CONVERGENCE RATES OF INERTIAL SPLITTING SCHEMES FOR NONCONVEX COMPOSITE OPTIMIZATION

Patrick R. Johnstone and Pierre Moulin

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Email: prjohns2@illinois.edu, moulin@ifp.uiuc.edu

ABSTRACT

We study the convergence properties of a general inertial first-order proximal splitting algorithm for solving nonconvex nonsmooth optimization problems. Using the Kurdyka–Łojaziewicz (KL) inequality we establish new convergence rates which apply to several inertial algorithms in the literature. Our basic assumption is that the objective function is semialgebraic, which lends our results broad applicability in the fields of signal processing and machine learning. The convergence rates depend on the exponent of the "desingularizing function" arising in the KL inequality. Depending on this exponent, convergence may be finite, linear, or sublinear and of the form $O(k^{-p})$ for $p > 1$.

***Index Terms*—** Kurdyka–Łojaziewicz Inequality, Inertial forward-backward splitting, heavy-ball method, convergence rate, first-order methods.

1. INTRODUCTION

We are interested in solving the following optimization problem

$$\min_{x\in\mathbb{R}^n} \Phi(x) = f(x) + g(x) \tag{1}$$

where $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous (l.s.c.) and $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable with Lipschitz continuous gradient. We also assume that Φ is *semialgebraic* [1], meaning there are integers $p, q \geq 0$ and polynomial functions $P_{ij}, Q_{ij} : \mathbb{R}^{n+1} \to \mathbb{R}$ such that

$$\{(x, y) : y \geq f(x)\} = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \{z \in \mathbb{R}^{n+1} : P_{ij}(z) = 0, Q_{ij}(z) < 0\}.$$

We make no assumption of convexity. Semialgebraic objective functions in the form of (1) are widespread in machine learning, image processing, compressed sensing, matrix completion, and computer vision [2, 3, 4, 5, 6]. In this paper we focus on the application of Prob. (1) to *sparse least-squares*, which includes compressed sensing and regression. This problem arises when looking for a sparse solution to a set of underdetermined linear equations. Suppose we observe $y = Ax + b$ where b is noise and wish to recover x which is known to be sparse, however the matrix A is "fat" or poorly conditioned. One approach is to solve (1) with f a loss function modeling the noise b and g a regularizer modeling prior knowledge of x, in this case sparsity. The correct choice for f will depend on the noise model. A common choice is the least-squares function $\frac{1}{2}\|Ax - b\|_2^2$ which is convex, smooth, and semialgebraic. Examples of appropriate nonconvex semialgebraic choices for g are the ℓ_0 pseudo-norm, and the smoothly clipped absolute deviation (SCAD) [7]. The prevailing convex choice is the ℓ_1 norm which is also semialgebraic. SCAD has the advantage over the ℓ_1-norm that it leads to nearly unbiased estimates of large coefficients. Furthermore unlike the ℓ_0 norm SCAD leads to a solution which is continuous in the data matrix A [7]. Nevertheless ℓ_1-based methods continue to be the standard throughout the literature due to convexity and computational simplicity.

For Problem (1), *first-order methods* have been found to be computationally inexpensive, simple, and effective solvers [8]. In this paper we are interested in first order methods of the *inertial* type, also known as *momentum* methods. These methods generate the next iterate using more than one previous iterate so as to mimic the inertial dynamics of a model differential equation. In many instances both in theory and in practice, inertial methods have been shown to converge faster than noninertial ones [9, 10, 11]. Furthermore for nonconvex problems it has been observed that using inertia can help the algorithm escape local minima and saddle points that would capture other first-order algorithms [12, 13, Sec 4.1]. A prominent example of the use of inertia in nonconvex optimization is in training neural networks, which goes under the name of *back propagation with momentum* [14]. In convex optimization a prominent example is the heavy ball method [9].

Over the past decade the KL inequality has come to prominence in the optimization community as a powerful tool for studying both convex and nonconvex problems. It is very general, applicable to almost all problems encountered in real applications, and powerful because it allows researchers to precisely understand the local convergence properties of first-order methods. The inequality goes back to [15, 16]. In [17, 18, 19, 20] the KL inequality was used to derive convergence rates of descent-type first order methods. The KL inequality was used to study convex optimization problems in [21, 22].

