LINEAR CONVERGENCE OF PROXIMAL GRADIENT ALGORITHM WITH EXTRAPOLATION FOR A CLASS OF NONCONVEX NONSMOOTH MINIMIZATION PROBLEMS

BO WEN[†], XIAOJUN CHEN[‡], AND TING KEI PONG[‡]

Abstract. In this paper, we study the proximal gradient algorithm with extrapolation for minimizing the sum of a Lipschitz differentiable function and a proper closed convex function. Under the error bound condition used in [19] for analyzing the convergence of the proximal gradient algorithm, we show that there exists a threshold such that if the extrapolation coefficients are chosen below this threshold, then the sequence generated converges R-linearly to a stationary point of the problem. Moreover, the corresponding sequence of objective values is also R-linearly convergent. In addition, the threshold reduces to 1 for convex problems and, as a consequence, we obtain the R-linear convergence of the sequence generated by FISTA with fixed restart. Finally, we present some numerical experiments to illustrate our results.

Key words. linear convergence, extrapolation, error bound, accelerated gradient method, nonconvex nonsmooth minimization, convex minimization

AMS subject classifications. 90C30, 65K05, 90C25, 90C26

1. Introduction.

In this paper, we consider the following optimization problem:

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + g(x), \tag{1.1}$$

where g is a proper closed convex function and f is a possibly nonconvex function that has a Lipschitz continuous gradient. We also assume that the proximal operator of μg, i.e.,

$$u \mapsto \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \left\{ g(x) + \frac{1}{2\mu}\|x-u\|^2 \right\}$$

is easy to compute for all $\mu > 0$ and any $u \in \mathbb{R}^n$, where arg min denotes the *unique* minimizer. We also assume that the optimal value of (1.1) is finite and is attained. Problem (1.1) arises in many important contemporary applications including compressed sensing [8, 13], matrix completion [7] and image processing [9]. Since the problem instances are typically of large scale, first-order methods such as the proximal gradient algorithm [17] are used for solving them, whose main computational efforts per iteration are the evaluations of the gradient of f and the proximal mapping of μg. For the proximal gradient algorithm, when f is in addition convex, it is known that

$$F(x^k) - \inf_{x\in\mathbb{R}^n} F(x) = O\left(\frac{1}{k}\right),$$

[†] Department of Mathematics, Harbin Institute of Technology, Harbin, P.R. China. Current address: Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, P.R. China. (bo.wen@ connect.polyu.hk). This author's work was supported in part by the NSFC 11471088 and Hong Kong Research Grants Council PolyU5002/13p.

[‡] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, P.R. China.(xiaojun.chen@polyu.edu.hk, tk.pong@polyu.edu.hk). The second author's work was supported in part by Hong Kong Research Grants Council PolyU153001/14p. The third author's work was supported in part by Hong Kong Research Grants Council PolyU253008/15p.

where $\{x^k\}$ is generated by the proximal gradient algorithm; see, for example, [32, Theorem 1(a)]. However, the proximal gradient algorithm, in its original form, can be slow in practice; see, for example, [12, Section 5].

Various attempts have thus been made to accelerate the proximal gradient algorithm. One simple and often efficient strategy is to perform extrapolation, where *momentum* terms involving the previous iterations are added to the current iteration. A prototypical algorithm takes the following form

$$\begin{cases} y^k = x^k + \beta_k(x^k - x^{k-1}), \\ x^{k+1} = \underset{x\in\mathbb{R}^n}{\arg\min} \left\{ \langle \nabla f(y^k), x\rangle + \frac{1}{2\mu}\|x - y^k\|^2 + g(x) \right\}, \end{cases} \tag{1.2}$$

where $\mu > 0$ is a constant that depends on the Lipschitz continuity modulus of ∇f, and the extrapolation coefficients β_k satisfy $0 \le \beta_k \le 1$ for all k. A recent example is the fast iterative shrinkage-thresholding algorithm (FISTA) proposed by Beck and Teboulle [2], which is based on Nesterov's extrapolation techniques [22, 23, 24, 26] and is designed for solving (1.1) with f being convex and g being continuous. Their analysis can be directly extended to the case when g is a proper closed convex function. The same algorithm was also independently proposed and studied by Nesterov [25]. FISTA takes the form (1.2) and requires $\{\beta_k\}$ to satisfy a certain recurrence relation. It was shown in [2, 25] that this algorithm exhibits a faster convergence rate than the proximal gradient algorithm, which is

$$F(x^k) - \inf_{x\in\mathbb{R}^n} F = O\left(\frac{1}{k^2}\right),$$

where $\{x^k\}$ is generated by FISTA. Many accelerated proximal gradient algorithms based on Nesterov's extrapolation techniques have been proposed since then, and we refer the readers to [4, 5, 32] and the references therein for an overview of these algorithms.

The faster convergence rate of FISTA in terms of objective values motivates subsequent studies on the extrapolation scheme (1.2); see, for example, [1, 10, 12, 16, 33]. Particularly, O'Donoghue and Candès [12] proposed an adaptive restart scheme for β_k based on FISTA for solving (1.1) with f being convex and $g = 0$. Specifically, instead of following the recurrence relation of β_k in FISTA for all k, they reset $\beta_k = \beta_0$ every K iterations, where K is a positive number. They established *global* linear convergence of the function values when f is strongly convex if K is sufficiently large. Their algorithm is robust against errors in the estimation of the strong convexity modulus of f; see the discussion in [12, Section 2.1]. Later, Attouch and Chbani [1], and independently, Chambolle and Dossal [10], established the convergence of the whole sequence generated by (1.2) for solving (1.1) when f is convex and $\beta_k = \frac{k-1}{k+\alpha-1}$ for any fixed $\alpha > 3$. More recently, Tao, Boley and Zhang [33] established local linear convergence of FISTA applied to the LASSO (i.e., f is a least squares loss function and g is a positive multiple of the ℓ_1 norm) under the assumption that the problem has a unique solution that satisfies strict complementarity. Johnstone and Moulin [16] considered (1.1) with f being convex, and showed that the whole sequence generated by (1.2) is convergent by assuming that the extrapolation coefficients β_k satisfy $0 \le \beta_k \le \bar\beta$ for some $\bar\beta < 1$. Moreover, by imposing uniqueness of the optimal solution together with a technical assumption, they showed that the sequence generated by (1.2) is locally linearly convergent when applied to the LASSO for a particular choice of $\{\beta_k\}$.

Despite the rich literature, we note that the local linear convergence of (1.2) is only established for a certain type of convex problems with *unique* optimal solutions for some specific choices of $\{\beta_k\}$, which can be restrictive for practical applications. Thus, in this paper, we further study the behavior of the sequence $\{x^k\}$ generated by (1.2). Specifically, we discuss local linear convergence under more general conditions in the possibly nonconvex case.

In details, under the same error bound condition used in [19] for analyzing convergence of the proximal gradient algorithm, we show that there is a threshold $\widetilde{\beta}$ depending on f so that if $\sup_k \beta_k < \widetilde{\beta}$, then the sequence $\{x^k\}$ generated by (1.2) converges R-linearly to a stationary point of (1.1) and the sequence of the objective value $\{F(x^k)\}$ is also R-linearly convergent. In particular, if f is in addition convex, then $\widetilde{\beta}$ reduces to 1 and we can conclude that the sequence $\{x^k\}$ generated by FISTA with fixed restart is R-linearly convergent to an optimal solution of (1.1); see Section 3.3. The error bound condition is satisfied for a wide range of problems including the LASSO, and hence our linear convergence result concerning (1.2) with a fixed μ is more general than those discussed in [16].

The rest of this paper is organized as follows. Section 2 presents some basic notation and preliminary materials. In Section 3, we establish linear convergence of the iterates generated by the proximal gradient algorithm with extrapolation under the same error bound condition used in [19]. Linear convergence of the corresponding sequence of function values is also established. FISTA with restart is discussed in Section 3.3. In Section 4, we perform numerical experiments to illustrate our results.

2. Notation and preliminaries. Throughout this paper, we use $\mathbb{R}^n$ to denote the n-dimensional Euclidean space, with its standard inner product denoted by $\langle \cdot, \cdot \rangle$. The Euclidean norm is denoted by $\|\cdot\|$, the ℓ_1 norm is denoted by $\|\cdot\|_1$ and the ℓ_∞ norm is denoted by $\|\cdot\|_\infty$. The vector of all ones is denoted by e, whose dimension should be clear from the context. For a matrix $A \in \mathbb{R}^{m\times n}$, we use $A^\top$ to denote its transpose. Finally, for a symmetric matrix $A \in \mathbb{R}^{n\times n}$, we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote its largest and smallest eigenvalue, respectively.

For a nonempty closed set $\mathcal{C} \subseteq \mathbb{R}^n$, its indicator function is defined by

$$\delta_{\mathcal{C}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{C}, \\ +\infty & \text{if } x \notin \mathcal{C}. \end{cases}$$

Moreover, we use $\mathrm{dist}(x, \mathcal{C})$ to denote the distance from x to $\mathcal{C}$, where $\mathrm{dist}(x, \mathcal{C}) = \inf_{y\in\mathcal{C}} \|x - y\|$. When $\mathcal{C}$ is in addition convex, we use $\mathrm{Proj}_{\mathcal{C}}(x)$ to denote the unique closest point on $\mathcal{C}$ to x.

