Newton-type Alternating Minimization Algorithm for Convex Optimization

Lorenzo Stella, Andreas Themelis and Panagiotis Patrinos

Abstract—We propose NAMA (Newton-type Alternating Minimization Algorithm) for solving structured nonsmooth convex optimization problems where the sum of two functions is to be minimized, one being strongly convex and the other composed with a linear mapping. The proposed algorithm is a line-search method over a continuous, real-valued, exact penalty function for the corresponding dual problem, which is computed by evaluating the augmented Lagrangian at the primal points obtained by alternating minimizations. As a consequence, NAMA relies on exactly the same computations as the classical alternating minimization algorithm (AMA), also known as the dual proximal gradient method. Under standard assumptions the proposed algorithm possesses strong convergence properties, while under mild additional assumptions the asymptotic convergence is superlinear, provided that the search directions are chosen according to quasi-Newton formulas. Due to its simplicity, the proposed method is well suited for embedded applications and large-scale problems. Experiments show that using limited-memory directions in NAMA greatly improves the convergence speed over AMA and its accelerated variant.

I. INTRODUCTION

We consider convex optimization problems of the form

\[
\min_{x \in \mathbb{R}^n} f(x) + g(Ax), \tag{P}
\]

where \( f \) is strongly convex, \( g \) is convex and \( A \) is a linear mapping. Problems of this form are quite general and appear in various areas of applications, including optimal control [1], system identification [2] and machine learning [3], [4]. For example, whenever \( g \) is the indicator function of a convex set \( C \), then (P) models a constrained convex problem: if \( C \) is a box, then in particular (P) amounts to minimizing a strongly convex function subject to polyhedral constraints.

A general approach to the solution of (P) is based on the dual proximal gradient method, or forward-backward splitting, also known as alternating minimization algorithm (AMA) [5]. This is the dual application of an algorithm introduced by Lions and Mercier [6] for finding the zero of the sum of two maximal monotone operators, one of which is assumed to be co-coercive. The alternating minimization algorithm is intimately tied to the framework of augmented Lagrangian methods, and its global convergence and complexity bounds are well covered in the literature, see [5]: a global convergence rate of order \( O(1/\sqrt{k}) \) holds for the primal iterates of AMA under very general assumptions, and can be improved to the optimal rate \( O(1/k) \) using a simple acceleration technique due to Nesterov, see [7]–[9].

As with all first order methods, the performance of (fast) AMA is severely affected by ill-conditioning of the problem [1]. One way to deal with this issue, which is extensively used in classical smooth, unconstrained optimization, is to precondition the problem using (approximate) second-order information on the cost function, as in (quasi-) Newton methods. However, both (P) and its dual are nonsmooth in general. This motivates considering the concept of alternating minimization envelope (AME): this is a real-valued (as opposed to extended real-valued) exact merit function for the dual problem, and is precisely the augmented Lagrangian associated with (P) evaluated at the primal points computed by AMA. Under mild assumptions on (P), the AME is continuously differentiable around the set of dual solutions and even strictly twice differentiable there. As a consequence, the AME allows to extend classical, smooth unconstrained optimization algorithms to the solution of the dual problem to (P), which is nonsmooth in general. In this work we propose a dual line-search method, which uses the AME as merit function to compute the stepsizes. The convergence properties of the proposed algorithm greatly improve over AMA when fast-converging directions, computed by means of quasi-Newton formulas, are followed. Furthermore, we show that the AME is equivalent to the forward-backward envelope (FBE, see [10]–[12]) associated with the dual problem.

A. Related works

The FBE, as a tool for extending smooth unconstrained algorithms to nonsmooth problems, has first been introduced in [10]: there, two semismooth Newton methods are proposed for minimizing the sum of two convex functions, one of which is smooth and the other having an efficiently computable proximal mapping. This is the classical setting in which the proximal gradient method (and its accelerated variant) can be applied. In [11] the convexity assumption on the smooth term is relaxed, and the authors propose a line-search method with global sublinear rate (in the convex case) and asymptotic superlinear rate when quasi-Newton directions are used: the algorithm relies on descent directions over the FBE which is required to be everywhere differentiable. In [13] classical gradient-based line-search methods are considered for minimizing the FBE, see also [14]. In [12] the most general framework, where both summands are allowed to be nonconvex, is taken into account. In this case differentiability of the FBE cannot be assumed: a new algorithm is proposed which com-
computes fast convergent directions with no need for gradient information on the FBE.

A similar approach was used in [15], [16] to accelerate other splitting algorithms, namely the Douglas-Rachford splitting and its dual counterpart ADMM.

### B. Contributions and organization of the paper

In the present paper we deal with the case where $g$ in $(P)$ is composed with a linear mapping. In this case, even though $g$ may possess an efficiently computable proximal mapping, $g \circ A$ in general does not. This motivates addressing the dual problem of $(P)$ instead. The contributions and organization of the present work can be summarized as follows.

- We propose the Newton-type Alternating Minimization Algorithm (NAMA, Section II, Algorithm 1), a generalization of the alternating minimization algorithm that performs a line-search step over the FBE: the proposed algorithm relies on the very same alternating minimization operations of AMA.

- We show that the AME is equivalent to the FBE of the primal problem terms (Section IV).

- We propose the Newton-type Alternating Minimization Algorithm (NAMA, Section II, Algorithm 1), a generalization of

- We show that the AME is equivalent to the FBE of the dual problem (Section III). This observation extends a classical result by Rockafellar, relating the Moreau envelope and the augmented Lagrangian, to our setting where an additional constraint is present.

- We show that the proposed method enjoys global sublinear convergence under standard assumptions, and local linear convergence assuming calmness of the subdifferentials of the problem terms (Section IV).

- We analyze the first- and second-order properties of the AME, by linking them to generalized second-order properties of the primal functions $f$ and $g$ (Section V).

- We show that the proposed method converges asymptotically superlinearly when the dual problem has a (unique) strong dual minimum, and the line-search directions are selected so as to satisfy the Dennis-Moré condition, as it is the case when quasi-Newton update formulas are adopted (Section VI). The effectiveness of our approach is demonstrated by numerical simulations on linear MPC problems (Section VII).

Differently from the approaches in [11], [13], [14], NAMA does not require the gradient of the envelope function, therefore no second-order information on the smooth term is needed; this would severely limit its applicability in the present setting where the dual problem is solved. Furthermore, with respect to the approaches of [13], [14], the algorithm presented here possesses strong global convergence properties which are not typical of classical line-search methods. Differently from [12], despite the fact that the selected directions may not be descent directions and the line search is performed on the envelope function, NAMA is a descent method for the dual objective; this allows to simplify the convergence analysis of the method, and to show the global sublinear convergence rate for the dual cost and the primal iterates.

### C. Notation

In what follows $\langle \cdot, \cdot \rangle$ denotes an inner product over a Euclidean space (whose nature will be clear from the context) and $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$ is the associated norm. For a linear $A : \mathbb{R}^n \to \mathbb{R}^m$, $\| A \|$ is the operator norm induced by the inner products over $\mathbb{R}^n$ and $\mathbb{R}^m$. For a set $C$, we denote by $\text{ri}(C)$ its relative interior, and by $\Pi_C(x) = \text{argmin}_{y \in C} \| y - x \|$ the projection onto $C$ in the considered norm. We denote the extended real line by $\mathbb{R} = \mathbb{R} \cup \{ \infty \}$, and by $\Gamma_0(\mathbb{R}^n)$ the set of proper, closed, convex functions defined over $\mathbb{R}^n$ with values in $\mathbb{R}$. For $h \in \Gamma_0(\mathbb{R}^n)$ its Fenchel conjugate $h^*$, defined as $h^*(y) = \sup_{x \in \mathbb{R}^n} \{ \langle x, y \rangle - h(x) \}$ is also proper, closed and convex. Properties of conjugate functions are well described for example in [17]–[20]. Among these we recall the Fenchel-Young inequality [19, Prop. 13.13]

$$\langle x, y \rangle \leq h(x) + h^*(y) \quad \forall x, y,$$  

with

$$y \in \partial h(x) \iff \langle x, y \rangle = h(x) + h^*(y) \iff x \in \partial h^*(y),$$  

see [17, Thm. 23.5]. For any $\gamma > 0$, the proximal mapping associated with $h$, with stepsize $\gamma$, is denoted as

$$\text{prox}_{\gamma h}(x) = \text{argmin}_z \{ h(z) + (1/2\gamma)\| z - x \|^2 \}. $$

This satisfies the Moreau identity [19, Thm. 14.3(ii)]

$$ h^*(x) = \text{prox}_{\gamma h}(y) + \gamma \text{prox}_{\gamma^{-1} h^*}(\gamma^{-1} y) \quad \forall y. $$

The value function of the problem defining $\text{prox}_{\gamma h}$ is the Moreau envelope

$$ h^*(x) = \min_z \{ h(z) + (1/2\gamma)\| z - x \|^2 \}. $$

An alternative formulation for $(P)$ is

$$ \text{minimize } f(x) + g(z) \quad \text{subject to } Ax = z. \quad (P') $$

Therefore we can define the augmented Lagrangian associated with $(P)$, denoted as

$$ L(x, z, y) = f(x) + g(z) + \langle y, Ax - z \rangle + \frac{\tau}{2} \| Ax - z \|^2, $$

where $\tau \geq 0$. We indicate by $L \equiv L_0$ the ordinary Lagrangian function.

We follow the terminology of [20] when referring to the concepts of strict continuity and strict differentiability. We say that a mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ is strictly continuous at $\bar{x}$ if [20, Def. 9.1(b)]

$$ \lim \sup_{(x,y)\to(\bar{x},\bar{y})} \frac{\| F(y) - F(x) \|}{\| y - x \|} < \infty. $$

If $F$ is (Frechet) differentiable, we let $J_F : \mathbb{R}^n \to \mathbb{R}^{m \times n}$ denote the Jacobian of $F$. When $m = 1$ we indicate with $\nabla F = JF^\top$ the gradient of $F$ and with $\nabla^2 F = J\nabla F^\top$ its Hessian, whenever it makes sense. We say that $F$ is strictly differentiable at $\bar{x}$ if it satisfies the stronger limit [20, Eq. 9(7)]

$$ \lim_{(x,y) \to (\bar{x},\bar{y})} \frac{\| F(y) - F(x) - JF(\bar{x})(y - x) \|}{\| y - x \|} = 0. $$

Some results in the paper are based on generalized second-order properties of extended-real-valued functions.

**Definition I.1** ([20, Def. 13.6]). Function $h : \mathbb{R}^n \to \mathbb{R}$ is said to be twice epi-differentiable at $x$ for $v$, if the second-order difference quotient $\Delta^2_h h(x|v|d) = \frac{h(x + \tau d) - h(x) - \tau \langle v, d \rangle}{\tau^2/2}$
epi-converges as $\tau \searrow 0$ (i.e., its epigraph converges in the sense of Painlevé-Kuratowski, see [20, Def. 7.1]), the limit being the function $d^2 h(x|v)$ given by
\[
d^2 h(x|v)[d] = \lim_{\gamma \to 0} \Delta^2 h(x|v)[d^\gamma].
\]
In this case $d^2 h(x|v)[d]$, as a function of $d$, is said to be the second-order epi-derivative of $h$ at $x$ for $v$. If $\Delta^2 h(x|v)$ epi-converges as $\tau \searrow 0$, $\bar{x} \to x$ and $\bar{v} \to v$, then $h$ is said to be strictly twice epi-differentiable.