Nonconvex optimization has traditionally been challenging for researchers to study since iterative methods generally cannot distinguish a local minimum from a global minimum. Nevertheless, for some applications such as empirical risk minimization in machine learning, finding a good local minimum is all that is required of the optimization solver [23, Sec. 3]. In other problems local minima have been shown to be global minima [24].

Contributions: The main contribution of this paper is to determine for the first time the local convergence rate of a broad family of inertial proximal splitting methods for solving Prob. (1). The family of methods we study includes several algorithms proposed in the literature for which convergence rates are unknown. The family was proposed in [10], where it was proved that the iterates converge to a critical point. However the *convergence rate*, e.g. how fast the iterates converge, was not determined. In fact in [10], local linear convergence was shown under a partial smoothness assumption. In contrast we do not assume partial smoothness and our results are far more general. We use the KL inequality and show finite, linear, or sublinear convergence, depending on the KL exponent (see Sec. 2). The main inspiration for our work is [19] which studied convergence rates of several *noninertial* schemes using the KL property. However, the analysis of [19] cannot be applied to inertial methods. Our approach is to extend the framework of [19] to the inertial setting. This is done by proving convergence rates of a multistep Lyapunov potential function which upper bounds the objective function. We also include experiments to illustrate the derived convergence rates.

Notation: Given a closed set C and point x, define $d(x,C) \triangleq \min\{\|x-c\| : c \in C\}$. For a sequence $\{x_k\}_{k\in\mathbb{N}}$ let $\Delta_k \triangleq \|x_k - x_{k-1}\|$. We say that $x_k \to x^*$ linearly with convergence factor $q \in (0,1)$ if there exists $C > 0$ such that $\|x_k - x^*\| \leq Cq^k$.

2. MATHEMATICAL BACKGROUND

In this section we give an overview of the relevant mathematical concepts. We use the notion of the limiting subdifferential $\partial\Phi(x)$ of a l.s.c. function Φ. For the definition and properties we refer to [1, Sec 2.1]. A necessary (but not sufficient) condition for x to be a minimizer of Φ is $0 \in \partial\Phi(x)$. The set of critical points of Φ is $\text{crit}(\Phi) \triangleq \{x : 0 \in \partial\Phi(x)\}$. A useful notion is the *proximal operator* w.r.t. a l.s.c. proper function g, defined in [8, Def. 10.1]. Note that, unlike the convex case, the prox operator is not necessarily single-valued.

Definition A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to have the Kurdyka-Lojasiewicz (KL) property at $x^* \in \text{dom}\,\partial f$ if there exists $\eta \in (0,+\infty]$, a neighborhood U of x^*, and a differentiable and concave function $\varphi : [0,\eta) \to \mathbb{R}_+$ such that

(i) $\varphi(0) = 0$, and for all $s \in (0,\eta)$, and $\varphi'(s) > 0$.

(ii) For all $x \in U \cap \{x : f(x^*) < f(x) < f(x^*) + \eta\}$ the KL inequality holds:

$$\varphi'(f(x) - f(x^*))d(0, \partial f(x)) \geq 1. \tag{2}$$

If f is semialgebraic, then it has the KL property at all points in $\text{dom}\,\partial f$, and $\varphi(t) = \frac{c}{\theta}t^\theta$ for $\theta \in (0,1]$.

In the semialgebraic case we will refer to θ as the *KL exponent* (note that some other papers use $1-\theta$ [22]). For the special case where f is smooth, (2) can be rewritten as $\|\nabla(\varphi \circ (f(x) - f(x^*))\| \geq 1$, which shows why φ is called a "desingularizing function". The slope of φ near the origin encodes information about the "flatness" of the function about a point, thus the KL exponent provides a way to quantify convergence rates of iterative first-order methods.