The domain of an extended-real-valued function $h : \mathbb{R}^n \to [-\infty, \infty]$ is defined as $\mathrm{dom}\, h = \{x \in \mathbb{R}^n : h(x) < +\infty\}$. We say that h is proper if it never equals $-\infty$ and $\mathrm{dom}\, h \neq \emptyset$. Such a function is closed if it is lower semicontinuous. A proper closed function h is said to be level bounded if the lower level sets of h are bounded, i.e., the set $\{x \in \mathbb{R}^n : h(x) \le r\}$ is bounded for any $r \in \mathbb{R}$. For a proper closed convex function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, the subdifferential of h at $x \in \mathrm{dom}\, h$ is given by

$$\partial h(x) = \{\xi \in \mathbb{R}^n : h(u) - h(x) - \langle \xi, u - x\rangle \ge 0, \ \forall u \in \mathbb{R}^n\}.$$

We use $\mathrm{Prox}_h(v)$ to denote the proximal operator of a proper closed convex function h at any $v \in \mathbb{R}^n$, i.e.:

$$\mathrm{Prox}_h(v) = \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \left\{ h(x) + \frac{1}{2}\|x - v\|^2 \right\}.$$

We note that this operator is well defined for any $v \in \mathbb{R}^n$, and we refer the readers to [27, Chapter 1] for properties of the proximal operator.

For an optimal solution $\hat{x}$ of (1.1), the following first-order necessary condition always holds, thanks to [29, Exercise 8.8(c)]:

$$0 \in \nabla f(\hat{x}) + \partial g(\hat{x}), \tag{2.1}$$

where ∇f denotes the gradient of f. We say that $\tilde{x}$ is a stationary point of (1.1) if $\tilde{x}$ satisfies (2.1) in place of $\hat{x}$; in particular, any optimal solution $\hat{x}$ of (1.1) is a stationary point of (1.1). We use $\mathcal{X}$ to denote the set of stationary points of F.

Finally, we recall two notions of (local) linear convergence, which will be used in our convergence analysis. For a sequence $\{x^k\}$, we say that $\{x^k\}$ converges Q-linearly to x^* if there exist $c \in (0,1)$ and $k_0 > 0$ such that

$$\|x^{k+1} - x^*\| \le c\|x^k - x^*\|, \quad \forall k \ge k_0;$$

and we say that $\{x^k\}$ converges R-linearly to x^* if

$$\limsup_{k\to\infty} \|x^k - x^*\|^{\frac{1}{k}} < 1.$$

We state the following simple fact relating the two notions of linear convergence, which is an immediate consequence of the definitions of Q- and R-linear convergence. We will use this fact in our convergence analysis.

LEMMA 2.1. *Suppose that $\{a_k\}$ and $\{b_k\}$ are two sequences in $\mathbb{R}$ with $0 \le b_k \le a_k$ for all k, and $\{a_k\}$ is Q-linearly convergent to zero. Then $\{b_k\}$ is R-linearly convergent to zero.*

3. Convergence analysis of the proximal gradient algorithm with extrapolation.

In this section, we present the proximal gradient algorithm with extrapolation for solving (1.1), and discuss the convergence behavior of the sequence generated by the algorithm.

We recall that in our problem (1.1), the function g is proper closed convex and f has a Lipschitz continuous gradient; moreover, $\inf F > -\infty$ and $\mathcal{X} \neq \emptyset$. Furthermore, we observe that any function f whose gradient is Lipschitz continuous can be written as $f = f_1 - f_2$, where f_1 and f_2 are two convex functions with Lipschitz continuous gradients. For instance, one can decompose f as

$$f(x) = \underbrace{f(x) + \frac{c}{2}\|x\|^2}_{f_1(x)} - \underbrace{\frac{c}{2}\|x\|^2}_{f_2(x)},$$

for any $c \ge L_f$, where L_f is a Lipschitz continuity modulus of ∇f. It is then routine to show that both f_1 and f_2 are convex functions with Lipschitz continuous gradients.

Thus, without loss of generality, from now on, we assume that $f = f_1 - f_2$ for some convex functions f_1 and f_2 with Lipschitz continuous gradients. For concreteness, we denote a Lipschitz continuity modulus of ∇f_1 by $L > 0$, and a Lipschitz continuity modulus of ∇f_2 by $l \ge 0$. Moreover, by taking a larger L if necessary, we assume throughout that $L \ge l$. Then it is not hard to show that ∇f is Lipschitz continuous with a modulus L.

We are now ready to present our proximal gradient algorithm with extrapolation.

```
Algorithm 1: Proximal gradient algorithm with extrapolation
Input: $x^0 \in \mathrm{dom}\, g$, $\{\beta_k\} \subseteq \left[0, \sqrt{\frac{L}{L+l}}\right]$. Set $x^{-1} = x^0$.
    for $k = 0, 1, 2, \cdots$ do
        $y^k = x^k + \beta_k(x^k - x^{k-1}),$
        $x^{k+1} = \mathrm{Prox}_{\frac{1}{L}g}\left(y^k - \frac{1}{L}\nabla f(y^k)\right).$        (3.1)
    end for
```

We shall discuss the convergence behavior of Algorithm 1. We note first that it is immedidate from the definition of the proximal operator that the x-update in (3.1) is equivalently given by

$$x^{k+1} = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \left\{ \langle \nabla f(y^k), x \rangle + \frac{L}{2}\|x - y^k\|^2 + g(x) \right\}. \tag{3.2}$$

This fact will be used repeatedly in our convergence analysis below. Our analysis also relies heavily on the following auxiliary sequence:

$$H_{k,\alpha} = F(x^k) + \alpha\|x^k - x^{k-1}\|^2, \tag{3.3}$$

for a fixed $\alpha \in [\frac{L+l}{2}\bar{\beta}^2, \frac{L}{2}]$ with $\bar{\beta} := \sup_k \beta_k$, where $\{x^k\}$ is generated by Algorithm 1. We study the convergence properties of $\{H_{k,\alpha}\}$ in Section 3.1. The results will then be used in subsequent subsections for analyzing the convergence of $\{x^k\}$ and $\{F(x^k)\}$. The auxiliary sequence (3.3) was also used in [1, 10, 16] for analyzing (1.2).

3.1. Auxiliary lemmas.

We start by showing that $\{H_{k,\alpha}\}$ is nonincreasing and convergent.

LEMMA 3.1. *Let $\{x^k\}$ be a sequence generated by Algorithm 1 and $\alpha \in [\frac{L+l}{2}\bar{\beta}^2, \frac{L}{2}]$. Then the following statements hold.*

(i) *For any $z \in \mathrm{dom}\, g$, we have*

$$F(x^{k+1}) \le F(z) + \frac{L+l}{2}\|z - y^k\|^2 - \frac{L}{2}\|x^{k+1} - z\|^2. \tag{3.4}$$

(ii) *It holds that for all k,*

$$H_{k+1,\alpha} - H_{k,\alpha} \le \left(-\frac{L}{2} + \alpha\right)\|x^{k+1} - x^k\|^2 + \left(\frac{L+l}{2}\beta_k^2 - \alpha\right)\|x^k - x^{k-1}\|^2. \tag{3.5}$$

(iii) *The sequence $\{H_{k,\alpha}\}$ is nonincreasing.*

Proof. We first prove (i). Fix any $z \in \mathrm{dom}\, g$. Using the definition of x^{k+1} in (3.2) and the strong convexity of the objective in the minimization problem (3.2), we obtain upon rearranging terms that

$$\begin{aligned} g(x^{k+1}) \le g(z) + \langle -\nabla f(y^k), x^{k+1} - z\rangle + \frac{L}{2}\|z - y^k\|^2 \\ - \frac{L}{2}\|x^{k+1} - y^k\|^2 - \frac{L}{2}\|x^{k+1} - z\|^2. \end{aligned} \tag{3.6}$$

On the other hand, using the fact that ∇f is Lipschitz continuous with a Lipschitz continuity modulus L, we have

$$f(x^{k+1}) \le f(y^k) + \langle \nabla f(y^k), x^{k+1} - y^k\rangle + \frac{L}{2}\|x^{k+1} - y^k\|^2. \tag{3.7}$$

Summing (3.6) and (3.7), we see further that

$$\begin{aligned} f(x^{k+1}) + g(x^{k+1}) \le\ & f(y^k) + g(z) + \langle \nabla f(y^k), z - y^k\rangle \\ & + \frac{L}{2}\|z - y^k\|^2 - \frac{L}{2}\|x^{k+1} - z\|^2. \end{aligned} \tag{3.8}$$

Next, recall that $f = f_1 - f_2$. Hence, we have

$$\begin{aligned} & f(y^k) + \langle \nabla f(y^k), z - y^k\rangle \\ & = f_1(y^k) - f_2(y^k) + \langle \nabla f_1(y^k), z - y^k\rangle - \langle \nabla f_2(y^k), z - y^k\rangle. \end{aligned} \tag{3.9}$$

Since f_1 and f_2 are convex and their gradients are Lipschitz continuous with moduli L and l, respectively, the following two inequalities hold.

$$\begin{aligned} f_1(y^k) + \langle \nabla f_1(y^k), z - y^k\rangle &\le f_1(z), \\ f_2(z) - f_2(y^k) - \langle \nabla f_2(y^k), z - y^k\rangle &\le \frac{l}{2}\|z - y^k\|^2. \end{aligned}$$

Combining these relations with (3.9) and recalling that $f = f_1 - f_2$, we see further that

$$f(y^k) + \langle \nabla f(y^k), z - y^k\rangle \le f(z) + \frac{l}{2}\|z - y^k\|^2. \tag{3.10}$$

Summing (3.8) and (3.10), and recalling that $F = f + g$, we obtain (3.4) immediately. This proves (i).