Twice epi-differentiability is a mild requirement, and functions with this property are abundant. Refer to [21]–[25] and to [20, §7, §13] for examples and an in-depth account on epi-derivatives, epi-differentiability, and their connections with ordinary differentiability.

II. BACKGROUND AND PROPOSED ALGORITHM

Without further specifying it, throughout the paper we will work under the following basic assumption.

**Assumption 1.** The following hold for (P):
(i) (P) is feasible, i.e., $\text{dom } f \cap \text{dom } g \neq \emptyset$;
(ii) $f \in \mathcal{G}_0(\mathbb{R}^n)$ is strongly convex with modulus $\mu_f > 0$;
(iii) $g \in \mathcal{G}_0(\mathbb{R}^m)$.

**Remark II.1.** Assumption 1 guarantees, by strong convexity of $f$, that a solution to (P) exists and is unique, be it $x^\star$. Assumption 1(ii) also implies that $f^\ast$ is Lipschitz continuously differentiable with constant $\mu_f^{-1}$ [20, Th. 12.60]. Assumption 1(iii) ensures that $g^\ast$ is also proper, closed, convex [19, Cor. 13.33], and its Moreau envelope $(g^\ast)^\gamma$ is strictly convex [20, Ex. 10.32] with $\gamma^{-1}$-Lipschitz gradient
\[
\nabla (g^\ast)^\gamma(y) = \gamma^{-1}(y - \text{prox}_{g^\ast}(y)),
\]
as shown in [19, Prop. 12.29]. □

The Fenchel dual problem associated with (P) is
\[
\text{minimize } \psi(y) = f^\ast(-A^\top y) + g^\ast(y).
\]

Under Assumption 1 strong duality holds, see [26, Thm. 5.2.1(b)-(c)] and primal-dual solutions $(x_\ast, y_\ast)$ to (P)-(D) are characterized by the first-order optimality conditions
\[
\begin{align*}
-A^\top y_\ast &\in \partial f(x_\ast) \quad (\Leftrightarrow \ x_\ast = \nabla f^\ast(-A^\top y_\ast)) \quad (5a) \\
y_\ast &\in \partial g(Ax_\ast) \quad (\Leftrightarrow \ Ax_\ast \in \partial g^\ast(y_\ast)). \quad (5b)
\end{align*}
\]

A natural way to tackle (P) is to solve (D) by means of forward-backward splitting (or proximal gradient method): starting from an initial dual point $y^0 \in \mathbb{R}^m$, iterate
\[
y^{k+1} = T_\gamma(y^k) := \text{prox}_{g^\ast}(y^k + \gamma A^\top f^\ast(-A^\top y^k))
\]
for some positive stepsize parameter $\gamma$. If we define the associated fixed-point residual
\[
R_\gamma(y) := \gamma^{-1}(y - T_\gamma(y)),
\]
then dual optimality can be characterized as follows:
\[
y_\ast \in Y_\ast \Leftrightarrow y_\ast \in \text{fix } T_\gamma \Leftrightarrow y_\ast \in \text{zer } R_\gamma \forall \gamma > 0. \quad (7)
\]

Algorithm 1 Newton-type AMA (NAMA)

**Require:** $y^0 \in \mathbb{R}^m$, $\gamma \in (0, 2\mu_f/\|A\|^2]$, $\beta \in (0, 1)$

**Initialize:** $k = 0$
1: $x^k = \text{argmin}_{x} \{ f(x) + \langle y^k, Ax \rangle \}$
2: $z^k = \text{argmin}_{z} L_{\gamma}(x^k, z, y^k)$
3: Find the largest $\tau_k = \beta k$, $i_k \in \mathbb{N}$, such that
\[
L_{\gamma}(\tilde{x}^k, \tilde{z}^k, \tilde{y}^k) \geq L_{\gamma}(x^k, z^k, y^k),
\]
where
\[
\begin{align*}
\tilde{y}^k &= y^k + \tau_k d^k + \gamma(1 - \tau_k)(Ax^k - z^k) \\
\tilde{x}^k &= \text{argmin}_{x} \{ f(x) + \langle \tilde{y}^k, Ax \rangle \} \\
\tilde{z}^k &= \text{argmin}_{z} L_{\gamma}(\tilde{x}^k, z, \tilde{y}^k)
\end{align*}
\]
4: $y^{k+1} = \tilde{y}^k + \gamma(A\tilde{x}^k - \tilde{z}^k)$, $k = k + 1$, go to step 1

Iterations (6) are easily shown to be equivalent to the following scheme, the alternating minimization algorithm (AMA)
\[
x^k = x(y^k) = \text{argmin}_{x} \{ f(x) + \langle y^k, Ax \rangle \}, \quad (8a)
\]
\[
z^k = z(y^k) = \text{argmin}_{z} L_{\gamma}(x^k, z, y^k), \quad (8b)
\]
\[
y^{k+1} = y^k + \gamma(Ax^k - z^k).
\]

Note that step (8b) can be equivalently formulated as
\[
z^k = \text{prox}_{1-\gamma A^\top f^\ast}((1-\gamma) y^k + Ax^k).
\]

Using the notation of (8), $T_\gamma$ and $R_\gamma$ can be expressed as
\[
T_\gamma(y) = y + \gamma(Ax(y) - z(y))
\]
and
\[
R_\gamma(y) = z(y) - Ax(y).
\]

It can be shown that $x^k \to x_\ast$ in iterations (8), provided that $\gamma \in (0, 2\mu_f/\|A\|^2]$, see [5, Prop. 3]. Moreover, the dual cost in this case converges sublinearly to the optimum with global rate $O(1/k)$, and the extrapolation techniques introduced by Nesterov [8], [27], [28] allow to obtain accelerated versions of AMA with an optimal global rate $O(1/k^2)$, see [9]: we will here refer to this variant as fast AMA.

A. Newton-type alternating minimization algorithm

The convergence speed of (fast) AMA is affected by ill-conditioning of the problem, as it is the case for all first-order methods. To accelerate convergence, we propose Algorithm 1. An overview of the algorithm is as follows:

- **Algorithm 1** is composed by the very same operations as AMA: in fact, only alternating minimization steps with respect to $x$ and $z$ are performed.
- **Step 3** computes a new dual iterate $\tilde{y}^k$, by performing a line-search over the augmented Lagrangian associated with (P) evaluated at the alternating minimization primal points: we will see that this is equivalent to the forward-backward envelope function associated with the dual problem (D).
- The line-search is performed using a convex combination of the “nominal” residual direction $\gamma(Ax^k - z^k)$ and an “arbitrary” direction $d^k$, to be selected so as to ensure fast asymptotic convergence. This novel choice of direction ensures that the line-search is feasible at every iteration (i.e., condition (10))
holds for a sufficiently small stepsize) despite the fact that \( d^k \) may not be a direction of descent, as we will see.

- Step 4 will allow us to obtain global convergence rates, and it comes at no cost since vectors \( \tilde{y}^k, \hat{x}^k, \hat{z}^k \) have already been computed in the line-search. In a sense, this step robustifies the algorithmic scheme.

By appropriately choosing \( d^k \), the algorithm is able to greatly improve the convergence of AMA: we will prove that the algorithm converges with superlinear asymptotic rate when Newton-type directions are selected. For this reason we refer to Algorithm 1 as Newton-type Alternating Minimization Algorithm (NAMA).

**Remark II.2 (AMA as special case).** If in Algorithm 1 one sets \( d^k = 0 \) for all \( k \), then one can trivially select \( \tau_k = 1 \). In this case, \((\tilde{y}^k, \hat{x}^k, \hat{z}^k) = (y^k, x^k, z^k)\) and Algorithm 1 reduces to AMA, cf. (8).

**Remark II.3** (General equality constrained problems). For any proper, closed, convex \( h : \mathbb{R}^r \rightarrow \mathbb{R} \) and linear mapping \( B : \mathbb{R}^r \rightarrow \mathbb{R}^m \), a problem of the form

\[
\begin{align*}
\text{minimize } & \ f(x) + h(w) \quad \text{subject to } \ Ax+Bw = b \quad \text{(P')}
\end{align*}
\]

can be rewritten as (P) by letting

\[
g(z) = (Bh)(b-z) = \inf_{w \in \mathbb{R}^r} \{ h(w) \mid Bw = b-z \}. \quad (11)
\]

Function \((Bh)\) is the image of \( h \) under \( B \), see [17, Thm. 5.7] and discussion thereafter. If we further assume \( \text{ri}(\text{dom } h^*) \cap \text{range } (B^T) \neq \emptyset \), then \((Bh)\) is proper, closed, convex, see [17, Thm. 16.3], therefore \( g \) in (11) satisfies Assumption 1(iii) (if \( h \) is piecewise linear-quadratic then it is sufficient to assume \( \text{dom } h^* \cap \text{range } (B^T) \neq \emptyset \), see [20, Cor. 11.33(b)]). In this case steps (8b) and (8c) of AMA become

\[
\begin{align*}
\tilde{w}^k &= \arg\min_{w \in \mathbb{R}^r} \{ g(w) + \langle \tilde{y}^k, Bw \rangle + \frac{\gamma}{2} \|Ax^k + Bw - b\|^2 \} \\
y^{k+1} &= y^k + \gamma( Ax^k + Bw^k - b ).
\end{align*}
\]

Similar modifications allow to adapt NAMA to this more general setting: in light of these observations, what follows readily applies to problems of the form (P').

**B. Quasi-Newton directions**

There is freedom in selecting \( d^k \) in Algorithm 1. To accelerate the convergence of the iterates, one possible choice is to compute fast converging directions for the system of nonlinear equations \( R_\gamma(y) = 0 \) characterizing dual optimal points, cf. (7). Specifically, in Algorithm 1 one can set

\[
d^k = B_k^{-1}(Ax^k - z^k), \quad (12)
\]

for a sequence of nonsingular matrices \((B_k)_{k \in \mathbb{N}}\) approximating in some sense the Jacobian \( J_{R_\gamma} \) at the limit point of the dual iterates \((\tilde{y}^k)_{k \in \mathbb{N}}\). In quasi-Newton methods, starting from an initial nonsingular matrix \( B_0 \), the sequence of matrices \((B_k)_{k \in \mathbb{N}}\) is determined by low-rank updates that satisfy the secant condition: in Algorithm 1 fast asymptotic convergence can be proved if

\[
B_{k+1}p^k = q^k \quad \text{with} \quad \begin{cases} p^k = \tilde{y}^k - y^k, \\
q^k = (z^k - Ax^k) - (\hat{z}^k - Ax^k), \end{cases}
\]

as will be discussed in Section VI. Note that all quantities required to compute the vectors \( p^k, q^k \) are available as by-product of the iterations.