For example the 1D function $f(x) = |x|^p$ for $p \geq 2$ has desingluarizing function $\varphi(t) = t^{\frac{1}{p}}$. The larger p, the flatter f is around the origin, and the slower gradient-based methods will converge. In general, functions with smaller exponent θ have slower convergence near a critical point [19]. Thus, determining the KL exponent of an objective function holds the key to assessing convergence rates near critical points. Note that for most prominent optimization problems, determining the KL exponent is an open problem. Nevertheless many important examples have been determined recently, such as least-squares and logistic regression with an ℓ_1, ℓ_0, or SCAD penalty [22]. A very interesting recent work showed that for convex functions the KL property is equivalent to an error bound condition which is often easier to check in practice [21]. We now precisely state our assumptions on Problem (1), which will be in effect throughout the rest of the paper.

Assumption 1. The function $\Phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is semialgebraic, bounded from below, and has desingularizing function $\varphi(t) = \frac{c}{\theta}t^\theta$ where $c > 0$ and $\theta \in (0,1]$. The function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is l.s.c., and $f : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz continuous gradient with constant L.

3. A FAMILY OF INERTIAL ALGORITHMS

We study the family of inertial algorithms proposed in [10]. In what follows $s \geq 1$ is an integer, and $I = \{0, 1, \ldots, s-1\}$.

Algorithm 1 Multi-step Inertial Forward-Backward splitting (MiFB)

Require: $x_0 \in \mathbb{R}^n$, $0 < \underline{\gamma} \leq \overline{\gamma} < 1/L$.
  Set $x_{-s} = \ldots = x_{-1} = x_0$, $k = 1$
  repeat
    Choose $0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 1/L$, $\{a_{k,0}, a_{k,1}, \ldots\} \in (-1,1]^s$, $\{b_{k,0}, b_{k,1}, \ldots\} \in (-1,1]^s$.
    $y_{a,k} = x_k + \sum_{i\in I} a_{k,i}(x_{k-i} - x_{k-i-1})$
    $y_{b,k} = x_k + \sum_{i\in I} b_{k,i}(x_{k-i} - x_{k-i-1})$
    $x_{k+1} \in \text{prox}_{\gamma_k g}\left(y_{a,k} - \gamma_k \nabla f(y_{b,k})\right)$
    $k = k+1$
  until convergence

Note the algorithm as stated leaves open the choice of the parameters $\{a_{k,i}, b_{k,i}, \gamma_k\}$. For convergence conditions on the parameters we refer to Section 4 and [11, Thm. 1].

The algorithm is very general and covers several inertial algorithms proposed in the literature as special cases. For instance the inertial forward-backward method proposed in [12] corresponds to MiFB with $s = 1$, and $b_{k,0} = 0$. The well-known iPiano algorithm also corresponds to this same parameter choice, however the original analysis of this algorithm assumed g was convex [13]. The *heavy-ball method* is an early and prominent inertial first-order method which also corresponds to this parameter choice when $g = 0$. The heavy-ball method was originally proposed for strongly convex quadratic problems but was considered in the context of nonconvex problems in [25]. The analysis of [26] applies to MiFB for the special case when $s = 1$ and $a_{k,0} = b_{k,0}$. However [26] only derived convergence rates of the iterates and not the function values, which are our main interest. Note that if the objective function is not Lipschitz continuous, rates derived for the iterates do not immediately imply rates for the objective. Furthermore [26] used a different proof technique to the one used here. This same parameter choice has been considered for convex optimization in [11, 27], albeit without the sharp convergence rates derived here.

General convergence rates have not been derived for MiFB under nonconvexity and semialgebraicity assumptions. The convergence rate of iPiano has been examined in a limited situation where the KL exponent $\theta = 1/2$ in [22, Thm 5.2]. Note that the primary motivation for studying this framework is its generality - allowing our analysis to cover many special cases from the literature. However the case $s = 1$ is the most interesting in practice and corresponds to the most prominent inertial algorithms.

4. CONVERGENCE RATE ANALYSIS

Throughout the analysis, Assumption 1 is in effect. Before providing our convergence rate analysis, we need a few results from [10].