We now prove (ii). We note first from the definition of the y-update in (3.1) that $y^k - x^k = \beta_k(x^k - x^{k-1})$. Using this and (3.4) with $z = x^k$, we obtain that

$$F(x^{k+1}) - F(x^k) \le \frac{L+l}{2}\beta_k^2\|x^k - x^{k-1}\|^2 - \frac{L}{2}\|x^{k+1} - x^k\|^2.$$

From this and the definition of $H_{k,\alpha}$ from (3.3), we see further that

$$\begin{aligned} & H_{k+1,\alpha} - H_{k,\alpha} = F(x^{k+1}) + \alpha\|x^{k+1} - x^k\|^2 - F(x^k) - \alpha\|x^k - x^{k-1}\|^2 \\ & \le -\frac{L}{2}\|x^{k+1} - x^k\|^2 + \frac{L+l}{2}\beta_k^2\|x^k - x^{k-1}\|^2 + \alpha\|x^{k+1} - x^k\|^2 - \alpha\|x^k - x^{k-1}\|^2 \\ & = \left(-\frac{L}{2} + \alpha\right)\|x^{k+1} - x^k\|^2 + \left(\frac{L+l}{2}\beta_k^2 - \alpha\right)\|x^k - x^{k-1}\|^2, \end{aligned}$$

which is just (3.5). This proves (ii). Finally, since $\frac{L+l}{2}\bar\beta^2 \le \alpha \le \frac{L}{2}$ by our assumption, we have

$$-\frac{L}{2} + \alpha \le 0, \text{ and } \frac{L+l}{2}\beta_k^2 - \alpha \le \frac{L+l}{2}\bar\beta^2 - \alpha \le 0 \ \ \forall k.$$

Consequently, $H_{k+1,\alpha} - H_{k,\alpha} \le 0$, i.e., $\{H_{k,\alpha}\}$ is nonincreasing. This completes the proof. □

The following result is an immediate consequence of Lemma 3.1.

Corollary 3.2. *The sequence $\{x^k\}$ generated by Algorithm 1 is bounded if F is level bounded.*

Proof. From Lemma 3.1, the sequence $\{H_{k,\frac{L}{2}}\}$ is nonincreasing. This together with the definition of $H_{k,\frac{L}{2}}$ implies that

$$F(x^k) \le H_{k,\frac{L}{2}} \le H_{0,\frac{L}{2}} < \infty.$$

Since F is level bounded by assumption, we conclude that $\{x^k\}$ is bounded. □

LEMMA 3.3. *Let $\{x^k\}$ be a sequence generated by Algorithm 1, and $\alpha \in [\frac{L+l}{2}\bar{\beta}^2, \frac{L}{2}]$. Then the following statements hold.*

(i) *The sequence $\{H_{k,\alpha}\}$ is convergent.*

(ii) $\sum_{k=0}^{\infty}\left(\alpha - \frac{L+l}{2}\beta_{k+1}^2\right)\|x^{k+1}-x^k\|^2 < \infty$.

Proof. Recall that $\inf F > -\infty$. Hence, $H_{k,\alpha} = F(x^k)+\alpha\|x^k - x^{k-1}\|^2$ is bounded from below. This together with the fact that $\{H_{k,\alpha}\}$ is nonincreasing from Lemma 3.1 implies that $\{H_{k,\alpha}\}$ is convergent. This proves (i).

We now prove (ii). Since $-\frac{L}{2}+\alpha \le 0$, we have from (3.5) that

$$H_{k+1,\alpha} - H_{k,\alpha} \le -\left(\alpha - \frac{L+l}{2}\beta_k^2\right)\|x^k - x^{k-1}\|^2. \tag{3.11}$$

Summing both sides of (3.11) from 1 to N, we see further that

$$0 \le \sum_{k=1}^{N}\left(\alpha - \frac{L+l}{2}\beta_k^2\right)\|x^k - x^{k-1}\|^2 \le \sum_{k=1}^{N}(H_{k,\alpha} - H_{k+1,\alpha}) = H_{1,\alpha} - H_{N+1,\alpha}, \tag{3.12}$$

where the nonnegativity follows from the fact that $\alpha \ge \frac{L+l}{2}\bar{\beta}^2 \ge \frac{L+l}{2}\beta_k^2$ for all k. Since $\{H_{k,\alpha}\}$ is convergent by (i), letting $N \to \infty$ in (3.12), we conclude that the infinite sum exists and is finite, i.e.,

$$\sum_{k=1}^{\infty}\left(\alpha - \frac{L+l}{2}\beta_k^2\right)\|x^k - x^{k-1}\|^2 < \infty.$$

This completes the proof. □

In the next lemma, we show that when $\{\beta_k\}$ is chosen below a certain threshold, then any accumulation point of the sequence $\{x^k\}$ generated by Algorithm 1, if exists, is a stationary point of F. This result has been established in [16] when the function f is convex. Indeed, in the convex case, it was shown in [16, Theorem 4.1] that the whole sequence $\{x^k\}$ is convergent. However, the following convergence result is new when the function f is nonconvex.

LEMMA 3.4. *Suppose that $\bar{\beta} < \sqrt{\frac{L}{L+l}}$ and $\{x^k\}$ is a sequence generated by Algorithm 1. Then the following statements hold.*

(i) $\sum_{k=0}^{\infty}\|x^{k+1}-x^k\|^2 < \infty$.

(ii) *Any accumulation point of $\{x^k\}$ is a stationary point of F.*

Proof. Since $\bar{\beta} < \sqrt{\frac{L}{L+l}}$, one can choose $\alpha \in (\frac{L+l}{2}\bar{\beta}^2, \frac{L}{2})$. Then $\frac{L+l}{2}\beta_k^2 \le \frac{L+l}{2}\bar{\beta}^2 < \alpha$ for all k, and the conclusion in (i) follows immediately from Lemma 3.3 (ii).

We next prove (ii). Let $\bar{x}$ be an accumulation point. Then there exists a subsequence $\{x^{k_i}\}$ such that $\lim\limits_{i\to\infty} x^{k_i} = \bar{x}$. Using the first-order optimality condition of the minimization problem (3.2), we obtain

$$-L(x^{k_i+1} - y^{k_i}) \in \nabla f(y^{k_i}) + \partial g(x^{k_i+1}).$$

Combining this with the definition of y^{k_i}, which is $y^{k_i} = x^{k_i} + \beta_{k_i}(x^{k_i} - x^{k_i-1})$, we see further that

$$-L[(x^{k_i+1} - x^{k_i}) - \beta_{k_i}(x^{k_i} - x^{k_i-1})] \in \nabla f(y^{k_i}) + \partial g(x^{k_i+1}). \tag{3.13}$$

Passing to the limit in (3.13), and invoking $\|x^{k_i+1} - x^{k_i}\| \to 0$ from (i) together with the continuity of ∇f and the closedness of ∂g (see, for example, [6, Page 80]), we have

$$0 \in \nabla f(\bar{x}) + \partial g(\bar{x}),$$

meaning that $\bar{x}$ is a stationary point of F. This completes the proof. □

Let Ω be the set of accumulation points of the sequence $\{x^k\}$ generated by Algorithm 1. Then, from Corollary 3.2 and Lemma 3.4 (ii), we have $\emptyset \neq \Omega \subseteq \mathcal{X}$ when F is level bounded. We prove in the next proposition that F is constant over Ω if $\{\beta_k\}$ is chosen below a certain threshold. Since F is only assumed to be lower semicontinuous, this conclusion is nontrivial when F has stationary points that are not globally optimal.

Proposition 3.5. *Suppose that $\bar{\beta} < \sqrt{\frac{L}{L+l}}$ and $\{x^k\}$ is a sequence generated by Algorithm 1 with its set of accumulation points denoted by Ω. Then $\zeta := \lim\limits_{k\to\infty} F(x^k)$ exists and $F \equiv \zeta$ on Ω.*

Proof. Fix any $\alpha \in (\frac{L+l}{2}\bar{\beta}^2, \frac{L}{2})$, which exists because $\bar{\beta} < \sqrt{\frac{L}{L+l}}$. Then, in view of Lemmas 3.3 and 3.4, the sequence $\{H_{k,\alpha}\}$ is convergent and $\|x^{k+1} - x^k\| \to 0$. These together with the definition of $H_{k,\alpha}$ imply that $\lim\limits_{k\to\infty} F(x^k)$ exists. We denote this limit by ζ.