In [29] the modified Broyden update is proposed, that prescribes rank-one updates of the form

\[
\begin{align*}
\text{Broyden } & \quad B_{k+1} = B_k + \theta_k \frac{(q^k - B_k p^k)(p^k)^T}{\|p^k\|^2}, \\
\text{BFGS } & \quad B_{k+1} = B_k + \frac{q^k(q^k)^T}{\langle q^k, p^k \rangle} - \frac{B_k p_k (B_k p_k)^T}{\langle p^k, B_k p^k \rangle}. \quad (13)
\end{align*}
\]

Here, \((\theta_k)_{k \in \mathbb{N}} \subset [0, 2]\) is a sequence used to ensure that all terms in \((B_k)_{k \in \mathbb{N}}\) are nonsingular, so that (12) is well defined. The original Broyden method [30] is obtained with \( \theta_k \equiv 1 \).

Probably the most popular quasi-Newton scheme is BFGS, which prescribes the following rank-two updates

\[
B_{k+1} = B_k + \frac{q^k(q^k)^T}{\langle q^k, p^k \rangle} - \frac{B_k p_k (B_k p_k)^T}{\langle p^k, B_k p_k \rangle}. \quad (14)
\]

Note that in this case matrices \( B_k \) are symmetric, and in fact the fast asymptotic properties of BFGS are guaranteed only if the Jacobian \( J_{R_\gamma} \) is symmetric [31] at the problem solution. This is not the case in our setting (cf. Example V.3) although we have observed that (14) often outperforms other non-symmetric updates such as (13) in practice.

Using the Sherman-Morrison-Woodbury identity in (13) and (14) allows to directly store and update \( H_k = B_k^{-1} \), so that \( d^k \) can be computed without inverting matrices or solving linear systems.

Ultimately, instead of storing and operating on dense \( m \times m \) matrices, limited-memory variants of quasi-Newton schemes keep in memory only a few (usually 3 to 30) most recent pairs \((p^k, q^k)\) implicitly representing the approximate inverse Jacobian. Their employment considerably reduces storage and computations over the full-memory counterparts, and as such they are the methods of choice for large-scale problems. The most popular limited-memory method is probably L-BFGS, which is based on the update (14), but efficiently computes matrix-vector products with the approximate inverse Jacobian using a two-loop recursion procedure [32]–[34].

**III. ALTERNATING MINIMIZATION ENVELOPE**

The fundamental tool enabling fast convergence of Algorithm 1 is the alternating minimization envelope function associated with (P). This is precisely the (negative) augmented Lagrangian function, evaluated at the primal points given by the alternating minimization steps.

**Definition III.1** (Alternating minimization envelope). The alternating minimization envelope (AME) for (P), with parameter \( \gamma > 0 \), is the function

\[
\psi_\gamma(y) = -L_\gamma(x(y), z_\gamma(y), y).
\]

The first observation that we make relates the alternating minimization envelope in Definition III.1 with the concept of forward-backward envelope.

**Theorem III.2.** Function \( \psi_\gamma \) is the forward-backward envelope (cf. [11, Def. 2.1]) associated with the dual problem (D):

\[
\psi_\gamma(y) = f^*(-A^Ty) + g^*(T_\gamma(y)) + \frac{\gamma}{2} \|Ax(y) - z_\gamma(y)\|^2 + \gamma \langle Ax(y), z_\gamma(y) - Ax(y) \rangle. \quad (15)
\]
Proof. The optimality conditions for \(x(y)\) and \(z_\gamma(y)\) are
\[
\partial f(x(y)) \ni -A^T y, \quad (16a)
\]
\[
\partial g(z_\gamma(y)) \ni T_\gamma(y) = y + \gamma(Ax(y) - z_\gamma(y)). \quad (16b)
\]
From these, using (2), we obtain
\[
f(x(y)) + f^*(-A^T y) = -\langle Ax(y), y \rangle \quad (17a)
\]
\[
g(z_\gamma(y)) + g^*(T_\gamma(y)) = (z_\gamma(y), T_\gamma(y)) \quad (17b)
\]
Summing (17) and rearranging the terms we get (15). \(\square\)

An alternative expression for \(\psi_\gamma\) in terms of the Moreau envelope of \(\gamma\) is as follows, see [10]:
\[
\psi_\gamma(y) = f^*(-A^T y - \frac{3}{2} \|Ax(y) - z_\gamma(y)\|^2 + (\gamma^*)^\gamma(y + \gamma Ax(y)). \quad (18)
\]

The AME enjoys several favorable properties, some of which we now summarize. For any \(\gamma > 0\), \(\psi_\gamma\) is (strictly) continuous over \(\mathbb{R}^m\), whereas if \(\gamma\) is small enough then the problem of minimizing \(\psi_\gamma\) is equivalent to solving (D). These properties are listed in the next result.

**Theorem III.3.** For any \(\gamma > 0\), \(\psi_\gamma\) is a strictly continuous function on \(\mathbb{R}^m\) satisfying

(i) \(\psi_\gamma(y) \leq \psi(y) + \frac{\gamma^2}{2\mu_f} \|Ax(y) - z_\gamma(y)\|^2\),

(ii) \(\psi_\gamma(y) \geq \psi(T_\gamma(y)) + \frac{1}{2\mu_f} \|A^T(\gamma Ax(y) - z_\gamma(y))\|^2\),

for any \(y \in \mathbb{R}^m\). In particular, if \(\gamma < \mu_f/\|A\|^2\), then the following also holds

(iii) \(\inf \psi_\gamma = \inf \psi \quad \text{and} \quad \argmin \psi_\gamma = \argmin \psi\).

**Proof.** Strict continuity of \(\psi_\gamma\) follows immediately by the expression (18).

\(\blacklozenge \) III.3(ii): follows by Lem. A.1 using \(w = y\).

\(\blacklozenge \) III.3(ii): due to strong convexity of \(f\), \(f^*\) has 1/\(\mu_f\)-Lipschitz gradient, and consequently
\[
f^*(-A^T T_\gamma(y)) \leq f^*(-A^T y - \langle Ax(y), T_\gamma(y) - y \rangle)
\]
\[
+ \frac{1}{2\mu_f} \|A^T(T_\gamma(y) - y)\|^2
\]
\[
= f^*(-A^T y) - \gamma(Ax(y), Ax(y) - z_\gamma(y))
\]
\[
+ \frac{\gamma^2}{2\mu_f} \|A^T(Ax(y) - z_\gamma(y))\|^2. \quad (19)
\]
Combining (15) with (19):
\[
\psi_\gamma \geq \psi(T_\gamma(y)) - \frac{\gamma^2}{2\mu_f} \|A^T(Ax(y) - z_\gamma(y))\|^2
\]
\[
+ \frac{1}{2\mu_f} \|Ax(y) - z_\gamma(y)\|^2
\]
\[
\geq \psi(T_\gamma(y)) + \frac{\gamma^2}{2\mu_f} \|A^T(Ax(y) - z_\gamma(y))\|^2.
\]
\(\blacklozenge \) III.3(iii): easily follows combining III.3(i) and III.3(ii) with \(y = y_\ast \in Y_\ast\), in light of the dual optimality condition (7). \(\square\)

A. Analogy with the dual Moreau envelope

Theorem III.2 highlights a clear connection between the augmented Lagrangian, the forward-backward envelope and the alternating minimization algorithm. This closely resembles the one, first noticed by Rockafellar [35], [36], relating the augmented Lagrangian, the Moreau envelope and the method of multipliers (also known as augmented Lagrangian method) by Hestenes and Powell [37], [38]. Consider the general linear equality constrained convex problem

\[
\begin{align*}
\text{minimize} & \quad g(z) \\
\text{subject to} & \quad Bz = b,
\end{align*}
\]

where \(g : \mathbb{R}^m \to \mathbb{R}\) is proper, closed, convex, \(B \in \mathbb{R}^{m \times k}\) and \(b \in \mathbb{R}^m\). When applied to the dual of (20), namely
\[
\begin{align*}
\text{minimize} & \quad \omega(y) = g^*(-B^T y) + \langle b, y \rangle,
\end{align*}
\]
the proximal minimization algorithm [39, §5.2] is equivalent to the following augmented Lagrangian method
\[
\begin{align*}
z^{k+1} = \argmin_{z \in \mathbb{R}^m} \{g(z) + \langle y^k, Bz - b \rangle + \frac{\gamma}{2} \|Bz - b\|^2 \}
\end{align*}
\]

Therefore the forward-backward and Moreau envelope functions have the same nice interpretation in terms of augmented Lagrangian, when they are applied to the dual of equality constrained convex problems: in a sense, Theorem III.2 extends and generalizes the classical result on the dual Moreau envelope, by allowing for an additional variable \(x\) and a strongly convex term \(f\) in the problem.

**IV. CONVERGENCE**

We now turn our attention to the global convergence properties of Algorithm 1. In light of Remark II.2, the results in this section directly apply to AMA, which is a special case of NAMA.

**Remark IV.1** (Termination of line-search). The line-search step 3 is well defined regardless of the choice of \(d^k\); at any iteration \(k\), condition (10) holds for \(i_k\) sufficiently large. To see this, suppose that \(\|Ax^k - z^k\| > 0\) (otherwise \((x^k, y^k)\) is a primal-dual solution). Then, since \(\gamma < \mu_f/\|A\|^2\), Theorem III.3 implies that
\[
\psi_\gamma(T_\gamma(y^k)) < \psi_\gamma(y^k). \quad (21)
\]
Since \(\tilde{y}^k \to T_\gamma(y^k)\) as \(\tau_k \to 0\) and \(\psi_\gamma\) is continuous, then necessarily \(\psi_\gamma(\tilde{y}^k) \leq \psi_\gamma(y^k)\) for \(\tau_k\) sufficiently small. \(\square\)

**Remark IV.2** (Bounded iteration complexity). In the best case where \(\tau_k = 1\) is accepted in step 3, exactly two alternating minimizations are performed at iteration \(k\). In practice, one can also impose a lower bound \(\tau_{\min} > 0\) for \(\tau_k\) when \(\tau_k < \tau_{\min}\) then the ordinary AMA update \(y^{k+1} = y^k + \gamma(Ax^k - z^k)\) is executed and the algorithm proceeds to the next iteration. This strategy results in a bounded iteration complexity for NAMA, and does not affect the convergence results of this and later sections.

Theorem III.3 ensures that the following chain of inequalities, which will be fundamental for convergence results, holds in Algorithm 1:
\[
\begin{align*}
\psi(y^{k+1}) & \leq \psi_\gamma(\tilde{y}^k) \quad (22a) \\
& \leq \psi_\gamma(y^k) \quad (22b) \\
& \leq \psi(y^k) - \frac{\gamma}{2} \|Ax^k - z^k\|^2. \quad (22c)
\end{align*}
\]
In particular, Algorithm 1 is a descent method for \(\psi\).
We now prove that the iterates of (1) converge to the dual optimal cost and to the primal solution. Moreover, global convergence rates are provided.