Theorem 1. *Fix $s \geq 1$ and recall $I = \{0, 1, \ldots, s-1\}$. Fix $\{\gamma_k\}$, $\{a_{k,i}\}$ and $\{b_{k,i}\}$ for $k \in \mathbb{N}$ and $i \in I$. Fix $\mu, \nu > 0$ and define*

$$\beta_k \triangleq \frac{1 - \gamma_k L - \mu - \nu\gamma_k}{2\gamma_k}, \quad \underline{\beta} \triangleq \liminf_{k\in\mathbb{N}} \beta_k,$$
$$\alpha_{k,i} \triangleq \frac{s a_{k,i}^2}{2\gamma_k \mu} + \frac{s b_{k,i}^2 L^2}{2\nu}, \quad \overline{\alpha}_i \triangleq \limsup_{k\in\mathbb{N}} \alpha_{k,i},$$

and $z_k \triangleq (x_k^\top, x_{k-1}^\top, \ldots, x_{k-s}^\top)^\top$. Recall $\Delta_k \triangleq \|x_k - x_{k-1}\|$. Define the multi-step Lyapunov function as

$$\Psi(z_k) \triangleq \Phi(x_k) + \sum_{i\in I}\left(\sum_{j=i}^{s-1}\overline{\alpha}_j\right)\Delta_{k-i}^2. \tag{3}$$

and

$$\delta \triangleq \underline{\beta} - \sum_{i\in I}\overline{\alpha}_i > 0. \tag{4}$$

If the parameters are chosen so that $\delta > 0$ then

(i) for all k, $\Psi(z_{k+1}) \leq \Psi(z_k) - \delta\Delta_{k+1}^2$,

(ii) for all k, there is a $\sigma > 0$ such that $d(0, \partial\Psi(z_k)) \leq \sigma\sum_{j=k+1-s}^{k}\Delta_j$,

(iii) If $\{x_k\}$ is bounded there exists $x^* \in crit(\Phi)$ such that $x_k \to x^*$ and $\Phi(x_k) \to \Phi(x^*)$.

Proof. Statements (i) and (ii) are shown in [10, Lemma A.5] and [10, Fact (R.2)] respectively. The fact that $\Phi(x_k) \to \Phi(x^*)$ is shown in [10, Lemma A.6]. The fact that $x_k \to x^*$ is the main result of [10, Thm 2.2]. ■

The assumption that $\{x_k\}$ is bounded is standard in the analysis of algorithms for nonconvex optimization and is guaranteed under ordinary conditions such as coercivity. Since the set of semialgebraic functions is closed under addition, Ψ is semialgebraic [28]. We now give our convergence result.

Theorem 2. *Assume $\{x_k\}$ is bounded and the parameters of MiFB are chosen such that $\delta > 0$ where δ is defined in (4). Thus there exists a critical point x^* such that $x_k \to x^*$. Let θ be the KL exponent of Ψ defined in (3).*

(a) If $\theta = 1$, then x_k converges to x^* in a finite number of iterations.

(b) If $\frac{1}{2} \leq \theta < 1$, then $\Phi(x_k) \to \Phi(x^*)$ linearly.

(c) If $0 < \theta < 1/2$, then $\Phi(x_k) - \Phi(x^*) = O\left(k^{\frac{1}{2\theta-1}}\right)$.

Proof. The starting point is the KL inequality applied to the multi-step Lyapunov function defined in (3). Let $z^* \triangleq ((x^*)^\top, \ldots, (x^*)^\top)^\top$. Suppose $\Psi(z_K) = \Psi(z^*)$ for some $K > 0$. Then the descent property of Thm. 1(i), along with the fact that $\Psi(z_k) \to \Psi(z^*)$, implies that $\Delta_{K+1} = 0$ and therefore $\Psi(z_k) = \Psi(z^*)$ holds for all $k > K$. Therefore assume $\Psi(z_k) > \Psi(z^*)$. Now since $z_k \to z^*$ and $\Psi(z_k) \to \Psi(z^*)$, there exists $k_0 > 0$ such that for $k > k_0$ (2) holds with $f = \Psi$. Assume $k > k_0$. Squaring both sides of (2) yields

$$\varphi'^2(\Psi(z_k) - \Psi(z^*))d(0, \partial\Psi(z_k))^2 \geq 1, \tag{5}$$

Now substituting Thm.1 (ii) into (5) yields

$$\sigma^2\varphi'^2(\Psi(z_k) - \Psi(z^*))\left(\sum_{j=k+1-s}^{k}\Delta_j\right)^2 \geq 1. \tag{6}$$