We now show that $F \equiv \zeta$ on Ω. If $\Omega = \emptyset$, then the conclusion holds trivially. Otherwise, take any $\hat{x} \in \Omega$. Then there exists a convergent subsequence $\{x^{k_i}\}$ with $\lim\limits_{i\to\infty} x^{k_i} = \hat{x}$. From the lower semicontinuity of F and the definition of ζ, we have

$$F(\hat{x}) \leq \liminf_{i\to\infty} F(x^{k_i}) = \zeta. \tag{3.14}$$

On the other hand, using the definition of x^{k_i} as the minimizer in (3.2), we see that

$$g(x^{k_i}) + \langle \nabla f(y^{k_i-1}), x^{k_i} - \hat{x}\rangle + \frac{L}{2}\|x^{k_i} - y^{k_i-1}\|^2 \leq g(\hat{x}) + \frac{L}{2}\|\hat{x} - y^{k_i-1}\|^2. \tag{3.15}$$

Adding $f(x^{k_i})$ to both sides of (3.15), we obtain further that

$$f(x^{k_i})+g(x^{k_i})+\langle \nabla f(y^{k_i-1}), x^{k_i}-\hat{x}\rangle+\frac{L}{2}\|x^{k_i}-y^{k_i-1}\|^2 \leq f(x^{k_i})+g(\hat{x})+\frac{L}{2}\|\hat{x}-y^{k_i-1}\|^2. \tag{3.16}$$

Next, recall that $y^{k_i-1} = x^{k_i-1} + \beta_{k_i-1}(x^{k_i-1} - x^{k_i-2})$. Thus, we have

$$\begin{aligned} \|x^{k_i} - y^{k_i-1}\| &= \|x^{k_i} - x^{k_i-1} - \beta_{k_i-1}(x^{k_i-1} - x^{k_i-2})\| \\ &\leq \|x^{k_i} - x^{k_i-1}\| + \bar{\beta}\|x^{k_i-1} - x^{k_i-2}\|. \end{aligned} \tag{3.17}$$

In addition, we also have

$$\begin{aligned} \|\hat{x} - y^{k_i-1}\| &= \|\hat{x} - x^{k_i} + x^{k_i} - y^{k_i-1}\| \\ &\leq \|\hat{x} - x^{k_i}\| + \|x^{k_i} - y^{k_i-1}\|. \end{aligned} \tag{3.18}$$

Since $\|x^{k+1} - x^k\| \to 0$ and $\lim\limits_{i\to\infty} x^{k_i} = \hat{x}$, it follows from (3.17) and (3.18) that

$$\|x^{k_i} - y^{k_i-1}\| \to 0 \text{ and } \|\hat{x} - y^{k_i-1}\| \to 0,$$

and hence $\nabla f(y^{k_i-1}) \to \nabla f(\hat{x})$. From these and (3.16), we obtain that

$$\zeta = \limsup_{i\to\infty} F(x^{k_i}) \le F(\hat{x}). \tag{3.19}$$

Thus $F(\hat{x}) = \lim\limits_{i\to\infty} F(x^{k_i}) = \zeta$ from (3.14) and (3.19). Since $\hat{x} \in \Omega$ is arbitrary, we see that $F \equiv \zeta$ on Ω. This completes the proof. □

3.2. Linear convergence of $\{x^k\}$ and $\{F(x^k)\}$.

In this subsection, we establish local linear convergence of $\{x^k\}$ and $\{F(x^k)\}$ under the following assumption.

Assumption 3.1.

(i) **(Error bound condition)** *For any $\xi \ge \inf_{x\in\mathbb{R}^n} F(x)$, there exist $\epsilon > 0$ and $\tau > 0$ such that*

$$\operatorname{dist}(x, \mathcal{X}) \le \tau \left\| \operatorname{Prox}_{\frac{1}{L}g}\left(x - \frac{1}{L}\nabla f(x)\right) - x \right\|,$$

whenever $\|\operatorname{Prox}_{\frac{1}{L}g}(x - \frac{1}{L}\nabla f(x)) - x\| < \epsilon$ and $F(x) \le \xi$.

(ii) *There exists $\delta > 0$, such that $\|x - y\| \ge \delta$ whenever $x, y \in \mathcal{X}$, $F(x) \ne F(y)$.*

The above assumption has been used in the convergence analysis of many algorithms, including the gradient projection and block coordinate gradient descent method, etc; see, for example, [3, 18, 19, 20, 30, 31, 32] and the references therein. The assumption consists of two parts: the first part is an error bound condition, while the second part states that when restricted to $\mathcal{X}$, the isocost surfaces of F are properly separated. Under our blanket assumptions on F, Assumption 3.1 is known to be satisfied for many choices of f and g, including:

- $f(x) = h(Ax)$, and g is a polyhedral function, where h is twice continuously differentiable on $\mathbb{R}^n$ with a Lipschitz continuous gradient, and on any compact convex set, h is strongly convex; see, [18, Theorem 2.1] and [31, Lemma 6]. This covers the well-known LASSO;
- f is a possibly nonconvex quadratic function, and g is a polyhedral function; see, for example, [31, Theorem 4].

The first example is convex, while the second one is possibly nonconvex. We refer the readers to [31, 32, 34] and the references therein for more examples and discussions on the error bound condition.

We next show that $\{H_{k,\alpha}\}$ is Q-linearly convergent under Assumption 3.1. Our analysis uses ideas from the proof of [31, Theorem 2], which studied a block coordinate gradient descent method.

Lemma 3.6. *Suppose that $\bar\beta < \sqrt{\frac{L}{L+l}}$, $\alpha \in (\frac{L+l}{2}\bar\beta^2, \frac{L}{2})$ and that Assumption 3.1 holds. Let $\{x^k\}$ be a sequence generated by Algorithm 1. Then the following statements hold.*

(i) $\lim\limits_{k\to\infty} \operatorname{dist}(x^k, \mathcal{X}) = 0$.

(ii) *The sequence $\{H_{k,\alpha}\}$ is Q-linearly convergent.*

Proof. First we prove (i). Observe that

$$\begin{aligned}
&\left\| \operatorname{Prox}_{\frac{1}{L}g}\left(x^k - \frac{1}{L}\nabla f(x^k)\right) - x^k \right\| \\
&\le \left\| \operatorname{Prox}_{\frac{1}{L}g}\left(x^k - \frac{1}{L}\nabla f(x^k)\right) - \operatorname{Prox}_{\frac{1}{L}g}\left(y^k - \frac{1}{L}\nabla f(y^k)\right) \right\| \\
&\quad + \left\| \operatorname{Prox}_{\frac{1}{L}g}\left(y^k - \frac{1}{L}\nabla f(y^k)\right) - y^k \right\| + \|y^k - x^k\|.
\end{aligned} \tag{3.20}$$

We now derive an upper bound for the first term on the right hand side of (3.20). To this end, using the nonexpansiveness property of the proximal operator (see, for example, [28, Page 340]), we have

$$\begin{aligned}&\left\|\mathrm{Prox}_{\frac{1}{L}g}\left(x^k - \frac{1}{L}\nabla f(x^k)\right) - \mathrm{Prox}_{\frac{1}{L}g}\left(y^k - \frac{1}{L}\nabla f(y^k)\right)\right\| \\ &\le \left\|x^k - \frac{1}{L}\nabla f(x^k) - y^k + \frac{1}{L}\nabla f(y^k)\right\| \\ &\le \|x^k - y^k\| + \frac{1}{L}\|\nabla f(x^k) - \nabla f(y^k)\| \le 2\|x^k - y^k\|,\end{aligned} \tag{3.21}$$

where the last inequality follows from the fact that ∇f is Lipschitz continuous with modulus L. Combining (3.20), (3.21) and invoking the definition of x^{k+1} in Algorithm 1, we see further that

$$\begin{aligned}&\left\|\mathrm{Prox}_{\frac{1}{L}g}\left(x^k - \frac{1}{L}\nabla f(x^k)\right) - x^k\right\| \le 3\|x^k - y^k\| + \|x^{k+1} - y^k\| \\ &\le 4\|x^k - y^k\| + \|x^{k+1} - x^k\| \le 4\bar{\beta}\|x^k - x^{k-1}\| + \|x^{k+1} - x^k\|,\end{aligned} \tag{3.22}$$

where the last inequality follows from the definition of y^k in (3.1) and the definition of $\bar{\beta}$. Since $\|x^{k+1} - x^k\| \to 0$ by Lemma 3.4, we conclude from (3.22) that

$$\left\|\mathrm{Prox}_{\frac{1}{L}g}\left(x^k - \frac{1}{L}\nabla f(x^k)\right) - x^k\right\| \to 0. \tag{3.23}$$

Let $\xi = H_{0,\alpha}$. Since $\{H_{k,\alpha}\}$ is nonincreasing by Lemma 3.1, we must have $H_{k,\alpha} \le \xi$ for all k, and consequently $F(x^k) \le \xi$ for all k. In view of this, (3.23) and Assumption 3.1 (i), we see that for $\xi = H_{0,\alpha}$, there exist $\tau > 0$ and a positive integer K so that for all $k \ge K$, we have

$$\mathrm{dist}(x^k, \mathcal{X}) \le \tau \left\|\mathrm{Prox}_{\frac{1}{L}g}\left(x^k - \frac{1}{L}\nabla f(x^k)\right) - x^k\right\|. \tag{3.24}$$

Thus from (3.23) and (3.24), we immediately obtain the conclusion in (i).

We now prove (ii). Take an arbitrary $z \in \mathcal{X}$, we have from (3.4) that

$$\begin{aligned}F(x^{k+1}) &\le F(z) + \frac{L+l}{2}\|z - y^k\|^2 - \frac{L}{2}\|x^{k+1} - z\|^2 \\ &\le F(z) + \frac{L+l}{2}\|z - y^k\|^2 \\ &= F(z) + \frac{L+l}{2}\|z - x^k + x^k - y^k\|^2 \\ &\le F(z) + (L+l)\|z - x^k\|^2 + (L+l)\|x^k - y^k\|^2.\end{aligned} \tag{3.25}$$

Choose z in (3.25) to be an $\bar{x}^k \in \mathcal{X}$ so that $\|\bar{x}^k - x^k\| = \mathrm{dist}(x^k, \mathcal{X})$. Then we obtain

$$F(x^{k+1}) - F(\bar{x}^k) \le (L+l)\mathrm{dist}^2(x^k, \mathcal{X}) + (L+l)\|x^k - y^k\|^2. \tag{3.26}$$

In addition, recall that $\|x^{k+1} - x^k\| \to 0$ by Lemma 3.4. This together with (3.23) and (3.24) shows that $\|\bar{x}^{k+1} - \bar{x}^k\| \to 0$. In view of this and Assumption 3.1 (ii), it must then hold true that $F(\bar{x}^k) \equiv \zeta$ for some constant ζ for all sufficiently large k. Thus, for all sufficiently large k, we have from (3.26) that

$$F(x^{k+1}) - \zeta \le (L+l)\mathrm{dist}^2(x^k, \mathcal{X}) + (L+l)\|x^k - y^k\|^2. \tag{3.27}$$