**Theorem IV.3** (Global convergence). In Algorithm 1:

(i) \(x^k \rightarrow x_\star, z^k \rightarrow Ax_\star\), and all cluster points of \((y^k)_{k \in \mathbb{N}}\) are dual optimal, i.e., they belong to \(Y_\star\);

(ii) if \(0 \in \text{int}(\text{dom } g - \text{A dom } f)\) then \(\psi(y^k) \searrow \inf \psi \) with global rate \(O(1/k), \) and \(x^k \rightarrow x_\star\) with global rate \(O(1/\sqrt{k})\);

(iii) if \(f\) and \(g\) are piecewise linear-quadratic then \(\psi(y^k) \searrow \inf \psi\) with global Q-linear rate, and \(x^k \rightarrow x_\star\) with global R-linear rate.

**Proof.**

\(\star\) IV.3(i): by (22c), for all \(i \geq 0\) we have

\[ \inf \psi \leq \psi(y^{i+1}) \leq \psi(y^i) - \frac{\gamma}{2} \|Ax^i - z^i\|^2. \]

By summing up the inequality for \(i = 1, \ldots, k\) we obtain

\[ \inf \psi \leq \psi(y^{k+1}) \leq \psi(y^1) - \frac{\gamma}{2} \sum_{i=1}^k \|Ax^i - z^i\|^2 \]

(the sum starts from \(i = 1\) since \(y^0\) may be dual infeasible). In particular (cf. (9)) \(R_\tau(y^k) = z^k - Ax^k \rightarrow 0\), and since \(R_\tau\) is continuous, necessarily all cluster points of \((y^k)_{k \in \mathbb{N}}\) are optimal. Thus, it follows from Lem. A.2 that the sequence \((x^k)_{k \in \mathbb{N}}\) is bounded. Let \(K \subseteq \mathbb{N}\) and \(\bar{x}\) be such that \((x^k)_{k \in K} \rightarrow \bar{x};\) then, since \(Ax_k - z_k \rightarrow 0\) we also have that \((x^k)_{k \in K} \rightarrow \bar{Ax}\). By multiplying (16b) on the left by \(A^T\) and summing (16a) we obtain \(\gamma A^T (Ax_k - z_k) \in \partial f(x_k) + A^T \partial g(z_k)\). By letting \(K \ni k \rightarrow \infty\), from outer semicontinuity of the subdifferential we obtain that

\[ 0 \in \partial f(\bar{x}) + A^T \partial g(\bar{Ax}) \subseteq \partial (f + g \circ A)(\bar{x}) \]

where the last inclusion follows from [17, Thms. 23.8 and 23.9]. Thus, \(\bar{x}\) is optimal, and being \(x_\star\) the unique primal optimal (due to strong convexity), necessarily \(\bar{x} = x_\star\). From the arbitrariness of the cluster point we conclude that \(x^k \rightarrow x_\star\) and \(z^k \rightarrow Ax_\star\).

\(\star\) IV.3(ii): the assumed condition is equivalent to \(Y_\star\) being nonempty and compact, see [26, Thm. 5.2.1], which implies that \(\psi\) has bounded level sets [20, Prop. 3.23]. The proof proceeds similarly to that of [8, Thm. 4]. Let \(D > 0\) be such that \(\text{dist}(y, Y_\star) < D\) for all points \(y \in \{y \in \mathbb{R}^m \mid \psi(y) \leq \psi(y^0)\}\). From [11, Prop. 2.5] we know that \(\psi_0\) is \(\psi^\star\) (the Moreau envelope of \(\psi\)). Therefore,

\[ \psi(y^{k+1}) \leq \psi(y^k) \leq \psi^\star(y^k) = \min_{w \in \mathbb{R}^m} \left\{ \psi(w) + \frac{1}{2\gamma} \|w - y^k\|^2 \right\} \]

and in particular, for \(y_\star \in \text{argmin } \psi \),

\[ \psi(y^{k+1}) \leq \min_{\alpha \in [0,1]} \left\{ \psi(\alpha y_\star + (1 - \alpha) y^k) + \frac{\alpha^2}{2\gamma} \|y^k - y_\star\|^2 \right\} \]

where in last inequality we used convexity of \(\psi\). In case \(\psi(y^0) - \inf \psi \geq D^2/\gamma\), then the optimal solution of the latter problem for \(k = 0\) is \(\alpha = 1\), and \(\psi(y^1) - \inf \psi \leq D^2/2\gamma\). Otherwise, the optimal solution is

\[ \alpha = \frac{\gamma}{2\gamma} (\psi(y^k) - \inf \psi) \leq \frac{\gamma}{2\gamma} (\psi(y^0) - \inf \psi) \leq 1 \]

and we obtain

\[ \psi(y^{k+1}) \leq \psi(y^k) - \frac{\gamma}{2\gamma} (\psi(y^k) - \inf \psi)^2. \]

By letting \(\lambda_k = \frac{\gamma}{2\gamma} \inf \psi\) in the last inequality becomes

\[ \lambda_k^{k+1} \leq \lambda_k^{k} - \frac{\gamma}{2\gamma} \lambda_k^{k} - \frac{\gamma}{2\gamma} \lambda_k^{k+1}. \]

By multiplying both sides by \(\lambda_k \lambda_{k+1}\) and rearranging,

\[ \lambda_k^{k+1} \geq \lambda_k + \frac{\gamma}{2\gamma} \lambda_k^{k+1} \geq \lambda_k + \frac{\gamma}{2\gamma} \]

where the latter inequality follows from the fact that the sequence \((\psi(y^k))_{k \in \mathbb{N}}\) is nonincreasing, as shown in (22). By telescoping the inequality we obtain

\[ \lambda_k \geq \lambda_0 + k \frac{\gamma}{2\gamma} \geq k \frac{\gamma}{2\gamma}, \]

and therefore \(\psi(y^k) - \inf \psi \leq 2D^2/k\gamma\). This, together with Lem. A.2, proves IV.3(ii).

\(\star\) IV.3(iii): since the primal optimum is finite (see Rem. II.1), if \(f\) and \(g\) are piecewise linear-quadratic then \(Y_\star\) is nonempty, see [20, Thm. 11.42, Ex. 11.43]. Using (22) we have that

\[ \psi(y^k) - \psi(y^{k+1}) \geq \frac{\gamma}{2} \|Ax^k - z^k\|^2. \]

Furthermore, using Lem. A.1 with \(w = y_\star = \Pi Y_\star y^k\) and \(y = y_k\), we obtain

\[ \psi(y^{k+1}) - \inf \psi \leq \psi(y^k) - \inf \psi \leq \psi(y^k) - \psi(y^{k+1}) \]

\[ \leq \|Ax^k - z^k\|^2 \leq \frac{1}{2\gamma} \|Ax^k - z^k\|^2, \]

where first inequality is due to (22c). This implies

\[ \psi(y^{k+1}) - \inf \psi \leq \|Ax^k - z^k\|^2 \left( \frac{\text{dist}(y^k, Y_\star)}{\|Ax^k - z^k\|^2} - \frac{\gamma}{2} \right), \]

which, by using (23), yields

\[ \psi(y^{k+1}) - \inf \psi \leq \left( 1 - \gamma \frac{\text{dist}(y^k, Y_\star)}{\|Ax^k - z^k\|^2} \right) (\psi(y^k) - \inf \psi). \]

It follows from [20, Thm. 11.14] that \(f^*\) and \(g^*\) are convex piecewise linear-quadratic in this case, and so is \(\psi\). Therefore by [40, Thm. 2.7] \(\psi\) enjoys the following quadratic growth condition: for any \(\nu > 0\) there is \(\alpha > 0\) such that

\[ \frac{\alpha}{2} \text{dist}^2(y, Y_\star) \leq \psi(y) - \inf \psi \quad \forall y : \psi(y) - \inf \psi \leq \nu, \]

which by [41, Cor. 3.6] is equivalent to the following error bound condition for some \(\beta > 0\)

\[ \text{dist}(y, Y_\star) \leq \beta \|Ax(y) - z_\star(y)\|. \]

In general we can prove local linear convergence of Algorithm 1 provided that \(\partial f\) and \(\partial g\) are calm, according to the following definition (see [42, Sec. 3H, Ex. 3H.4]).

**Definition IV.4** (Calmness of a mapping). A multi-valued mapping \(F : \mathbb{R}^m \rightrightarrows \mathbb{R}^n\) is said to be calm at \(y \in \mathbb{R}^m\) if there is a neighborhood \(U \ni y \in F(y)\) such that

\[ F(y) \cap U \subseteq F(y \pm O(\|y - \bar{y}\|)), \quad \forall y \in \mathbb{R}^m. \]

We simply say that \(F\) is calm at \(y \in \mathbb{R}^m\) (with no mention of \(x\)) if it is calm at \(y \in \mathbb{R}^m\) for all \(x \in F(y)\).

Calmness is a very common property of the subdifferential mapping. The subdifferential of all piecewise linear-quadratic functions is calm everywhere, as follows from [42, Prop. 3H.1]. Other examples include the nuclear and spectral norms [43]. Smooth functions, i.e., with Lipschitz gradient, clearly
have calm subdifferential: this includes Moreau envelopes of closed, convex functions, such as the Huber loss for robust estimation, and commonly used loss functions such as the squared Euclidean norm and the logistic loss.

Calmness is equivalent to metric subregularity of the inverse mapping ([42, Thm. 3H.3]: from [44, Prop. 6, Prop. 8] we then deduce that the indicator functions of $\ell_1$, $\ell_\infty$ and Euclidean norm balls all have calm subdifferentials.

The following result holds. Its proof is analogous to the one of [41, Thm. 4.2], although our assumption of calmness is equivalent to metric subregularity of $\partial f^*$ and $\partial g^*$, which is implied by the firm convexity assumed in [41].

**Theorem IV.5** (Local linear convergence). **Suppose that the following hold for (P):**

(i) $0 \in \text{int}(\text{dom } g - A \text{ dom } f)$ (nonempty, compact $Y$);

(ii) $0 \in \partial f(f + g \circ A)(x_\ast)$ (strict complementarity).

**Suppose also that $\partial f$ is calm at $x_\ast$ and $\partial g$ is calm at $Ax_\ast$. Then in Algorithm 1 eventually $\psi(y) \rightarrow \inf \psi$ with Q-linear rate and $x^k \rightarrow x_\ast$ with R-linear rate.**

**Proof.** As discussed in the proof of Thm. IV.3(iii), it suffices to show that an error bound of the form (25) holds for some $\beta, \nu > 0$.