Now

$$\begin{aligned}\left(\sum_{j=k+1-s}^{k}\Delta_j\right)^2 &\leq s\sum_{j=k+1-s}^{k}\Delta_j^2\\ &\leq \frac{s}{\delta}\sum_{j=k+1-s}^{k}\left(\Psi(z_{j-1}) - \Psi(z_j)\right)\\ &= \frac{s}{\delta}\left(\Psi(z_{k-s}) - \Psi(z_k)\right),\end{aligned}$$

where in the first inequality we have used the fact that $(\sum_{i=1}^s a_i)^2 \leq s\sum_{i=1}^n a_i^2$, and in the second inequality we have used Thm. 1(i). Substituting this into (6) yields

$$\frac{\sigma^2 s}{\delta}\varphi'^2(\Psi(z_k) - \Psi(z^*))\left(\Psi(z_{k-s}) - \Psi(z_k)\right) \geq 1,$$

from which convergence rates can be derived by extending the arguments in [19, Thm 4].

Proceeding, let $r_k \triangleq \Psi(z_k) - \Psi(z^*)$, and $C_1 \triangleq \frac{\delta}{\sigma^2 c^2 s}$, then using $\varphi'(t) = ct^{\theta-1}$, we get

$$r_{k-s} - r_k \geq C_1 r_k^{2(1-\theta)}. \tag{7}$$

If $\theta = 1$, then the recursion becomes $r_{k-s} - r_k \geq C_1, \quad \forall k > k_0$. Since by Theorem 1 (iii), r_k converges, this would require $C_1 = 0$, which is a contradiction. Therefore there exists k_1 such that $r_k = 0$ for all $k > k_1$.

Suppose $\theta \geq 1/2$, then since $r_k \to 0$, there exists k_2 such that for all $k > k_2$, $r_k \leq 1$, and $r_k^{2(1-\theta)} \geq r_k$. Therefore for all $k > k_2$,

$$\begin{aligned}r_{k-s} - r_k \geq C_1 r_k \implies r_k &\leq (1 + C_1)^{-1} r_{k-s}\\ &\leq (1 + C_1)^{-p_1} r_{k_2},\end{aligned} \tag{8}$$

where $p_1 \triangleq \lfloor\frac{k-k_2}{s}\rfloor$ Note that $p_1 > \frac{k-k_2-s}{s}$. Therefore $r_k \to 0$ linearly. Note that if $\theta = \frac{1}{2}$, $2(1-\theta) = 1$ and (8) holds for all $k \geq k_0$.

Finally suppose $\theta < 1/2$. Define $\phi(t) \triangleq \frac{D}{1-2\theta}t^{2\theta-1}$ where $D > 0$, so $\phi'(t) = -Dt^{2\theta-2}$. Now

$$\phi(r_k) - \phi(r_{k-s}) = \int_{r_{k-s}}^{r_k}\phi'(t)dt = D\int_{r_k}^{r_{k-s}}t^{2\theta-2}dt.$$

Therefore since $r_{k-s} \geq r_k$ and $t^{2\theta-2}$ is nonincreasing,

$$\phi(r_k) - \phi(r_{k-s}) \geq D(r_{k-s} - r_k)r_{k-s}^{2\theta-2}.$$

Now we consider two cases.

Case 1: suppose $2r_{k-s}^{2\theta-2} \geq r_k^{2\theta-2}$, then

$$\phi(r_k) - \phi(r_{k-s}) \geq \frac{D}{2}(r_{k-s} - r_k)r_k^{2\theta-2} \geq \frac{C_1 D}{2}. \tag{9}$$

where in the second inequality we have used (7).