On the other hand, since $\bar{x}^k$ is a stationary point of (1.1) so that $-\nabla f(\bar{x}^k) \in \partial g(\bar{x}^k)$, we have for all k that,

$$g(\bar{x}^k) - g(x^k) \le \langle -\nabla f(\bar{x}^k), \bar{x}^k - x^k\rangle.$$

Using this and the definitions of F, $H_{k,\alpha}$ and ζ, we see that for all sufficiently large k,

$$\begin{aligned}
&\zeta - H_{k,\alpha} = F(\bar{x}^k) - F(x^k) - \alpha\|x^k - x^{k-1}\|^2\\
&= f(\bar{x}^k) + g(\bar{x}^k) - f(x^k) - g(x^k) - \alpha\|x^k - x^{k-1}\|^2\\
&\le f(\bar{x}^k) - f(x^k) + \langle -\nabla f(\bar{x}^k), \bar{x}^k - x^k\rangle - \alpha\|x^k - x^{k-1}\|^2\\
&= -f(x^k) - [-f(\bar{x}^k)] - \langle -\nabla f(\bar{x}^k), x^k - \bar{x}^k\rangle - \alpha\|x^k - x^{k-1}\|^2\\
&\le \frac{L}{2}\|x^k - \bar{x}^k\|^2 - \alpha\|x^k - x^{k-1}\|^2,
\end{aligned}$$

where the last inequality follows from the Lipschitz continuity of $-\nabla f$. Using this, the fact that $\|x^{k+1} - x^k\| \to 0$ by Lemma 3.4 and the fact that $\|\bar{x}^k - x^k\| = \mathrm{dist}(x^k, \mathcal{X}) \to 0$ by (i), we deduce that

$$\zeta \le \lim_{k\to\infty} H_{k,\alpha} = \inf_k H_{k,\alpha}, \tag{3.28}$$

where the equality follows from Lemma 3.1 (iii).

Now, from (3.22), (3.24) and (3.27), we see that for all sufficiently large k,

$$\begin{aligned}
F(x^{k+1}) - \zeta &\le (L+l)\mathrm{dist}^2(x^k, \mathcal{X}) + (L+l)\|x^k - y^k\|^2\\
&\le (L+l)\tau^2(4\bar{\beta}\|x^k - x^{k-1}\| + \|x^{k+1} - x^k\|)^2 + (L+l)\|x^k - y^k\|^2\\
&\le (L+l)\tau^2(4\bar{\beta}\|x^k - x^{k-1}\| + \|x^{k+1} - x^k\|)^2 + (L+l)\bar{\beta}^2\|x^k - x^{k-1}\|^2\\
&\le C(\|x^k - x^{k-1}\|^2 + \|x^{k+1} - x^k\|^2),
\end{aligned}$$

for some positive constant C, where the third inequality follows from the definition of y^k in (3.1) and the definition of $\bar{\beta}$. Combining this with the definition of $H_{k,\alpha}$, we obtain further that

$$0 \le H_{k+1,\alpha} - \zeta \le \eta(\|x^k - x^{k-1}\|^2 + \|x^{k+1} - x^k\|^2), \tag{3.29}$$

where $\eta = C + \alpha$, and the nonnegativity is a consequence of (3.28). On the other hand, let $\delta = \min\left\{\frac{L}{2} - \alpha, \alpha - \frac{L+l}{2}\bar{\beta}^2\right\}$. Then $\delta > 0$ and we see from (3.5) that

$$(H_{k+1,\alpha} - \zeta) - (H_{k,\alpha} - \zeta) \le -\delta(\|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2). \tag{3.30}$$

Combining (3.30) and (3.29), we obtain further that

$$(H_{k+1,\alpha} - \zeta) - (H_{k,\alpha} - \zeta) \le -\frac{\delta}{\eta}(H_{k+1,\alpha} - \zeta). \tag{3.31}$$

Reorganizing (3.31), we see that for all sufficiently large k,

$$0 \le H_{k+1,\alpha} - \zeta \le \frac{1}{1+\frac{\delta}{\eta}}(H_{k,\alpha} - \zeta),$$

which implies that the sequence $\{H_{k,\alpha}\}$ is Q-linearly convergent. This completes the proof. □

We are now ready to prove the local linear convergence of the sequences $\{x^k\}$ and $\{F(x^k)\}$, using the Q-linear convergence of $\{H_{k,\alpha}\}$.

THEOREM 3.7. *Suppose that $\bar\beta < \sqrt{\frac{L}{L+l}}$ and that Assumption 3.1 holds. Let $\{x^k\}$ be a sequence generated by Algorithm 1. Then the following statements hold.*
(i) *The sequence $\{x^k\}$ is R-linearly convergent to a stationary point of F.*
(ii) *The sequence $\{F(x^k)\}$ is R-linearly convergent.*

Proof. Fix any $\alpha \in (\frac{L+l}{2}\bar\beta^2, \frac{L}{2})$, which exists because $\bar\beta < \sqrt{\frac{L}{L+l}}$. Then, in view of Lemma 3.6, the sequence $\{H_{k,\alpha}\}$ is Q-linearly convergent. For notational simplicity, we denote its limit by ζ. Let $\delta = \min\left\{\frac{L}{2} - \alpha, \alpha - \frac{L+l}{2}\bar\beta^2\right\}$. Then $\delta > 0$ and we obtain from (3.5) that

$$\|x^{k+1} - x^k\|^2 \le \frac{1}{\delta}(H_{k,\alpha} - \zeta) - \frac{1}{\delta}(H_{k+1,\alpha} - \zeta) \le \frac{1}{\delta}(H_{k,\alpha} - \zeta), \tag{3.32}$$

where the last inequality follows from the fact that the sequence $\{H_{k,\alpha}\}$ is nonincreasing and convergent to ζ, thanks to Lemmas 3.1 and 3.3. Using the above inequality and the fact that the sequence $\{H_{k,\alpha}\}$ is Q-linearly convergent, we see that there exist $0 < c < 1$ and $M > 0$ such that

$$\|x^{k+1} - x^k\| \le Mc^k \tag{3.33}$$

for all k. Consequently, for any $m_2 > m_1 \ge 1$, we have

$$\|x^{m_2} - x^{m_1}\| \le \sum_{k=m_1}^{m_2-1} \|x^{k+1} - x^k\| \le \frac{Mc^{m_1}}{1-c},$$

showing that $\{x^k\}$ is a Cauchy sequence and hence convergent. Denoting its limit by $\hat x$ and passing to the limit as $m_2 \to \infty$ in the above relation, we see further that

$$\|x^{m_1} - \hat x\| \le \frac{Mc^{m_1}}{1-c}.$$

This means that the sequence $\{x^k\}$ is R-linearly convergent to its limit, which is a stationary point of F according to Lemma 3.4. This proves (i).

Next, we prove (ii). Notice that for any $k \ge 1$, we have from the definition of $H_{k,\alpha}$ that

$$\begin{aligned} |F(x^k) - \zeta| &= |H_{k,\alpha} - \zeta - \alpha\|x^k - x^{k-1}\|^2| \le H_{k,\alpha} - \zeta + \alpha\|x^k - x^{k-1}\|^2 \\ &\le H_{k,\alpha} - \zeta + \frac{\alpha}{\delta}(H_{k-1,\alpha} - \zeta), \end{aligned}$$

where the first inequality follows from the triangle inequality and the fact that the sequence $\{H_{k,\alpha}\}$ is nonincreasing and convergent to ζ according to Lemmas 3.1 and 3.3, and the second inequality follows from (3.32). This together with the Q-linear convergence of $\{H_{k,\alpha}\}$ and Lemma 2.1 implies the R-linear convergence of $\{F(x^k)\}$. □

3.3. FISTA with restart: a special case of Algorithm 1. In this subsection, we discuss FISTA with restart. Restart schemes for FISTA were proposed recently in O'Donoghue and Candès [12], where they adopted as a heuristic an adaptive restart technique, and established global linear convergence of the objective value when applying their method to (1.1) with f being strongly convex and $g = 0$. The restart techniques have also been adopted in the popular software, TFOCS [5]. While they did not prove any linear convergence results for convex nonsmooth problems such as the LASSO, they stated that for the LASSO, "after a certain number of iterations adaptive restarting can provide linear convergence"; see [12, Page 728]. In this subsection, we will explain that FISTA equipped with the aforementioned restart schemes is a special case of Algorithm 1. Moreover, when both of their restart schemes are used for the LASSO, both the sequences $\{x^k\}$ and $\{F(x^k)\}$ are R-linearly convergent.

To proceed, we first present FISTA [2, 25] for solving (1.1) with f being in addition convex.

```
FISTA  Input: $x^0 \in \mathrm{dom}\, g$, $\theta_{-1} = \theta_0 = 1$. Set $x^{-1} = x^0$.
    for $k = 0, 1, 2 \cdots$ do
        $\beta_k = \frac{\theta_{k-1} - 1}{\theta_k}$,
        $y^k = x^k + \beta_k(x^k - x^{k-1})$,
        $x^{k+1} = \mathrm{Prox}_{\frac{1}{L}g}\left(y^k - \frac{1}{L}\nabla f(y^k)\right)$,
        $\theta_{k+1} = \frac{1 + \sqrt{1 + 4\theta_k^2}}{2}$.
    end for
```

As one of the many variants of Nesterov's accelerated proximal gradient algorithms, FISTA uses a specific choice of $\{\beta_k\}$. According to the formula for updating β_k in FISTA above, it holds that $0 \leq \beta_k < 1$ for all k.[1] On the other hand, since f is convex, we can choose $l = 0$ and thus $\sqrt{\frac{L}{L+l}} = 1$ in Algorithm 1. Consequently, FISTA can be viewed as a special case of Algorithm 1.