The assumed calmness properties of $\partial f$ and $\partial g$ are equivalent to metric subregularity of $\partial f^*$ at $-A^\top y_\ast$ for $x_\ast$, and of $\partial g^*$ at $y_\ast$ for $Ax_\ast$, see [42, Thm. 3H.3], for all $y_\ast \in Y$. This can be seen, using [45, Thm. 3.3], to be equivalent to the following: there exist $c_{\ast} > 0$ and a neighborhood $U_{y_\ast}$ of $y_\ast$, such that for all $y \in U_{y_\ast}$,

\[
\begin{align*}
f^*(-A^\top y) &\geq f^*(-A^\top y_\ast) + \langle x_\ast, A^\top (y_\ast - y) \rangle + \frac{c_{\ast}}{2} \text{dist}^2(-A^\top y, (\nabla f^*)^{-1}(x_\ast)), \\
g^*(y) &\geq g^*(y_\ast) + \langle Ax_\ast, y - y_\ast \rangle + \frac{c_{\ast}}{2} \text{dist}^2(y, (\nabla g^*)^{-1}(Ax_\ast)).
\end{align*}
\]

Since $Y_\ast \subseteq \bigcup_{y_\ast \in Y} U_{y_\ast}$ and $Y_\ast$ is nonempty and compact (due to IV.5(i)), we may select a finite subset $W \subseteq Y_\ast$, such that $Y_\ast \subseteq U_{Y_\ast} = \bigcup_{y_\ast \in W} U_{y_\ast}$. Summing the above inequalities for all $y_\ast \in W$, and denoting $c = \min \{c_{\ast}, \ |y_\ast \in W\} > 0$, we obtain

\[
\psi(y) \geq \inf \psi + \frac{c}{2} \left[ \text{dist}^2(-A^\top y, \partial f(x_\ast)) + \text{dist}^2(y, \partial g(Ax_\ast)) \right] \quad (26)
\]

for all $y \in U_{Y_\ast}$, where we have also used $(\nabla f^*)^{-1} = \partial f$ and $(\nabla g^*)^{-1} = \partial g$. Note that IV.5(i) implies strict feasibility, and therefore from Lem. A.3, and the fact that for any $a, b \in \mathbb{R}$, $a^2 + b^2 \geq 2ab$, we obtain that (26) implies

\[
\psi(y) \geq \inf \psi + \frac{c}{2} \text{dist}^2(y, Y_\ast), \quad \forall y \in U_{Y_\ast},
\]

for some $\kappa > 0$, i.e., $\psi$ satisfies the quadratic growth condition, which by [41, Cor. 3.6] is equivalent to the error bound condition (25). This completes the proof. □

**Remark IV.6** (Backtracking on $\gamma$). In practice, no prior knowledge of the global Lipschitz constant $||A||^2/\mu_f$ is required for Algorithm 1: instead of a fixed parameter $\gamma$, one can adaptively determine a sequence $\langle \gamma_k \rangle_{k \in \mathbb{N}}$ essentially ensuring that inequalities (21) (which guarantees termination of the line-search step 3) and (22a) (which guarantees descent) hold at every iteration. This is done as follows. Select $\alpha \in (0, 1)$ and initialize $\gamma_0 > 0$. At iteration $k$, let $\bar{y}^k = y^k + \gamma_k (Ax^k - z^k)$ and $\bar{x}^k = x(\bar{y}^k)$, and if

\[
f(x^k) > f(\bar{x}^k) - \langle A^\top \bar{y}^k, x^k - \bar{x}^k \rangle + \frac{\alpha_k^2}{2} ||Ax^k - z^k||^2,
\]

then $\gamma_k \leftarrow \gamma_k/2$ and restart the iteration. Similarly if

\[
f(\bar{x}^k) > f(x^{k+1}) - \langle A^\top y^{k+1}, x^k - x^{k+1} \rangle + \frac{\alpha_k^2}{2} ||Ax^k - z^k||^2.
\]

As soon as $\gamma_k \leq \alpha_k/||A||^2$, the two inequalities above will never hold. As a consequence, $\gamma_k$ will be decreased only a finite number of times and will be constant starting from some iteration $k$. The inequalities above are obtained by imposing the usual quadratic upper bound on $f^*(\cdot - A^\top)$, due to smoothness, and applying the conjugate subgradient theorem (2) in light of (16a). This procedure of adaptively adjusting $\gamma_k$ is analogous to what is done in practice in (fast) AMA, see [9, Rem. 3.4] and [7, §3, §4], and does not affect the validity of Thms IV.3 and IV.5.

**V. First- and Second-Order Properties**

Algorithm 1 is a line-search method for the unconstrained minimization of $\psi_\gamma$, which, by Theorem III.3(iii), is equivalent to solving (D). To enable fast convergence of the iterates, we can apply ideas from smooth unconstrained optimization in selecting the sequence $\langle \delta \rangle_{k \in \mathbb{N}}$ of directions. To this end, differentiability of $\psi_\gamma$ around dual solutions $y_\ast$ is a desirable property. We will now see that this is implied by generalized second-order properties of $f$ around $x_\ast$, which are introduced in the following assumption. Analogous assumptions on $g$ further ensure that $\psi_\gamma$ is (strictly) twice differentiable at $y_\ast$. The interested reader is referred to [20] for an extensive discussion on (second-order) epi-differentiability.

**Assumption 2**. The following hold with respect to a primal-dual solution $(x_\ast, y_\ast)$ to (P)-(D):

(i) $f$ is strictly twice epi-differentiable at all $x \in \text{dom } f$ close enough to $x_\ast$, and in particular the second-order epi-derivative at $x_\ast$, for $-A^\top \psi$, is, for $w \in \mathbb{R}^n$,

\[
d^2 f(x_\ast - A^\top \psi)[w] = \langle H_f \psi, w \rangle + \delta_{S_f}(w),
\]

(27) where $S_f$ is a linear subspace of $\mathbb{R}^n$ and $H_f \in \mathbb{R}^{n \times n}$;

(ii) $g$ is (strictly) twice epi-differentiable at $Ax_\ast$ for $y_\ast$, with

\[
d^2 g(Ax_\ast, y_\ast)[w] = \langle H_g w, w \rangle + \delta_{S_g}(w),
\]

(28) for all $w \in \mathbb{R}^m$, where $S_g$ is a linear subspace of $\mathbb{R}^m$ and $H_g \in \mathbb{R}^{m \times m}$.

When the stronger condition in parenthesis holds we will say that the assumptions are strictly satisfied.

Without loss of generality, we consider $H_f$ and $H_g$ symmetric and positive semidefinite, satisfying $\text{range}(H_f) = S_f$, $\text{null}(H_f) = S_f^\perp$, $\text{range}(H_g) \subseteq S_g$ and $\text{null}(H_g) \supseteq S_g^\perp$.

The requirements on $H_f$ and $H_g$ can indeed be made without loss of generality: matrix $H_f = \frac{1}{2} \Pi_{S_f}(H_f + H_f^\top) \Pi_{S_f}$ has the desired properties and satisfies (27) provided $H_f$ does, and similarly for $H_g$. In particular, it holds that

\[
H_f = \Pi_{S_f} H_f \Pi_{S_f} \quad \text{and} \quad H_g = \Pi_{S_g} H_g \Pi_{S_g}.
\]

(29)
Theorem V.1 (Differentiability of $\psi_{\gamma}$). Suppose that Assumption 2(i) holds for a primal-dual solution $(x_*, y_*)$. Then $\psi_{\gamma}$ is of class $C^1$ around $y_*$, with

$$\nabla \psi_{\gamma}(y) = Q_{\gamma}(y) R_{\gamma}(y)$$

where $Q_{\gamma}(y) = I - \gamma \alpha V^2 f^* (-A^T y) A^T$.

Proof. From Lem. A.4 it follows that $f = f^* \circ (-A^T)$ is of class $C^2$ around $y_*$. The claim now easily follows from the chain rule of differentiation applied to (18), by using (4).

Twice differentiability of $\psi_{\gamma}$ at a dual solution $y_*$ is very important: when Newton-type directions are used, this implies that eventually unit stepsize will be accepted and fast asymptotic convergence will take place. In other words, unlike standard nonsmooth merit functions for constrained optimization, $\psi_{\gamma}$ does not prevent the acceptance of unit stepsize.

Theorem V.2 ( Twice differentiability of $\psi_{\gamma}$). Suppose that Assumption 2 (strictly) holds with respect to a primal-dual solution $(x_*, y_*)$. Then, (i) $R_{\gamma}(y)$ is (strictly) differentiable at $y_*$ with Jacobian

$$JR_{\gamma}(y_*) = \gamma^{-1} [I - P_{\gamma}(y_*)Q_{\gamma}(y_*)];$$

where $Q_{\gamma}$ is as in Theorem V.1 and $P_{\gamma}(y_*) = J \text{prox}_{\gamma g^*} (y_* + \gamma A \nabla f^*(-A^T y_*)) = \Pi_S [I + \gamma H_{g}^T]^{-1} \Pi_S$ (31) with $S = S_\gamma + \text{range}(H_g)$; (ii) $\psi_{\gamma}$ is (strictly) twice differentiable at $y_*$ with symmetric Hessian

$$\nabla^2 \psi_{\gamma}(y) = \gamma^{-1} Q_{\gamma}(y_*)[I - P_{\gamma}(y_*)Q_{\gamma}(y_*)].$$

Proof. Let $\hat{f} = f^* \circ (-A^T)$ and $L_\hat{f} = \nu_{\gamma}/\|A\|^2$. We know from [25, Thms. 3.8, 4.1] and [20, Thm. 13.21] that $\text{prox}_{\gamma g^*}$ is (strictly) differentiable at $y_* - \gamma \nabla \hat{f}(y_*)$ if and only if $g$ (strictly) satisfies Assumption 2(ii); in fact, by (5) we know that $Ax_* = -\nabla \hat{f}(y_*)$. Moreover, due to Lem. A.4, $\hat{f} \in C^2$ in a neighborhood of $y_*$ and in particular $\nabla \hat{f}$ is strictly differentiable at $y_*$. The formula for $JR_{\gamma}(y_*)$ follows from (4) and the chain rule of differentiation.

We now prove the claimed expression for $P_{\gamma}(y_*)$. We may invoke Lem. A.5 and apply [20, Ex. 13.45] to the tilted function $g + \langle \nabla \hat{f}(y_*) \cdot \cdot \rangle$ which tells us that for all $d \in \mathbb{R}^m$

$$P_{\gamma}(y_*) d = \text{prox}_{\gamma/2d^2g^*}(y_*, |Ax_*|)(d) = \text{argmin}_{d \in S} \left\{ \frac{1}{2} \|d'\|^2 + \frac{1}{\gamma} \|d - d'\|^2 \right\} = \Pi_S \text{argmin}_{d' \in \mathbb{R}^n} \left\{ \frac{1}{2} \langle \Pi_S d', H_{g}^T \Pi_S d' \rangle + \frac{1}{\gamma} \| \Pi_S d' - d \|^2 \right\} = \Pi_S \Pi_S[I + \gamma H_{g}^T] \Pi_S d' - \Pi_S$$

where $\dagger$ indicates the pseudo-inverse. Observe now that, since $\text{range}(H_{g}) = \text{range}(H_g) \subseteq \mathbb{S}$, we have

$$\Pi_S[I + \gamma H_{g}^T] \Pi_S = AB \quad \text{for} \quad A = I + \gamma H_{g}^T \quad \text{and} \quad B = \Pi_S.$$}

Moreover,

$$\text{range}(A^T AB) \subseteq \text{range}(B), \quad \text{range}(B^T BA) \subseteq \mathbb{R}^n = \text{range}(A),$$

therefore we can apply [46, Facts 6.4.12 (i)-(ii) and 6.1.6 (xxiii)] to see that $(\Pi_S[I + \gamma H_{g}^T] \Pi_S)_{\dagger} = \Pi_S[I + \gamma H_{g}^T]^{-1}$, yielding (31).