Case 2: suppose that $2r_{k-s}^{2\theta-2} < r_k^{2\theta-2}$. Now $2\theta - 2 < 2\theta - 1 < 0$, therefore $(2\theta-1)/(2\theta-2) > 0$, thus $r_k^{2\theta-1} > qr_{k-s}^{2\theta-1}$ where $q = 2^{\frac{2\theta-1}{2\theta-2}} > 1$. Thus

$$\begin{aligned}\phi(r_k) - \phi(r_{k-s}) &= \frac{D}{1-2\theta}\left(r_k^{2\theta-1} - r_{k-s}^{2\theta-1}\right)\\ &> \frac{D}{1-2\theta}(q-1)r_{k-s}^{2\theta-1}\end{aligned}$$

$$\geq \quad \frac{D}{1-2\theta}(q-1)r_{k_0}^{2\theta-1} \triangleq C_2. \quad (10)$$

Thus putting together (9) and (10) yields $\phi(r_k) \geq \phi(r_{k-s}) + C_3$ where $C_3 = \max(C_2, \frac{C_1 D}{2})$. Therefore

$$\phi(r_k) \quad \geq \quad \phi(r_k) - \phi(r_{k-p_2 s}) \geq p_2 C_3$$

where $p_2 \triangleq \lfloor \frac{k-k_0}{s} \rfloor$. Therefore

$$r_k \leq \left(\frac{1-2\theta}{D}\right)^{\frac{1}{2\theta-1}} (p_2 C_3)^{\frac{1}{2\theta-1}} \leq C_4 \left(\frac{k-s-k_0}{s}\right)^{\frac{1}{2\theta-1}}.$$

where $C_4 = \left(\frac{C_3(1-2\theta)}{D}\right)^{\frac{1}{2\theta-1}}$. To end the proof, note that $\Phi(x_k) \leq \Psi(z_k)$. ■

In the case where f and g are also convex, we can use parameter choices specified in [11, Thm. 1]. In the long version of this paper we also prove corresponding convergence rates for the iterates [29, Thm. 4]. Furthermore we prove that if Φ has KL exponent $\theta \in (0, 1/2]$, then the Lyapunov function Ψ has the same KL exponent [29, Thm. 5].

5. NUMERICAL RESULTS

5.1. One Dimensional Polynomial

This simple experiment verifies the convergence rates derived in Theorem 2 for MiFB. Consider the one dimensional function $f(x) = |x|^p$ for $p > 2$. Use $g(x) = +\infty$ if $|x| > 1$ and 0 otherwise. The proximal operator is simple projection and f is $p(p-1)$-smooth on this set. The function $\Phi = f + g$ is semialgebraic with $\varphi(t) = pt^{1/p}$, i.e. $\theta = 1/p$. Therefore Theorem 2 predicts $O\left(k^{-\frac{p}{p-2}}\right)$ rates for MiFB, which is verified in Fig. 1 for three parameter choices in the cases $p = 4, 18$. For simplicity we ignore constants and focus on the sublinear order. For $p \leq 4$ this convergence rate is better than that of Nesterov's accelerated method [30], for which only $O(1/k^2)$ worst-case rate is known. Faster rates are achievable due to the additional knowledge of the KL exponent.

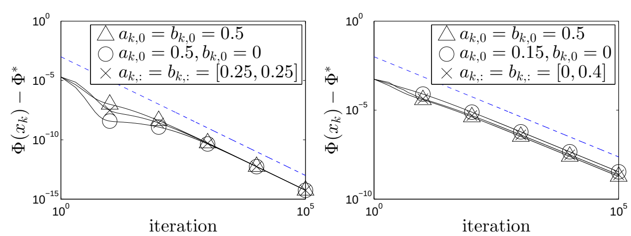

Fig. 1: (Left) $p = 4$, (Right) $p = 18$, $\Phi^* = 0$. The dotted line is the slope of the predicted $O\left(k^{-\frac{p}{p-2}}\right)$ rate (i.e. ignoring constants). Note $a_{k,:} \triangleq [a_{k,0}, a_{k,1}]$ and these are log-log plots.