FISTA with restart (see, for example, [5, 12]) is based on FISTA. Here, we adopt the same restart schemes as in [12]: fixed restart and adaptive restart. In the fixed restart scheme, we choose a positive integer K and reset $\theta_{k-1} = \theta_k = 1$ every K iterations, while in the adaptive restart (gradient scheme),[2] we reset $\theta_k = \theta_{k+1} = 1$ whenever $\langle y^k - x^{k+1}, x^{k+1} - x^k \rangle > 0$; see [12, Eq. 13]. Clearly, whenever the fixed restart scheme is invoked, we will have $\bar{\beta} < 1$. Thus, we have the following immediate corollary of Theorem 3.7.

Corollary 3.8. *Suppose that f in* (1.1) *is convex and Assumption 3.1 holds. Let $\{x^k\}$ be a sequence generated by FISTA with the fixed restart scheme or both the fixed and adaptive restart schemes. Then*

(i) *$\{x^k\}$ converges R-linearly to a globally optimal solution of* (1.1).

[1] Since $\theta_{k+1} = \left(1 + \sqrt{1 + 4\theta_k^2}\right)/2$ and $\theta_{-1} = \theta_0 = 1$ in FISTA, by induction, it is routine to show that $\theta_k \geq \frac{3}{2}$ and $\theta_{k-1} - 1 < \theta_k$ whenever $k \geq 1$. Combining these with the definition of β_k in FISTA, we see that $0 \leq \beta_k < 1$ for all k.

[2] There is also another scheme based on function values. It was discussed in [12, Section 3.2] that the two schemes perform similarly empirically and that the gradient scheme has advantages over the function value scheme. Thus, in this paper, we focus on the gradient scheme.

(ii) $\{F(x^k)\}$ *converges R-linearly to the globally optimal value of* (1.1).

From the discussion following Assumption 3.1, we see that the objective function in the LASSO satisfies Assumption 3.1. Thus, by Corollary 3.8, when the fixed or both the fixed and adaptive restart schemes are used for the LASSO, both the sequences $\{x^k\}$ and $\{F(x^k)\}$ are R-linearly convergent.

Before ending this subsection, we would like to point out two crucial differences between our Corollary 3.8 and the conclusion in [12]. First, they concluded *global* linear convergence of function values for a special case of (1.1) where f is strongly convex and $g = 0$, while we obtain *local* linear convergence for (1.1) for both $\{x^k\}$ and $\{F(x^k)\}$ with f being convex. Second, their global linear convergence is only guaranteed if K is chosen sufficiently large; see [12, Eq. 6]. On the other hand, we do not have any restrictions on the number K, the width of the restart interval.

4. Numerical experiments.

In this section, we conduct numerical experiments to study Algorithm 1 under different choices of $\{\beta_k\}$. We consider three different types of problems: the ℓ_1 regularized logistic regression problem, the LASSO, and the problem of minimizing a nonconvex quadratic function over a simplex. The first two problems are convex optimization problems, while the third problem is possibly nonconvex. We consider three different algorithms for each class of problems. For the convex problems, we consider Algorithm 1 with $\beta_k \equiv 0$ (proximal gradient algorithm), β_k chosen as in FISTA, and β_k chosen as in FISTA with both the fixed and the adaptive restart schemes. On the other hand, for the nonconvex problems, we consider Algorithm 1 with $\beta_k \equiv 0$ (proximal gradient algorithm) and $\beta_k \equiv 0.98\sqrt{\frac{L}{L+l}}$. We also consider FISTA as a heuristic.

All the numerical experiments are performed in Matlab 2014b on a 64-bit PC with an Intel(R) Core(TM) i7-4790 CPU (3.60GHz) and 32GB of RAM.

4.1. ℓ_1 regularized logistic regression.

In this subsection, we consider the ℓ_1 regularized logistic regression problem:

$$v_{\rm log} := \min_{\tilde{x}\in\mathbb{R}^n, x_0\in\mathbb{R}} \sum_{i=1}^m \log(1+\exp(-b_i(a_i^\top \tilde{x} + x_0))) + \lambda\|\tilde{x}\|_1, \tag{4.1}$$

where $a_i \in \mathbb{R}^n$, $b_i \in \{-1,1\}$, $i = 1, 2, \cdots, m$, with b_i not all the same, $m < n$ and $\lambda > 0$ is the regularization parameter. It is easy to see that (4.1) is in the form of (1.1) with

$$f(x) = \sum_{i=1}^m \log(1+\exp(-b_i(Dx)_i)), \qquad g(x) = \lambda\|\tilde{x}\|_1, \tag{4.2}$$

where $x := (\tilde{x}, x_0) \in \mathbb{R}^{n+1}$, and D is the matrix whose ith row is given by $(a_i^\top \;\; 1)$. Moreover, one can show that ∇f is Lipschitz continuous with modulus $0.25\lambda_{\max}(D^\top D)$. Thus, in our algorithms below, we take $L = 0.25\lambda_{\max}(D^\top D)$ and $l = 0$.

Before applying Algorithm 1, we need to show that $v_{\rm log} > -\infty$ and the solution set $\mathcal{X}$ of (4.1) is nonempty. To this end, we first recall that the dual problem of (4.1) is given by

$$\begin{array}{rl} \max\limits_{u\in\mathbb{R}^m} & d_{\rm log}(u) := -\sum_{i=1}^m [-b_i u_i \log(-b_i u_i) + (1+b_i u_i)\log(1+b_i u_i)] \\ \text{s.t.} & \|A^\top u\|_\infty \le \lambda, \;\; e^\top u = 0, \end{array} \tag{4.3}$$

where A is the matrix whose ith row is $a_i^\top$. It can be shown that the optimal values of (4.1) and (4.3) are the same, and that an optimal solution of (4.3) exists; see, for example, [6, Theorem 3.3.5]. In addition, we note that because $\lambda > 0$ and b_i are not all the same, the generalized Slater condition is satisfied for (4.3), i.e., there exists $\tilde{u}$ satisfying $\|A^\top \tilde{u}\|_\infty < \lambda$, $e^\top \tilde{u} = 0$ and $-1 < b_i\tilde{u}_i < 0$ for $i = 1, \ldots, m$. Hence, by [28, Corollary 28.2.2], an optimal solution of (4.1) exists. Consequently, $v_{\rm log} > -\infty$ and the solution set $\mathcal{X}$ of (4.1) is nonempty.

Thus, Algorithm 1 is applicable. In addition, from the discussion following Assumption 3.1, Assumption 3.1 is satisfied for (4.2). Hence, one should expect R-linear convergence of the sequences $\{x^k\}$ and $\{F(x^k)\}$ generated by FISTA with restart, in view of Corollary 3.8.

We now perform numerical experiments to study Algorithm 1 under three choices of $\{\beta_k\}$: $\beta_k \equiv 0$ as in the proximal gradient algorithm (PG), β_k chosen as in FISTA, and β_k chosen as in FISTA with both the fixed and the adaptive restart schemes, where we perform a fixed restart every 500 iterations (FISTA-R500). We choose $\lambda = 5$ in (4.1) and initialize all three algorithms at the origin. As for the termination, we make use of the fact that for any $\bar{x} \in \mathcal{X}$, $\nabla p(D\bar{x})$ is an optimal solution of (4.3) (see, for example, [28, Theorem 31.3]). Specifically, we define

$$u^k = \min\left\{1, \frac{\lambda}{\|A^\top \nabla p(Dx^k)\|_\infty}\right\} \nabla p(Dx^k),$$

and terminate the algorithms once the duality gap and the dual feasibility violation are small, i.e.,

$$\max\left\{\frac{|f(x^k) + g(x^k) - d_{\rm log}(u^k)|}{\max\{f(x^k) + g(x^k), 1\}}, \frac{50|e^T u^k|}{\max\{\|u^k\|, 1\}}\right\} \le 10^{-6}.$$

We also terminate the algorithms when the number of iterations hits 5000.

We consider random instances for our experiments. For each $(m, n, s) = (300, 3000, 30)$, $(500, 5000, 50)$ and $(800, 8000, 80)$, we generate an $m \times n$ matrix A with i.i.d. standard Gaussian entries. We then choose a support set T of size s uniformly at random, and generate an s-sparse vector $\hat{x}$ supported on T with i.i.d. standard Gaussian entries. The vector b is then generated as $b = \mathrm{sign}(A\hat{x} + ce)$, where c is chosen uniformly at random from $[0, 1]$.

Our computational results are presented in Figures 4.1, 4.2 and 4.3. In the plot (a) of each figure, we plot $\|x^k - x^*\|$ against the number of iterations, where x^* denotes the approximate solution obtained at termination of the respective algorithm; while in the plot (b) of each figure, we plot $|F(x^k) - F_{\min}|$ against the number of iterations, where $F_{\min}$ denotes the minimum of the three objective values obtained from the three algorithms. We see that both $\{x^k\}$ and $\{F(x^k)\}$ generated by FISTA with both fixed and adaptive restart schemes are R-linearly convergent, which conforms with our theory. Moreover, compared with FISTA and the proximal gradient algorithm, the algorithm with restart performs better.