Since $R_{\gamma}(y_*) = 0$ from [11, Lem. 6.2] it follows that $\nabla \psi_{\gamma} = Q_{\gamma} R_{\gamma}$ is (strictly) differentiable at $y_*$, provided that $Q_{\gamma}$ is (strictly) continuous at $y_*$ and $R_{\gamma}$ is (strictly) differentiable at $y_*$. A simple application of the chain rule of differentiation concludes the proof of V.2(ii).

To better understand the requirements of Assumption 2, let us consider the following simple but significant example: when $f$ is $C^2$ and $g \circ A$ models linear inequality constraints, Assumption 2 is implied by strict complementarity.

Example V.3 ($C^2$ functions subject to polyhedral constraints). Consider problems of the form

$$\min_{x \in \mathbb{R}^n} f(x) + \delta_C(Ax),$$

where $g = \delta_C$ is the indicator of $C = \{ z \in \mathbb{R}^m \mid z \leq b \}, \ b \in \mathbb{R}^m, \ f \in C^2$. In this case Assumption 2(ii) holds with $H_f = \nabla^2 f(x_*)$, $S_f = \mathbb{R}^n$ (therefore $\Pi_{S_f} = \text{Id}$), see [20, Ex. 13.8]. Regarding Assumption 2(ii), one can use [20, Ex. 13.17] to see that

$$d^2 g(Ax_*, y_*)[w] = \delta_{K(Ax_*, y_*)}(w),$$

where $K$ is the critical cone. Denoting as $T_C(y)$ the tangent cone of set $C$ at $y \in C$, and as $J = \{ i \mid (Ax_i) = b_i \}$ the set of active constraints at the solution $x_*$, the critical cone is given by

$$K(Ax_*, y_*) = \{ w \in T_C(Ax_*) \mid (y_*, w) = 0 \} = \{ w \mid (y_*, w) = 0, w_i \leq 0 \forall i \in J \}.$$
(b) $\nabla^2 \psi_\gamma(y_*)$ is nonsingular (in fact, positive definite);
(c) $JR_\gamma(y_*)$ is nonsingular (in fact, similar to a symmetric and positive definite matrix);
(d) $y_*$ is a strong minimum for $\psi_\gamma$.

Proof. 
\* V.4(b) $\iff$ V.4(c): Let $P = P_\gamma(y_*)$ and $Q = Q_\gamma(y_*)$ for brevity. Notice first that, due to Thm. III.3(iii), $y_*$ minimizes $\psi_\gamma$ and therefore $\nabla^2 \psi_\gamma(y_*) \succeq 0$. Moreover, since $Q$ is symmetric and positive definite,

$$JR_\gamma(y_*) = \gamma^{-1}(I - PQ) \sim Q^{-1/2} \nabla^2 \psi_\gamma(y_*) Q^{-1/2}$$

the latter matrix being symmetric and positive semidefinite, where $\sim$ denotes the simililitude relation.

\* V.4(b) $\iff$ V.4(d): trivial since $\nabla^2 \psi_\gamma(y_*)$ exists.

\* V.4(d) $\iff$ V.4(a): the right implication is trivial since $\psi \leq \psi_\gamma$ and $\psi_\gamma(y_*) = \psi(y_*)$ as it follows from Thm. III.3. Suppose now that there exist $c, \varepsilon > 0$ such that $\psi(y) - \psi_\gamma(y) \geq c \|y - y_*\|^2$ for all $y \in B(y_*, \varepsilon)$. Since $g^*$ is convex, it follows that $\text{prox} \_\gamma^*$ is 1-Lipschitz continuous; combined with the fact that $\nabla f^* = \frac{1}{\mu^*}$-Lipschitz continuous, we obtain that the alternating minimization operator $T_\gamma$ is Lipschitz continuous with modulus $\frac{\|A\|^2}{\mu^*}$. Let $\varepsilon' = \mu^*/\|A\| \varepsilon$; since $T_\gamma(y_*) = y_*$, for all $y \in B(y_*, \varepsilon)$ necessarily $T_\gamma(y) \in B(y_*, \varepsilon)$. Therefore, letting $\varepsilon' = \min \left\{ c, \frac{\gamma(1 - \frac{\|A\|^2}{\mu^*})}{\gamma} \right\} > 0$, it follows from Thm. III.3(ii) that for all $y \in B(y_*, \varepsilon')$

$$\psi_\gamma(y) - \psi_\gamma(y_*) \geq \psi(T_\gamma(y)) - \psi_\gamma(y) - \frac{\gamma}{2} \left(1 - \frac{\|A\|^2}{\mu^*}\right) \|y - T_\gamma(y)\|^2 \geq \frac{\varepsilon'}{2} \|T_\gamma(y) - y_*\|^2 + \|y - y_*\|^2 \geq \frac{\varepsilon'}{2} \|y - y_*\|^2.$$ 

This shows that $y_*$ is a strong local minimum for $\psi_\gamma$.

In the context of Example V.3, notice that

$$JR_\gamma(y_*) \text{ is nonsingular } \iff A_1 H_{JJ} A_1^T \text{ is nonsingular.}$$

Since $\nabla^2 f(x_*) > 0$ by assumption, then $H_{JJ} > 0$ and nonsingularity of the Jacobian is equivalent to $A_1$ being full row rank, i.e., linear independence of the active constraints at $x_*$ (the LICQ assumption).

VI. Superlinear Convergence

The following definition (cf. [47, Eq. (7.5.2)]) gives the fundamental condition, on the sequence $(d_k)_{k \in \mathbb{N}}$ of directions, ensuring superlinear asymptotic convergence of Algorithm 1.

**Definition VI.1** (Superlinear directions). For $(y^k)_{k \in \mathbb{N}}$ converging to $y_*$, we say that $(d_k)_{k \in \mathbb{N}}$ is superlinearly convergent w.r.t. $(y^k)_{k \in \mathbb{N}}$ if

$$\lim_{k \to \infty} \frac{\|y^k + d_k - y_*\|}{\|y^k - y_*\|} = 0. \quad (33)$$

When $y_*$ is a strong minimizer, by [41, Cor. 3.6] the error bound (25) holds for some $\beta, \nu > 0$ and $Y_\star = \{y_*\}$. This, by Thm. IV.3(i), implies $y^k \to y_*$. Therefore we have the following result.

**Theorem VI.2.** Suppose that $f$ and $g$ satisfy Assumption 2, and that (D) has a (unique) strong minimizer $y_*$. If (33) holds in Algorithm 1, then

(i) the stepsize $\tau_k = 1$ for all $k$ sufficiently large,
(ii) the cost $\psi(y^k) \to \inf \psi Q$-superlinearly,
(iii) the dual iterates $y^k \to y_*$ Q-superlinearly,
(iv) the primal iterates $x^k \to x_*$ R-superlinearly.

Proof. We know from Thm.s V.2(ii) and V.4(b) that $\psi_\gamma$ is twice differentiable with symmetric and positive definite Hessian $H_* = \nabla^2 \psi_\gamma(y_*)$. We can expand $\psi_\gamma$ around $y_*$ and obtain

$$\frac{\psi_\gamma(y^k + d_k) - \inf \psi_\gamma(y^k)}{\psi_\gamma(y^k) - \inf \psi} \leq \langle H_\gamma(y^k + d_k - y_*), y^k + d_k - y_* \rangle + o(\|y^k + d_k - y_*\|^2)$$

$$\leq \|H_\gamma\| \left(\frac{\|y^k + d_k - y_*\|}{\|y^k - y_*\|}\right)^2 + o(\|y^k + d_k - y_*\|^2)$$

and

$$\lambda_{\min}(H_\gamma) + \left(\frac{o(\|y^k - y_*\|)}{\|y^k - y_*\|}\right)^2$$

which vanishes for $k \to \infty$. In particular, eventually $\psi_\gamma(y^k + d_k) \to \psi_\gamma(y_*)$ and $y^k \to y_*$. In turn, since eventually $\hat{y}^k = y^k + \tau_k d_k = y^k + d_k$, using Thm. III(iii) and (22b) we have

$$\frac{\psi_\gamma(y^k + 1) - \inf \psi_\gamma(y^k)}{\psi_\gamma(y^k + 1) - \inf \psi_\gamma(y^k)} \leq \frac{\psi_\gamma(\hat{y}^k) - \inf \psi_\gamma(y^k)}{\psi_\gamma(y^k) - \inf \psi} \to 0,$$

which proves VI.2(ii). Moreover, (33) reads

$$\|\hat{y}^k - y_*\|^2 / \|y^k - y_*\| \to 0. \quad (34)$$

Now, using nonexpansiveness of $T_\gamma$ (cf. the proof of [19, Thm. 25.8]) one has

$$\|y^{k+1} - y_*\| = \|T_\gamma(\hat{y}^k) - T_\gamma(y_*)\| \leq \|\hat{y}^k - y_*\|$$

which, with (34), proves VI.2(iii). VI.2(iv) follows from VI.2(ii) and Lem. A.2.

When quasi-Newton directions are computed as in (12), superlinear convergence holds provided that the sequence of matrices $(B_k)_{k \in \mathbb{N}}$ satisfies the Dennis-Moré condition given in the following result. Such condition is satisfied for example by the modified Broyden method (13) under standard assumptions of calm semiderivatability of $R_\gamma$, see [48, Thm. 6.8].

**Theorem VI.3** (Dennis-Moré condition). Suppose that $f$ and $g$ strictly satisfy Assumption 2, and that (D) has a (unique) strong minimizer $y_*$. If $(d_k)_{k \in \mathbb{N}}$ is selected according to (12), with

$$\lim_{k \to \infty} \frac{\|B_k - JR_\gamma(y_*) d_k\|}{\|d_k\|} = 0, \quad (35)$$

then $(d_k)_{k \in \mathbb{N}}$ is superlinearly convergent with respect to $(y^k)_{k \in \mathbb{N}}$. In particular, the conclusions of Theorem VI.2 hold.