5.2. SCAD and ℓ_1 regularized Least-Squares

We solve Prob. (1) with $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ and $g(x) = \sum_{i=1}^n r(x_i)$ where r is: 1) the SCAD regularizer defined in [7, Sec. 2.1] and 2) the absolute value $r(x_i) = \lambda|x_i|$ leading to the ℓ_1-norm. In both cases the proximal operator w.r.t. r is easily computed as in Eq. (2.8) and (2.6) of [7]. It was shown in [22, Sec. 5.2] and [21, Lemma 10] that both of these objective functions are KL functions with exponent $\theta = 1/2$.

We choose $A \in \mathbb{R}^{500\times1000}$ having i.i.d. $\mathcal{N}(0, 10^{-4})$ entries, and $b = Ax_0$, where $x_0 \in \mathbb{R}^{1000}$ has 50 nonzero $\mathcal{N}(0, 1)$-distributed entries. For SCAD we use $a = 5$ and $\lambda = 0.08$ and for the ℓ_1 norm we use $\lambda = 0.01$.

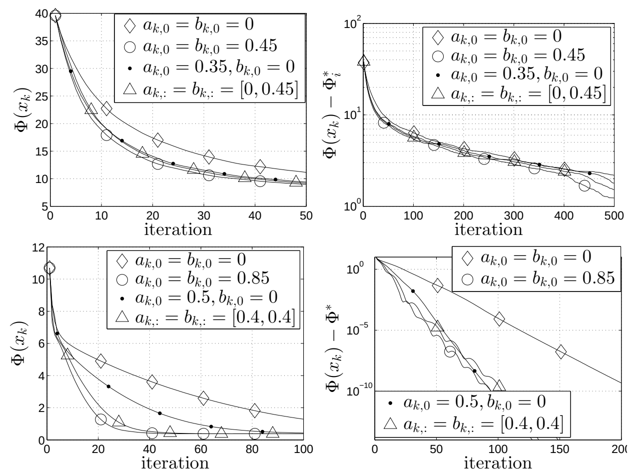

Fig. 2: (Top Left) Plot of $\Phi(x_k)$ for SCAD least-squares. (Top Right) Plot of $\Phi(x_k) - \Phi_i^*$ with a logarithmic y-axis for SCAD least-squares. As SCAD least-squares is a nonconvex problem, each of the four considered parameter choices may converge to a different objective function value Φ_i^* for $i = 1, 2, 3, 4$. (Bottom Left) Plot of $\Phi(x_k)$ for ℓ_1 least-squares. (Bottom Right) Plot of $\Phi(x_k) - \Phi^*$ with a logarithmic y-axis for ℓ_1 least-squares.

We consider four valid parameter choices. To isolate the effect of inertia, all choices used the same randomly chosen starting point and fixed stepsize, $\gamma_k = 0.1/L$ for SCAD and $\gamma_k = 1/L$ for ℓ_1. The inertial parameters were chosen so that $\delta > 0$ (defined in (4)) for SCAD and to satisfy [11, Thm. 1] for the ℓ_1 problem. The two figures on the right corroborate Theorem 2 in that all considered parameter choices converge linearly to their limit, which was estimated by using the attained objective function value after 1000 iterations. For the nonconvex SCAD this is a new result. For ℓ_1-regularized least squares, inertial methods have been shown to achieve *local* linear convergence in [11, 31] under additional strict complementarity or restricted strong convexity assumptions. However, our analysis, which is based on the KL inequality, does not explicitly require these additional assumptions, as the objective function always has a KL exponent of $1/2$ [21, Lemma 10]. Furthermore our result proves *global* linear convergence, in that the KL inequality (2) holds for all k, implying $k_0 = 1$ in (5) and (8) holds for all k. In addition the two left figures show that the inertial choices appear to provide acceleration relative to the standard non-inertial choice which for SCAD is a new observation. This does not conflict with Theorem 2 which only shows that both non-inertial and inertial methods will converge *linearly*, however the convergence factor may be different. Estimating the factor is beyond the scope of this paper and we leave it for future work. Finally we mention that FISTA [32] and other Nesterov-accelerated methods [30] are not applicable to SCAD as it is nonconvex.

6. REFERENCES

[1] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods," *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, 2013.

