cdot\delta}$. The CG method runs at maximum for 10 iterations or stops when $\|A'_{\cdot\delta} A_{\cdot\delta} d_\delta - c\| \leq \eta_\nu \|c\|$, where $\eta_\nu = \min\{0.5, \sqrt{\|c\|}\}\|c\|$. The stopping criterion on the residual can be shown to guarantee locally, superlinear convergence rate [9, Th. 7.5.5]. We have found that this stopping rule works well in practice, since when far from the solution, solving (9) accurately does not make much difference. However, when close to the solution, less than 10 iterations are usually enough to produce a direction that satisfies the stopping criterion for the residual.

Algorithms 1, 2 and FISTA are stopped as soon as the absolute value of the duality gap, $|\nabla f(x^\nu)' x + \lambda \|x^\nu\|_1|$ and maximum dual constraint violation, $\|\nabla f(x^\nu)\|_\infty - \lambda$ [2, Sec. 6], fall below 10^{-6}. For Algorithms l1_ls [27] and SpaRSA [28] the termination criterion is similar, but not exactly equivalent.

An instance of problem (8) was generated in the same way as in [2, Sec. 6], with $m = 4000$, $n = 1000$, $m^\star = 100$ (the number of nonzero elements of the optimal solution). Results for Algorithm 2 correspond to two different choices for the set of Newton iterations; $\mathcal{Y} = \{1, 2, 3, \ldots\}$ (PGNM–1) and $\mathcal{Y} = \{10, 20, 30, \ldots\}$ (PGNM–10).

The trajectories of the iterates in terms of $F(x^\nu) - F_\star$, where $F_\star$ is the (known) optimal cost, are given in Fig. 2. It can be observed that our algorithms are able to reach high

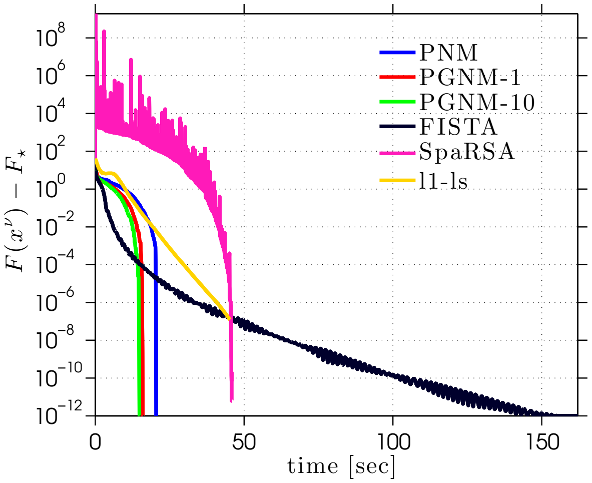

Fig. 2. Cost trajectories of Algorithms 1 (PNM), 2 (PGNM–1, PGNM–10), FISTA, SpaRSA and l1_ls.

accuracy faster. FISTA is able to reach medium accuracy faster but then it has difficulties reaching a high accuracy. The interior point solver l1_ls takes only 34 iterations, however it needs 45.2 seconds in total, since every iteration requires solving a much larger linear system than (9), again using CG. On the other hand PNM, PGNM–1 and PGNM–10 need 20.5, 16 and 14.9 seconds to fulfill the termination criterion with 10^{-6} accuracy. After 20 seconds FISTA has reached only 10^{-2} accuracy with respect to the duality gap and dual constraint violation and it takes in total 162 seconds to fullfil the termination criterion.

VIII. Conclusions & Future Work

The two algorithms presented in this paper can address a wide class of nonsmooth convex optimization problems. The main characteristic of the algorithms is the fast asymptotic convergence rates for the sequence of iterates. In addition, the bounds proved in Theorems 11 and 12 make the method appealing for real-time optimization. Future work includes embedding Newton iterations in accelerated versions of the proximal gradient method to construct algorithms with better global convergence rates.

References

[1] 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.

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

[3] 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.

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

[5] S. Becker and M. J. Fadili, “A quasi-Newton proximal splitting method,” *arXiv preprint arXiv:1206.1156*, 2012.

[6] 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, pp. 836–844. [Online]. Available: http://books.nips.cc/papers/files/nips25/NIPS2012_0388.pdf

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

[8] 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.

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

[10] 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.

[11] 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.

[12] P. Patrinos and A. Bemporad, “An accelerated dual gradient-projection algorithm for linear model predictive control,” *IEEE Transactions on Automatic Control*, In Press 2013.

[13] A. Bemporad and P. Patrinos, “Simple and certifiable quadratic programming algorithms for embedded linear model predictive control,” in *Proc. 4th IFAC Nonlinear Model Predictive Control Conference*, M. Lazar and F. Allgower, Eds., 2012.

[14] V. Nedelcu and I. Necoara, “Iteration complexity of an inexact augmented lagrangian method for constrained mpc,” in *IEEE 51st Annual Conference on Decision and Control*, 2012, pp. 650–655.

[15] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

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

[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] 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.

[20] D. Sun and J. Sun, “Semismooth matrix-valued functions,” *Mathematics of Operations Research*, vol. 27, no. 1, pp. 150–169, 2002.

[21] 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.

[22] 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.

[23] 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, SIAM, Philadelphia, PA*, M. Ferris and J. Pang, Eds. SIAM Publications, 1997, pp. 452–473.

[24] D. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.

[25] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer, 2003, vol. 87.

[26] Gurobi Optimization, Inc., “Gurobi optimizer reference manual,” 2012. [Online]. Available: http://www.gurobi.com

[27] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, “An interior-point method for large-scale ℓ_1-regularized least squares,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2007.

[28] M. Figueiredo, R. Nowak, and S. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, 2007.