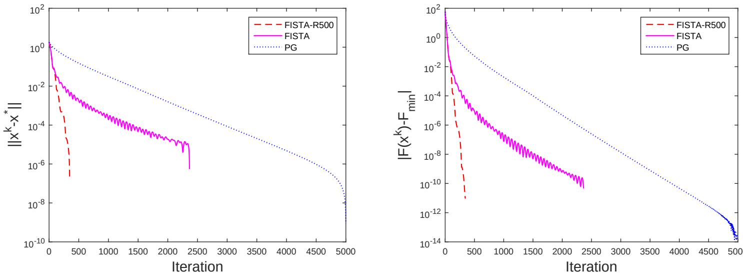
4.2. LASSO. In this subsection, we consider the LASSO:

$$v_{\rm ls} := \min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1, \tag{4.4}$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. We observe that (4.4) is in the form of (1.1) with

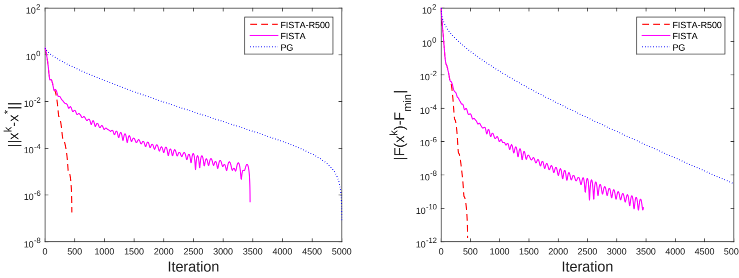
$$f(x) = \frac{1}{2}\|Ax - b\|^2, \quad g(x) = \lambda\|x\|_1. \tag{4.5}$$

Fig. 4.1. $l_1 - logistic: \; n = 3000, m = 300, s = 30$



(a) (b)

Fig. 4.2. $l_1 - logistic: \; n = 5000, m = 500, s = 50$



(a) (b)

It is clear that f has a Lipschitz continuous gradient and $f + g$ has compact lower level sets. Thus, we can apply Algorithm 1 to solving (4.4). Moreover, in view of the discussion following Assumption 3.1, Assumption 3.1 is satisfied for (4.5). Hence, according to Corollary 3.8, one should observe R-linear convergence of both the sequences $\{x^k\}$ and $\{F(x^k)\}$ generated by FISTA with restart. Finally, it is not hard to show that ∇f has a Lipschitz continuity modulus of $\lambda_{\max}(A^\top A)$. In view of this, in the algorithms below, we take $L = \lambda_{\max}(A^\top A)$ and $l = 0$.

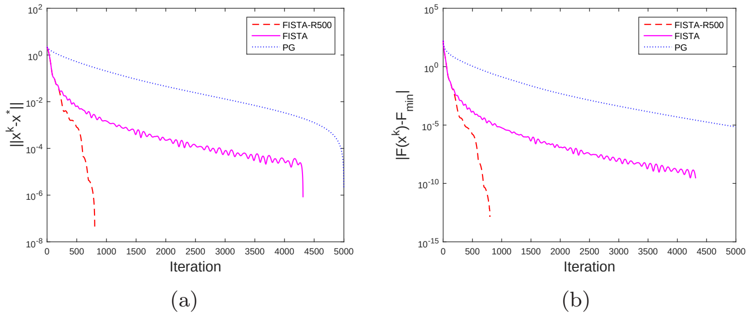
Before describing our numerical experiments, we recall that $f(x) = h(Ax) = \frac{1}{2}\|Ax - b\|^2$, where $h(v) = \frac{1}{2}\|v - b\|^2$. The conjugate function of h can then be easily computed as $h^*(u) := \sup_{v\in\mathbb{R}^m}\{u^\top v - h(v)\} = \frac{1}{2}\|u\|^2 + b^\top u$. Hence, the dual problem of (4.4) is given by

$$\begin{array}{rl} \max\limits_{u\in\mathbb{R}^m} & d_{\rm ls}(u) := -\frac{1}{2}\|u\|^2 - b^\top u \\ \text{s.t.} & \|A^\top u\|_\infty \le \lambda. \end{array} \tag{4.6}$$

It can be shown that the optimal values of (4.4) and (4.6) are the same, and moreover, an optimal solution of (4.6) exists; see, for example, [6, Theorem 3.3.5]. This dual problem will be used in developing termination criterion for our algorithms below.

Now we perform numerical experiments to study Algorithm 1 under the same three choices of $\{\beta_k\}$ as in the previous subsection. We choose $\lambda = 5$ in (4.4), initialize all three algorithms at the origin and use the duality gap to terminate the algorithms. Specifically, as in the previous subsection, we make use of the fact that for any optimal

FIG. 4.3. $l_1 - logistic: \ n = 8000, m = 800, s = 80$



(a) (b)

solution $\bar{x}$ of (4.4), $\nabla h(A\bar{x})$ is an optimal solution of (4.6). Hence, we define

$$u^k = \min\left\{1, \frac{\lambda}{\|A^\top \nabla h(Ax^k)\|_\infty}\right\} \nabla h(Ax^k),$$

and terminate the algorithms once the duality gap is small, i.e.,

$$\frac{|f(x^k) + g(x^k) - d_{1s}(u^k)|}{\max\{f(x^k) + g(x^k), 1\}} \leq 10^{-6}.$$

We also terminate them when the number of iterations hits 5000.

The problems used in our experiments are generated as follows. For each $(m, n, s) = (300, 3000, 30)$, $(500, 5000, 50)$ and $(800, 8000, 80)$, we generate an $m \times n$ matrix A with i.i.d. standard Gaussian entries. We then choose a support set T of size s uniformly at random, and generate an s-sparse vector $\hat{x}$ supported on T with i.i.d. standard Gaussian entries. The vector b is then generated as $b = A\hat{x} + 0.01\tilde{e}$, where $\tilde{e}$ has standard i.i.d. Gaussian entries.

The computational results are presented in Figures 4.4, 4.5 and 4.6. We plot $\|x^k - x^*\|$ against the number of iterations in part (a) of each figure, where x^* denotes the approximate solution obtained at termination of the respective algorithm; additionally, we plot $|F(x^k) - F_{\min}|$ against the number of iterations in part (b) of each figure, where $F_{\min}$ denotes the minimum of the three objective values obtained from the three algorithms. As in the previous subsection, we see from the figures that both $\{x^k\}$ and $\{F(x^k)\}$ generated by FISTA with both fixed and adaptive restart schemes are R-linearly convergent, which conforms with our theory. Additionally, the algorithm with restart performs better than FISTA and the proximal gradient algorithm.

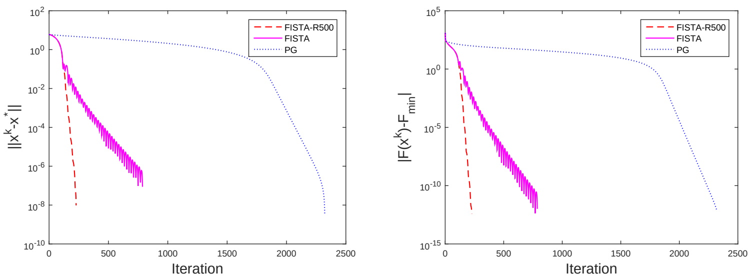
4.3. Nonconvex quadratic programming with simplex constraints.

In this subsection, we look at problems of the following form, which are possibly nonconvex:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \ & \frac{1}{2} x^\top A x - b^\top x \\ \text{s.t.} \ & e^\top x = s, \quad x \geq 0, \end{aligned} \tag{4.7}$$

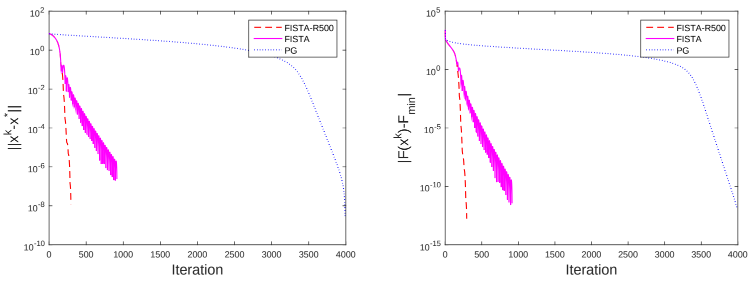
where $A \in \mathbb{R}^{n \times n}$ is a symmetric matrix that is not necessarily positive semidefinite, $b \in \mathbb{R}^n$ and s is a positive number. This is an example of nonconvex quadratic

Fig. 4.4. $l_1 - ls: \;\; n = 3000, m = 300, s = 30$



(a) (b)

Fig. 4.5. $l_1 - ls: \;\; n = 5000, m = 500, s = 50$



(a) (b)

programming problems, which is an important class of problems in global optimization [11, 14, 15, 21]. Notice that one can rewrite (4.7) in the form of (1.1) by defining

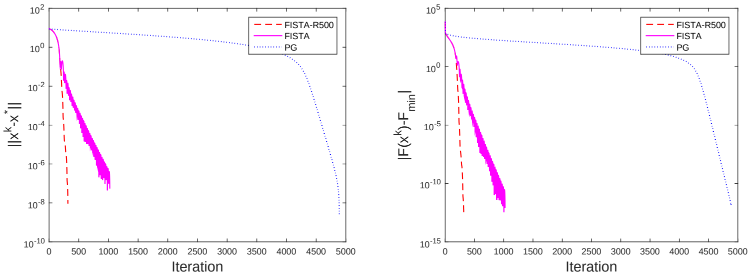
$$f(x) = \frac{1}{2}x^\top Ax - b^\top x, \qquad g(x) = \delta_{\mathcal{S}}(x), \tag{4.8}$$

where $\mathcal{S} = \left\{x \in \mathbb{R}^n : \; e^\top x = s, \; x \geq 0\right\}$. Moreover, it is clear that f has a Lipschitz continuous gradient and $f + g$ is level bounded. Hence, Algorithm 1 can be applied to solving (4.7). Furthermore, from the discussion following Assumption 3.1, Assumption 3.1 is satisfied for (4.8). Consequently, according to Theorem 3.7, one should expect to see R-linear convergence of both the sequences $\{x^k\}$ and $\{F(x^k)\}$ generated by Algorithm 1 when $\bar\beta < \sqrt{\frac{L}{L+l}}$. Finally, since $A = A_1 - A_2$, where A_1 and $-A_2$ are the projections of A onto the cone of positive semidefinite matrices and the cone of negative semidefinite matrices, respectively, we see that $f = f_1 - f_2$, where $f_1(x) = \frac{1}{2}x^\top A_1 x - b^\top x$ and $f_2(x) = \frac{1}{2}x^\top A_2 x$. In view of this, in our experiments below, we set $L = \max\{\lambda_{\max}(A), |\lambda_{\min}(A)|\}$ and $l = |\lambda_{\min}(A)|$ so that L and l are the Lipschitz continuity moduli of ∇f_1 and ∇f_2, respectively, and $L \geq l$.