Proof. From Thm.s V.2(ii) and V.4(c) we know that $R_\gamma$ is strictly differentiable, with nonsingular Jacobian $J_* = JR_\gamma(y_*)$. Let us denote $y^k = z^k - Ax^k = R_\gamma(y^k)$ for simplicity. By using (12) and (35), and by applying the reverse triangle inequality we obtain

$$0 \leq \frac{\|z^k - J_* d_k\|}{\|d_k\|} \geq \frac{\|J_* B_k z^k - r_k\|}{\|d_k\|} - \frac{\|r_k\|}{\|d_k\|} \geq \alpha - \frac{\|r_k\|}{\|d_k\|},$$

where $\alpha = \sqrt{\lambda_{\min}(J_* J_*)} > 0$ since $J_*$ is nonsingular. Therefore,

$$\lim_{k \to \infty} \frac{\|r_k\|}{\|d_k\|} \geq \alpha$$
and as a consequence $\|d^k\| \leq (2/\alpha) \|r^k\|$ for all $k$ sufficiently large. Since $r^k \to 0$ by Thm. IV.3(i), then $d^k \to 0$. We have

$$0 \leftarrow \frac{r^k - J_k d^k}{\|d^k\|} = \frac{r^k + J_k d^k - R_y(y^k + d^k)}{\|d^k\|} + \frac{R_y(y^k + d^k)}{\|d^k\|}.$$ 

The first summand in the above equation tends to zero because of strict differentiability of $R_y$ at $y*$; therefore

$$R_y(y^k + d^k)/\|d^k\| \to 0.$$ 

By nonsingularity of $J$, then $\|R_y(y)\| \geq \alpha \|y - y_*\|$ for all $y$ sufficiently close to $y_*$ and since $y^k + d^k \to y_*$ we have

$$0 \leftarrow \frac{\|R_y(y^k + d^k)\|}{\|d^k\|} \geq \frac{\alpha \|y^k + d^k - y_*\|}{\|d^k\|} \geq \frac{\alpha \|y + d^k - y_*\| + \|y^k - y_*\|}{\|d^k\|}.$$ 

This implies $\|y^k + d^k - y_*\|/\|y^k - y_*\| \to 0$, i.e., $(d^k)_{k \in \mathbb{N}}$ is superlinearly convergent with respect to $(y^k)_{k \in \mathbb{N}}$. □

VII. Simulations

We now present numerical results obtained with the proposed algorithm. The scripts reproducing the results in this section are available online.\footnote{https://github.com/kul-forbes/NAMA-experiments} In NAMA we used $\beta = 0.5$ and $\tau_{\text{min}} = 10^{-3}$ (see Remark IV.2). Furthermore, in all experiments we computed directions $(d^k)_{k \in \mathbb{N}}$ according to the L-BFGS method, with memory 20, which is able to scale with the problem dimension much better than full quasi-Newton update formulas. All experiments were performed using MATLAB 2016b (v9.1.0) on a MacBook Pro running macOS 10.12, with an Intel Core i5 CPU (2.7 GHz) and 8 GB of memory.

A. Linear MPC

We consider finite horizon, discrete time, linear optimal control problems of the form

$$\text{minimize} \sum_{i=0}^{N-1} \ell_i(x_i, u_i) + \ell_N(x_N) \tag{36a}$$

subject to $x_0 = x_{\text{init}}, x_{i+1} = \Phi_i x_i + \Gamma_i u_i + c_i, \; i = 0, \ldots, N-1, \tag{36b}$

where $x_0, \ldots, x_N \in \mathbb{R}^n_x$ and $u_0, \ldots, u_{N-1} \in \mathbb{R}^n_u$, and

$$\ell_i(x, u) = q_i(x, u) + g_i(L_i(x, u)), \tag{36c}$$

$$\ell_N(x) = q_N(x) + g_N(L_N x), \tag{36d}$$

Here the $q_i$ are strongly convex (typically quadratic), the $g_i$ are proper, closed, convex functions, while the $L_i$ are linear mappings, for $i = 0, \ldots, N$. For example, with a convex set $C$, one can set

$$g_i(\cdot) = \delta_C(\cdot) \quad \text{(hard constraints)}$$

$$g_i(\cdot) = \alpha \text{dist}_C(\cdot), \quad \alpha > 0, \quad \text{(soft constraints)}$$

Set $C$ here is typically the nonpositive orthant or a box, but can be any other convex set onto which one can efficiently project. When $C = [a_1, b_1] \times \cdots \times [a_d, b_d]$ is a $d$-dimensional box, then one can alternatively model soft constraints as

$$g_i(z) = \sum_{j=1}^d \alpha_j |z_j - \max \{a_j, \min \{b_j, z_j\}\}. \tag{37}$$

Problem (36) takes the form (P) by reformulating it as follows (see also [1], [49], [50]). Denote the full sequence of states and inputs as $\bar{x} = (x_0, u_0, x_1, u_1, \ldots, x_N)$, and let

$$S(p) = \{\bar{x} | \bar{x} = \Phi_i x_i + \Gamma_i u_i, x_0 = p\}$$

be the affine subspace of feasible trajectories of the system having initial state $p$. Then in (P)

$$f(\bar{x}) = \sum_{i=0}^{N-1} q_i(x_i, u_i) + g_N(x_N) + \delta_{S(\text{ref})}(\bar{x}),$$

$$g(\bar{z}) = \sum_{i=0}^{N} g_i(z_i), \quad A = \text{diag}(L_0, \ldots, L_N).$$

Let us further denote by $\bar{y} = (y_0, \ldots, y_N)$ the dual variable associated with this problem. In this case, in the alternating minimization step 1 of NAMA, the iterate $\bar{x}^k$ is obtained by solving

$$\text{minimize} \sum_{i=0}^{N-1} q_i(x_i, u_i) + g^k_i(L_i(x_i, u_i)) + q_N(x_N) + (y_N^k, L_N x_N).$$

subject to $x_{i+1} = \Phi_i x_i + \Gamma_i u_i + c_i, \; i = 0, \ldots, N-1, \tag{36e}$

This is an unconstrained LQR problem whose solution can be efficiently computed with a Riccati-like recursion procedure, in the typical case where $q_0, \ldots, q_N$ are quadratic, see [49, Alg.s 3, 4]. The expensive “factor” step only needs to be performed once, before the main loop of the algorithm takes place. At every iteration one needs to perform merely a forward-backward sweep and no matrix inversions are required. Furthermore

$$\bar{x}_i^k = \text{prox}_{\gamma^{-1} g_i}(\gamma^{-1} \bar{y}_i^k + L_i(x_i^k, u_i^k)), \; i = 0, \ldots, N-1,$$

$$\bar{x}_N^k = \text{prox}_{\gamma^{-1} g_N}(\gamma^{-1} \bar{y}_N^k + L_N(x_N^k)),$$

which in the case of hard/soft constraints essentially consist of projections onto the constrained sets.

1) Aircraft control: We applied the proposed method to the AFTI-16 aircraft control problem [50], [51] with $n_x = 4$ states and $n_u = 2$ inputs, for a sampling time $T_s = 0.05$ seconds. The objective is to drive the pitch angle from $0^\circ$ to $10^\circ$, and then back to $0^\circ$. We simulated the system for 4 seconds, at the sampling time $T_s = 0.05$, using $N = 50$ and quadratic costs

$$q_i(x, u) = \frac{1}{2} \|x - x_{\text{ref}}\|_Q^2 + \frac{1}{2} \|u\|_R^2, \; i = 0, \ldots, N-1,$$

$$q_N(x) = \frac{1}{2} \|x - x_{\text{ref}}\|_{Q_N}^2,$$

where $Q = \text{diag}(10^{-4}, 10^2, 10^{-3}, 10^2), \; Q_N = 100 \cdot Q$ and $R = \text{diag}(10^{-2}, 10^{-2})$. The reference was set $x_{\text{ref}} = (0, 0, 0, 10)$ for the first 2 seconds, and $x_{\text{ref}} = (0, 0, 0, 0)$ for the remaining 2 seconds. Furthermore, we imposed hard box constraints on the inputs, and soft box constraints (37) on the states, with weights $10^6$. Since soft constraints can be formulated into a QP, by adding linearly penalized nonnegative slack variables, we also compared against standard QP solvers.

The dual problem has a condition number of $10^9$. To improve the convergence of the algorithms we therefore considered scaling the dual variables according to the Jacobi scaling, which consists of a diagonal change of variable (in the dual space) enforcing the (dual) Hessian to have diagonal elements equal to one (see also [50], [52] on the problem of preconditioning fast dual proximal gradient methods). Note that a diagonal change of variable in the dual space simply corresponds to a scaling of the equality constraints, when the problem is equivalently formulated as (P').
We compared NAMA against fast AMA [53], which is also known as GPAD [49] in this context, qpOASES v3.2.0 [54] and the commercial QP solver MOSEK v7.1. We also compared against the cone solvers ECOS v2.0.4 [55], SDPT3 v4.0 [56] and SeDuMi v1.34 [57], all accessed through CVX v2.1 in MATLAB: note that the CPU time for these methods does not include the problem parsing and preprocessing by CVX, but only considers the actual running time of the solvers. The results of the simulations are reported in Table I. As termination criterion for NAMA and GPAD we used \( \|R_\gamma(y^k)\|_\infty \leq \epsilon_{\text{tol}} = 10^{-4} \). We also report the (average and maximum) number of \( x \)- and \( z \)-minimization steps performed by NAMA: due to the structure of \( f \), the \( x \)-update is a linear mapping, and consequently we can save its computation during the backtracking line-search. GPAD, in contrast, performs one alternating minimization per iteration.

Apparently, NAMA greatly improves the convergence performance with respect to GPAD. When the problem is prescaled, our method performs favorably also with respect to the other QP and cone solvers considered. One must keep in mind that NAMA was executed using a generic, high-level MATLAB implementation. As computation times become smaller and smaller, overheads due to the runtime environment get more and more relevant in the total CPU time. A tailored, low-level implementation of the same algorithm is assumed to be strongly convex, while the other is composed with a linear transformation. The method is an extension of the classical alternating minimization algorithm (AMA), performing an additional line-search step over the alternating minimization envelope associated with the problem. By appropriately selecting the line-search directions, for example according to quasi-Newton methods for solving the optimality conditions \( R_\gamma(y) = 0 \), we have shown that the algorithm converges superlinearly provided that ordinary second-order sufficiency conditions hold for the envelope function at the (unique) dual solution. At the same time, the algorithm possesses the same global sublinear and local linear convergence rates as AMA. Numerical experiments with the proposed method on linear MPC problems suggest that NAMA is able to significantly speed up the convergence of AMA, comparing favorably against its accelerated variant and other state-of-the-art solvers even when limited-memory methods, such as L-BFGS, are used to compute the search directions.

\[ \|R_\gamma(y^k)\|_\infty \leq \epsilon_{\text{tol}} = 10^{-4} \]

\[ \text{Lemma A.1. Let } y, w \in \mathbb{R}^m \text{ and } \gamma > 0. \text{ Then,} \]
\[ \psi(w) \geq \psi_\gamma(y) + \frac{\gamma}{2} \|Ax(y) - z_\gamma(y)\|^2 + \langle z_\gamma(y) - Ax(y), w - y \rangle. \]  

\[ \text{Proof. By (1) we have} \]
\[ f(x(y)) + f^*(-A^TW) \geq -\langle Ax(y), w \rangle, \]
\[ g(z_\gamma(y)) + g^*(w) \geq \langle z_\gamma(y), w \rangle. \]

By summing the two inequalities and using the definition of \( \psi_\gamma \), after manipulations one obtains the result. \( \square \)

\[ \text{Lemma A.2. For all } y \in \mathbb{R}^m \text{ it holds} \]
\[ \frac{\epsilon_\gamma^2}{2} \|x(y) - x_\gamma\|^2 \leq \psi(y) - \inf \psi. \]

\[ \text{Proof. From the optimality condition of the problem defining } x(y), \text{ one obtains } -A^Ty \in \partial f(x(y)). \text{ Then, by strong convexity of } f \text{ one gets } \]
\[ f(x(y)) - \langle A^Ty, x(y) - x_\gamma(y) \rangle + \frac{\epsilon_\gamma^2}{2} \|x(y) - x_\gamma\|^2 \leq f(x_\gamma). \]

By using (17a) in the above inequality we obtain
\[ \frac{\epsilon_\gamma^2}{2} \|x(y) - x_\gamma\|^2 - \langle Ax_\gamma, y \rangle \leq f(x_\gamma) + f^*(-A^Ty), \]
By using (1) on g we have instead
\[ (Ax_*, y) \leq g(Ax_*) + g^*(y). \]

By summing the last two inequalities one obtains
\[ \frac{\|x\|}{\|x\|} \|x - x_*\|^2 \leq f(x_*) + g(Ax_*) + \psi(y), \]
and the claimed bound follows by strong duality.

