Gradient methods for convex minimization: better rates under weaker conditions

Hui Zhang* Wotao Yin†

March 20, 2013

Abstract

The convergence behavior of gradient methods for minimizing convex differentiable functions is one of the core questions in convex optimization. This paper shows that their well-known complexities can be achieved under conditions weaker than the commonly accepted ones. We relax the common gradient Lipschitz-continuity condition and strong convexity condition to ones that hold only over certain line segments. Specifically, we establish complexities $O(\frac{R}{\epsilon})$ and $O(\sqrt{\frac{R}{\epsilon}})$ for the ordinary and accelerate gradient methods, respectively, assuming that ∇f is Lipschitz continuous with constant R over the line segment joining x and $x - \frac{1}{R}\nabla f$ for each $x \in \mathrm{dom} f$. Then we improve them to $O(\frac{R}{\nu}\log(\frac{1}{\epsilon}))$ and $O(\sqrt{\frac{R}{\nu}}\log(\frac{1}{\epsilon}))$ for function f that also satisfies the secant inequality $\langle \nabla f(x), x - x^*\rangle \geq \nu\|x - x^*\|^2$ for each $x \in \mathrm{dom} f$ and its projection x^* to the minimizer set of f. The secant condition is also shown to be necessary for the geometric decay of solution error. Not only are the relaxed conditions met by more functions, the restrictions give smaller R and larger ν than they are without the restrictions and thus lead to better complexity bounds. We apply these results to sparse optimization and demonstrate a faster algorithm.

Keywords: sublinear convergence, linear convergence, restricted Lipschitz continuity, restricted strong convexity, Nesterov acceleration, restart technique, skipping technique, sparse optimization.

1 Introduction

Owing much to the fast development in signal/image processing, compressive sensing, statistical and machine learning, and parallel computing, we have witnessed the (revived) popularity of gradient methods, which are easy to program, have relatively low per-iteration complexities, and are often among the best options for obtaining moderately accurate solutions for large-scale optimization problems.

This paper considers the convex unconstrained optimization problem:

$$f^* := \min_{x\in\mathbb{R}^n} f(x) \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable convex function. We assume throughout the paper that the set of optimal solutions $\mathcal{X}^*$ is nonempty and closed and thus $f^* \in \mathbb{R}$ is attainable. For simplicity, we assume $\mathrm{dom} f = \mathbb{R}^n$. Most of the discussions in this paper hold if we impose $x \in \mathrm{dom} f$ rather than $x \in \mathbb{R}^n$.

*Department of Mathematics and Systems Science, College of Science, National University of Defense Technology, Changsha, Hunan, China. Email: hhuuii.zhang@gmail.com

†Department of Computational and Applied Mathematics, Rice University, Houston, Texas, US. Email: wotao.yin@rice.edu

The gradient descent iteration is

$$x^{(k+1)} = x^{(k)} - \tau \nabla f(x^{(k)}). \tag{2}$$

Its convergence rates have been established for two major classes of functions [6, 7, 8]: The first class, denoted by $\mathcal{F}_L(\mathbb{R}^n)$, consists of the convex functions with Lipschitz continuous gradients, namely,

$$\begin{aligned} f \in \mathcal{F}_L(\mathbb{R}^n) \iff & f \text{ is differentiable and} \\ & \|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n, \end{aligned} \tag{3}$$

where $L > 0$ is the Lipschitz constant of ∇f; the second class, denoted by $\mathcal{S}_{L,\mu}(\mathbb{R}^n)$, is a subclass of $\mathcal{F}_L(\mathbb{R}^n)$ in which the functions are also strongly convex, namely,

$$\begin{aligned} f \in \mathcal{S}_{\mu,L}(\mathbb{R}^n) \iff & f \in \mathcal{F}_L(\mathbb{R}^n) \text{ and} \\ & \langle \nabla f(x) - \nabla f(y), x - y \rangle \ge \mu \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n, \end{aligned} \tag{4}$$

where $\mu > 0$ is the convex modulus of f. Geometrically, if $f \in \mathcal{F}_L$, ∇f cannot change too quickly; the curvature of f (assuming $f \in C^2$) is upper bounded by L. If $f \in S_{\mu,L}$, ∇f cannot change too slowly either; the curvature of f (assuming $f \in C^2$) is lower bounded by μ. One might be more familiar certain equivalent conditions of (3) and (4).