[2] T. Blumensath and M. E. Davies, "Iterative thresholding for Sparse approximations," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 629–654, Dec. 2008.

[3] D. Lazzaro, "A nonconvex approach to low-rank matrix completion using convex optimization," *Numerical Linear Algebra with Applications*, 2016.

[4] H. H. Zhang, J. Ahn, X. Lin, and C. Park, "Gene selection using support vector machines with non-convex penalty," *Bioinformatics*, vol. 22, no. 1, pp. 88–95, 2006.

[5] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM review*, vol. 57, no. 2, pp. 225–251, 2015.

[6] H. Ji, C. Liu, Z. Shen, and Y. Xu, "Robust video denoising using low rank matrix completion.," in *CVPR*. Citeseer, 2010, pp. 1791–1798.

[7] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

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

[9] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[10] J. Liang, J. Fadili, and G. Peyré, "A multi-step inertial forward–backward splitting method for non-convex optimization," *arXiv preprint arXiv:1606.02118*, 2016.

[11] P. R. Johnstone and P. Moulin, "Local and global convergence of a general inertial proximal splitting scheme," *arXiv preprint arXiv:1602.02726*, 2016.

[12] R. I. Boţ, E. R. Csetnek, and S. C. László, "An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions," *EURO Journal on Computational Optimization*, vol. 4, no. 1, pp. 3–25, 2016.

[13] P. Ochs, Y. Chen, T. Brox, and T. Pock, "iPiano: Inertial proximal algorithm for nonconvex optimization," *SIAM Journal on Imaging Sciences*, vol. 7, no. 2, pp. 1388–1419, 2014.

[14] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th international conference on machine learning (ICML-13)*, 2013, pp. 1139–1147.

[15] K. Kurdyka, "On gradients of functions definable in o-minimal structures," in *Annales de l'institut Fourier*, 1998, vol. 48, pp. 769–783.

[16] S. Lojasiewicz, "Une propriété topologique des sous-ensembles analytiques réels," *Les équations aux dérivées partielles*, pp. 87–89, 1963.

[17] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-lojasiewicz inequality," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 438–457, 2010.

[18] H. Attouch and J. Bolte, "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features," *Mathematical Programming*, vol. 116, no. 1-2, pp. 5–16, 2009.

[19] P. Frankel, G. Garrigos, and J. Peypouquet, "Splitting methods with variable metric for KL functions," *arXiv preprint arXiv:1405.1357*, 2014.

[20] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, "Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function," *Journal of Optimization Theory and Applications*, vol. 162, no. 1, pp. 107–132, 2014.

[21] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. Suter, "From error bounds to the complexity of first-order descent methods for convex functions," *arXiv preprint arXiv:1510.08234*, 2015.

[22] G. Li and T. K. Pong, "Calculus of the exponent of kurdyka-lojasiewicz inequality and its applications to linear convergence of first-order methods," *arXiv preprint arXiv:1602.02915*, 2016.

[23] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[24] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," *arXiv preprint arXiv:1605.07272*, 2016.

[25] S. Zavriev and F. Kostyuk, "Heavy-ball method in nonconvex optimization problems," *Computational Mathematics and Modeling*, vol. 4, no. 4, pp. 336–341, 1993.

[26] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on imaging sciences*, vol. 6, no. 3, pp. 1758–1789, 2013.

[27] D. A. Lorenz and T. Pock, "An inertial forward-backward algorithm for monotone inclusions," *Journal of Mathematical Imaging and Vision*, pp. 1–15, 2014.

[28] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.

[29] P. R. Johnstone and P. Moulin, "Convergence rates of inertial splitting schemes for nonconvex composite optimization," *arXiv preprint*, September 2016.

[30] Y. Nesterov, *Introductory lectures on convex optimization: a basic course*, Springer, 2004.

[31] J. Liang, J. Fadili, and G. Peyré, "Activity identification and local linear convergence of inertial forward-backward splitting," *arXiv preprint arXiv:1503.03703*, 2015.

[32] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *Image Processing, IEEE Transactions on*, vol. 18, no. 11, pp. 2419–2434, 2009.