Now we perform numerical experiments to study Algorithm 1 with two choices of $\{\beta_k\}$: $\beta_k \equiv 0$ (PG) and $\beta_k \equiv 0.98\sqrt{\frac{L}{L+l}}$ (PG$_\mathrm{e}$). In addition, we also perform the same experiments on FISTA.[3] We initialize all three algorithms at the origin, and

[3] We would like to point out that FISTA applied to the nonconvex problem (4.7) is not known to

Fig. 4.6. $l_1 - ls:\ \ n = 8000, m = 800, s = 80$



(a) (b)

terminate them when the successive changes of the iterates are small, i.e.,

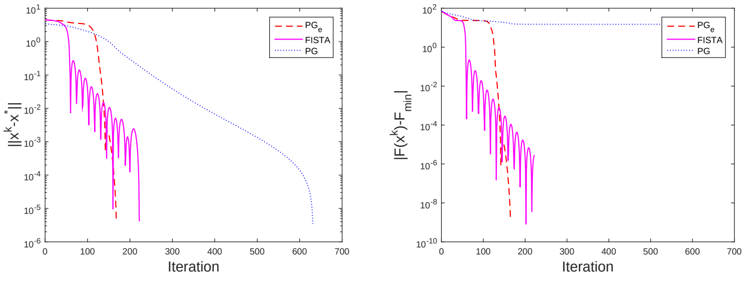
$$\frac{\|x^k - x^{k-1}\|}{\max\{\|x^k\|, 1\}} \leq 10^{-6}.$$

We also terminate the algorithms when the number of iterations hits 5000.

Our test problem is generated as follows. We generate a 2000×2000 matrix D with i.i.d. standard Gaussian entries. We then generate a symmetric matrix $A = D + D^\top$. Finally, the vector b is generated with i.i.d. standard Gaussian entries, and s is generated as $\max\{1, 10t\}$, with t chosen uniformly at random from $[0, 1]$.

The computational results are presented in Figure 4.7. We plot $\|x^k - x^*\|$ against the number of iterations in Figure 4.7 (a), where x^* denotes the approximate solution obtained at termination of the respective algorithm; in addition, we plot $|F(x^k) - F_{\min}|$ against the number of iterations in Figure 4.7 (b), with $F_{\min}$ being the minimum of the three objective values obtained from the three algorithms. We can see from Figure 4.7 (a) that the sequence $\{x^k\}$ generated by Algorithm 1 with $\beta_k \equiv 0.98\sqrt{\frac{L}{L+l}}$ is R-linearly convergent, which conforms with our theory. However, from Figure 4.7 (b), one can see that not all the algorithms are approaching $F_{\min}$. This is likely because the iterates generated by the algorithm got stuck at local minimizers.

Fig. 4.7. *Nonconvex Quadratic Problem*



(a) (b)

converge, unlike the other two algorithms which have convergence guarantee by our theory.

To further evaluate the quality (in terms of function values at termination) of the approximate solution obtained by the algorithms, we perform a second experiment. In this second experiment, we generate random instances as follows: we generate an $n \times n$ matrix D with i.i.d. standard Gaussian entries and symmetrize it to form $A = D + D^\top$; moreover, we generate a vector b with i.i.d. standard Gaussian entries, and an $s = \max\{1, 10t\}$, where t is chosen uniformly at random from $[0, 1]$.

In our test, for each $n = 500$, 1000, 1500, 2000 and 2500, we generate 50 random instances as described above. The computational results are reported in Table 4.1, where we present the number of iterations averaged over the 50 instances for each n (iter), and the function value at termination (fval), also averaged over the 50 instances. One can see that while Algorithm 1 with $\beta_k \equiv 0.98\sqrt{\frac{L}{L+l}}$ (i.e., $\mathrm{PG_e}$) is always the fastest algorithm, the function values obtained can be slightly compromised for some instances.

Table 4.1
Comparing $\mathrm{PG_e}$*, FISTA and PG on random instances.*

	$\mathrm{PG_e}$		FISTA		PG	
n	iter	fval	iter	fval	iter	fval
500	120	−56.02	175	−56.90	322	−57.96
1000	171	−69.77	274	−66.79	636	−66.93
1500	166	−66.29	270	−63.71	560	−65.29
2000	215	−80.72	271	−80.43	635	−81.21
2500	284	−81.70	359	−80.13	813	−83.81

5. Conclusion.

In this paper, we study the proximal gradient algorithm with extrapolation for solving a class of nonconvex nonsmooth optimization problems. Based on the error bound condition, we establish the R-linear convergence of both the sequence $\{x^k\}$ generated by the algorithm and the corresponding sequence of objective values $\{F(x^k)\}$ if the extrapolation coefficients are below the threshold $\sqrt{\frac{L}{L+l}}$. We further demonstrate that our theory can be applied to analyzing the convergence of FISTA with the fixed restart scheme for convex problems. Finally, we perform some numerical experiments to illustrate our results.

Acknowledgments. We would like to thank the anonymous referees for their insightful comments and helpful suggestions that helped improve the manuscript.

REFERENCES

[1] H. Attouch and Z. Chbani. Fast inertial dynamics and FISTA algorithms in convex optimization. Perturbation aspects. *arXiv preprint arXiv:* 1507.01367v1.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2: 183–202, 2009.

[3] A. Beck and M. Teboulle. A linearly convergent dual-based gradient projection algorithm for quadratically constrained convex minimization. *Mathematics of Operations Research*, 31: 398–417, 2006.

[4] S. Becker, J. Bobin, and E.J. Candès. NESTA: a fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4: 1–39, 2011.

[5] S. Becker, E.J. Candès, and M.C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3: 165–218, 2011.

[6] J.M. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization.* 2nd edition, Springer, 2006.

[7] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9: 717–772, 2009.

[8] E.J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51: 4203–4215, 2005.

[9] A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20: 89–97, 2004.

[10] A. Chambolle and Ch. Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization Theory and Applications*, 166: 968–982, 2015.

[11] X. Chen, J. Peng, and S. Zhang. Sparse solutions to random standard quadratic optimization problems. *Mathematical programming, Series A*, 141: 273–293, 2013.

[12] B. O'Donoghue and E.J. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15: 715–732, 2015.

[13] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52: 1289–1306, 2006.

[14] L.E. Gibbons, D.W. Hearn, P.M. Pardalos, and M.V. Ramana. Continuous characterizations of the maximal clique problem. *Mathematics of Operations Research*, 22: 754–768, 1997.

[15] T. Ibaraki and N. Katoh. *Resource Allocation Problems: Algorithmic Approaches.* MIT Press, Cambridge, 1988.

[16] P.R. Johnstone and P. Moulin. Local and global convergence of an inertial version of forward-backward splitting. *arXiv preprint arXiv*: 1502.02281v4.

[17] P.L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16: 964–979, 1979.

[18] Z.-Q. Luo and P. Tseng. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30: 408–425, 1992.

[19] Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46: 157–178, 1993.

[20] Z.-Q. Luo and P. Tseng. On the convergence rate of dual ascent methods for linearly constrained convex minimization. *Mathematics of Operations Research*, 18: 846–867, 1993.

[21] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7: 77–91, 1952.

[22] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(\frac{1}{k^2})$. *Soviet Mathematics Doklady*, 27: 372–376, 1983.

[23] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Kluwer Academic Publishers, Boston, 2004.

[24] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming, Series A*, 103: 127–152, 2005.

[25] Y. Nesterov. Gradient methods for minimizing composite objective function. *CORE Discussion Paper*, 2007.

[26] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical programming, Series B*, 109: 319–344, 2007.

[27] N. Parikh and S. Boyd. Proximal Algorithms. *Foundations and Trends in Optimization*, 1: 123–231, 2013.

[28] R.T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, 1970.

[29] R.T. Rockafellar and R.J-B. Wets. *Variational Analysis.* Springer, 1998.

[30] P. Tseng and S. Yun. A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. *Computational Optimization and Applications*, 47: 179–206, 2010.

[31] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical programming, Series B*, 117: 387-423, 2009.

[32] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming, Series B*, 125: 263-295, 2010.

[33] S. Tao, D. Boley, and S. Zhang. Local linear convergence of ISTA and FISTA on the LASSO problem. *SIAM Journal on Optimization*, 26: 313-336, 2016.

[34] Z. Zhou, and A. M.-C. So. A unified approach to error bounds for structured convex optimization problems. *arXiv preprint arXiv*: 1512.03518v1.