**Lemma A.3.** Suppose that the following hold for (P):

(i) \( \text{ri}(\text{dom} f) \cap \text{ri}(\text{dom} g) \neq \emptyset \) (strict feasibility);
(ii) \( 0 \in \text{ri} \partial(f + g \circ A)(x_*) \) (strict complementarity).

Then for any compact set \( U \) there is \( \kappa > 0 \) such that
\[ \text{dist}(y, Y_*) \leq \kappa [\text{dist}(A^\top y, \partial f(x_*)) + \text{dist}(y, \partial g(Ax_*))] \]
holds for all \( y \in U \).

**Proof.** From Lem. A.3(ii) it follows that
\[ 0 \in \text{ri} \left[ \partial f(x_*) + A^\top \partial g(Ax_*) \right] = \text{ri} \partial f(x_*) + A^\top \text{ri} \partial g(Ax_*). \]
In fact, the first inclusion is due to \[17\], Thm. 23.9 in light of Lem. A.3(ii), and the equality is due to \[17\], Thm. 6.6).
Consider \( W = \{ w \mid -A^\top w \in \partial f(x_*) \} \subseteq \mathbb{R}^m \). From (5),
\[ Y_* = W \cap \partial g(Ax_*). \]
Furthermore, using (40) we obtain
\[ \emptyset \neq \{ w \mid -A^\top w \in \text{ri} \partial f(x_*) \} = \text{ri} W, \]
where the equality is due to \[17\], Thm. 6.7, and the fact that \( \text{ri} W \cap \text{ri} \partial g(Ax_*) \neq 0 \). By \[58\], Cor. 5 then, we conclude that \( W \) and \( \partial g(Ax_*) \) are boundedly linearly regular: for any compact set \( U \) there is \( \alpha > 0 \) such that for all \( y \in U \)
\[ \text{dist}(y, Y_*) \leq \alpha [\text{dist}(y, W) + \text{dist}(y, \partial g(Ax_*))]. \]

Similarly, (40) implies with \[58\], Cor. 5 that the sets \( L = \{ (w, -A^\top w) \mid w \in \mathbb{R}^m \} \) and \( M = \mathbb{R}^m \times \partial f(x_*) \) are boundedly linearly regular. Observe that
\[ L \cap M = \{ (w, -A^\top w) \mid -A^\top w \in \partial f(x_*) \}. \]
Therefore, there is \( \beta > 0 \) such that for all \( y \in U \)
\[ \text{dist}(y, W) \leq \text{dist}(y, -A^\top y, L \cap M) \]
\[ \leq \beta [\text{dist}(y, -A^\top y, L) + \text{dist}(y, -A^\top y, M)] \]
\[ = \beta \text{dist}(y, A^\top y, \partial f(x_*)), \]
where the second inequality is due to bounded linear regularity of \( L \) and \( M \), while the equality holds since \( (y, -A^\top y) \in L \) and \( \text{dist}(y, -A^\top y, M) = \text{dist}(-A^\top y, \partial f(x_*)) \) for any \( y \). Using the above inequality in (41) yields the result.

**Lemma A.4** (Twice differentiability of \( f^* \)). Suppose that \( f \) satisfies Assumption 2(ii) for the primal-dual solution \((x_*, y_*)\). Then \( f^* \) is of class \( C^2 \) around \( y_* \), with
\[ \nabla^2 f^*(y_*) = [d^2 f(x_*)]^\top. \]

**Proof.** From \[20\], Thm. 13.21 we know that \( f^* \) is twice epi-differentiable at \( v \) for \( x \in \partial f^*(v) \) iff \( f \) is twice epi-differentiable at \( x \) for \( v \), with the relation
\[ d^2 f^*(v|x) = [d^2 f(x_*)]^\top. \]

The cited proof trivially extends to strict twice differentiability, and in fact \( f^* \) turns out to be strictly twice epi-differentiable at \( x_* \). Since \( \text{range}(H_f) + S_f^\perp = \mathbb{R}^n \), by applying (42) to (27) and conjugating \( d^2 f(x_*) \) with means of \[18\], Prop. E.3.2.1 we obtain that function \( f^* \) has purely quadratic second epi-derivative (as opposed to generalized quadratic)
\[ d^2 f^*(−A^\top y_*, x_*) = \langle (\Pi_{S_f^\perp} H_f \Pi_{S_f^\perp}) y, w \rangle, \quad w \rangle = \langle H_f^\top, w \rangle \]
which is everywhere finite in particular. The proof now follows from \[25\], Cor. 4.7.

With similar reasonings, the following result easily follows.

**Lemma A.5** (Twice epi-differentiability of \( g^* \)). Suppose that \( g \) (strictly) satisfies Assumption 2(ii) for a primal-dual solution \((x_*, y_*)\). Then \( g^* \) is (strictly) twice epi-differentiable at \( y_* \) for \( A_{x_*} \). More precisely, letting \( \delta = S_f^\perp + \text{range}(H_g) \),
\[ d^2 g^*(y_*, A_{x_*}) = [d^2 g(A_{x_*}|y_*)]^\top = \langle H_g^\top, \cdot \rangle + \delta S. \]

**Acknowledgment**

The authors would like thank Dmitriy Drusvyatskiy for his contribution to the proof of Lemma A.3 in the Appendix.

**References**

[1] G. Stathopoulos, H. Shukla, A. Szucs, Y. Pu, and C. N. Jones, “Operator splitting methods in control,” Foundations and Trends in Systems and Control, vol. 3, no. 3, pp. 249–362, 2016.

[2] M. Fazel, T. K. Pong, D. Sun, and P. Tseng, “Hankel matrix rank minimization with applications to system identification and realization,” SIAM Journal on Matrix Analysis and Applications, vol. 34, no. 3, pp. 946–977, 2013.
Figure 2: Oscillating masses, average and maximum CPU time (in seconds) for increasing prediction horizon and 50 randomly selected initial states. First column: $K = 8$ actuators. Second column: $K = 16$ actuators. Fast AMA and NAMA were stopped as soon as $\|R_\lambda(y^k)\|_\infty \leq \epsilon_{tol} = 10^{-4}$.

[3] 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, p. 1–122, 2011.
[4] N. Parikh and S. Boyd, “Proximal algorithms,” Foundations and Trends in Optimization, vol. 1, no. 3, pp. 127–239, 2014.
[5] P. Tseng, “Applications of a splitting algorithm to decomposition in convex programming and variational inequalities,” SIAM Journal on Control and Optimization, vol. 29, no. 1, pp. 119–138, 1991.
[6] P.-L. Lions and B. Mercier, “Splitting algorithms for the sum of two non-linear operators,” SIAM Journal on Numerical Analysis, vol. 16, no. 6, pp. 964–979, 1979.
[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] Y. Nesterov, “Gradient methods for minimizing composite functions,” Mathematical Programming, vol. 140, no. 1, pp. 125–161, 2013.
[9] A. Beck and M. Teboulle, “A fast dual proximal gradient algorithm for convex minimization and applications,” Operations Research Letters, vol. 42, no. 1, pp. 1–6, 2014.
[10] P. Patrinos and A. Bemporad, “Proximal Newton methods for convex composite optimization,” in IEEE Conference on Decision and Control, 2013, pp. 2358–2363.
[11] L. Stella, A. Themelis, and P. Patrinos, “Forward-backward quasi-Newton methods for nonsmooth optimization problems,” Computational Optimization and Applications, vol. 67, no. 3, pp. 443–487, 2017.
[12] A. Themelis, L. Stella, and P. Patrinos, “Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone line-search algorithms,” arXiv preprint arXiv:1606.06256, 2016.
[13] T. Liu and T. K. Pong, “Further properties of the forward-backward envelope with applications to difference-of-convex programming,” Computational Optimization and Applications, vol. 67, no. 3, pp. 489–520, 2017.
[14] A. K. Sampathirao, P. Sopasakis, A. Bemporad, and P. Patrinos, “Proximal limited-memory quasi-Newton methods for scenario-based stochastic optimal control,” To appear in Proceedings of the 20th IFAC Congress, 2017.
[15] P. Patrinos, L. Stella, and A. Bemporad, “Douglas-Rachford splitting: Complexity estimates and accelerated variants,” in 53rd IEEE Conference on Decision and Control, 2014, pp. 4234–4239.
[16] A. Themelis, L. Stella, and P. Patrinos, “Douglas–Rachford splitting and ADMM for nonconvex optimization: new convergence results and accelerated versions,” arXiv preprint arXiv:1709.05747, 2017.
[17] R. T. Rockafellar, Convex Analysis. Princeton university press, 1997.
[18] J.-B. Hiriart-Urruty and C. Lemaréchal, Fundamentals of Convex Analysis. Springer Science & Business Media, 2001.
[19] H. H. Bauschke and P. L. Combettes, Convex analysis and monotone operator theory in Hilbert spaces. Springer, 2011.
[20] R. T. Rockafellar and R. J.-B. Wets, Variational analysis. Springer, 2011.
[21] R. T. Rockafellar, “First- and second-order epi-differentiability in nonlinear programming,” Transactions of the American Mathematical Society, vol. 307, no. 1, pp. 75–108, 1988.
[22] ——, “Second-order optimality conditions in nonlinear programming obtained by way of epi-derivatives,” Mathematics of Operations Research, vol. 14, no. 3, pp. 462–484, 1989.
[23] R. A. Poliquin and R. T. Rockafellar, “Amenable functions in optimization,” Nonsmooth optimization: methods and applications (Erice, 1991), pp. 338–353, 1992.
[24] ——, “Second-order nonsmooth analysis in nonlinear programming,” Recent advances in nonsmooth optimization, pp. 322–349, 1995.
[25] ——, “Generalized Hessian properties of regularized nonsmooth functions,” SIAM Journal on Optimization, vol. 6, no. 4, pp. 1121–1137, 1996.
[26] A. Auslender and M. Teboulle, Asymptotic cones and functions in optimization and variational inequalities. Springer, 2003.