function class	1st-order oracle lower bound	ordinary gradient method	accelerated gradient method
$\mathcal{F}_L(\mathbb{R}^n)$	$O\left(\sqrt{\frac{L}{\epsilon}}\right)$	$O\left(\frac{L}{\epsilon}\right)$	$O\left(\sqrt{\frac{L}{\epsilon}}\right)$
$\mathcal{S}_{L,\mu}(\mathbb{R}^n)$	$O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	$O\left(\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$	$O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$

Table 1: Complexities of minimizing a convex differentiable function to ϵ-accuracy

For any $f \in \mathcal{F}_L$, iteration (2) reduces $f^k = f(x^{(k)})$ at the rate of $O(\frac{L}{k})$; hence, it takes $O(\frac{L}{\epsilon})$ iterations to guarantee $f^k < f^* + \epsilon$. For any $f \in S_{\mu,L}$, the rate is improved to $O(\frac{L-\mu}{L+\mu})^{2k}$. Therefore, it only takes $O(\frac{L}{\mu}\log(\frac{1}{\epsilon}))$ iterations.

In the seminal paper [6], Nesterov presents an accelerated gradient descent iteration. For functions in $\mathcal{F}_L$, its complexity is $O(\sqrt{\frac{L}{\epsilon}})$. In papers [7, 9], he generalizes the method to more function classes. In particular, if $f \in S_{\mu,L}$, the complexity is $O(\sqrt{\frac{L}{\mu}}\log(\frac{1}{\epsilon}))$. He gives examples of functions on which no gradient-based methods can perform fundamentally better. So, his method has the optimal worst-case complexities; for more detail, see book [8]. The complexities discussed above are summarized in Table 1.

1.1 Contributions

We show that global Lipschitz continuity of ∇f is not necessary for deriving the sublinear bounds in Table 1. If ∇f is Lipschitz continuous with constant $R > 0$ restricted to the line segments joining x and $x - (1/R)\nabla f(x)$, for $x = x^{(0)}, x^{(1)}, \ldots$, or simply $x \in \mathbb{R}^n$, then the ordinary and accelerated gradient descent methods have complexities $O(R/\epsilon)$ and $O(\sqrt{R/\epsilon})$, respectively. We believe that some researchers, especially those who study line search methods, might be aware of this result though we do not find it in the literature.

Our analysis in fact hints a backtracking line search method that achieves the same complexities without the knowledge of R. It is worth noting that the recent paper [11] presents a skillful line search method that improves the Nesterov's accelerated gradient method.

On the other hand, the Lipschitz continuity of ∇f alone gives at best the rather weak $O(1/\epsilon)$ and $O(1/\sqrt{\epsilon})$ complexities. It is commonly know that the strong convexity of f enables the much better complexity of $O(\log(1/\epsilon))$. However, most convex functions are not strongly convex. Hence, it is interesting to relax the conditions and still establish a linear convergence rate. We show that an inequality resembling (4) but concerning just the secant between x and its projection to $\mathcal{X}^*$ is ultimately responsible for linear convergence. The inequality imposes a positive lower bound on the *average curvature* between x and the solution set and is shown to be both sufficient and necessary for the geometric decay of solution error.

1.2 Outline of the paper

The rest of the paper is organized as follows. Section 2 defines new properties of functions along with examples and discussions. Section 3 describes the convergence and complexity results. Section 4 applies these results to the augmented ℓ_1 model and presents numerical results of sparse signal recovery. Finally, Section 5 concludes this paper.

2 Weakened conditions

For any two vector $u, v \in \mathbb{R}^n$, we let the set of points on the line segment between u and v be denoted by $\lfloor u, v \rfloor$, i.e.,

$$\lfloor u, v \rfloor = \{w \in \mathbb{R}^n : w = \lambda u + (1-\lambda)v, 0 \le \lambda \le 1\}.$$

Definition 1 (Restricted Lipschitz-continuous gradient – RLG(R))**.** *A function $f(x) : \mathbb{R}^n \to \mathbb{R}$ has a restricted Lipschitz-continuous gradient (RLG) with constant $R \ge 0$ if it is differentiable and obeys*

$$\|\nabla f(x) - \nabla f(y)\| \le R\|x - y\|, \quad \forall (x, y) \in \Omega, \tag{5}$$

where

$$\Omega = \bigcup_{z \in \mathbb{R}^n} \{(x, y) : x, y \in \lfloor z, z - (1/R)\nabla f(z) \rfloor\}. \tag{6}$$

This definition requires ∇f not to change too quickly over the specified downhill line segments (6). Constant R can generally be smaller than the global Lipschitz constant L.

Definition 2 (Restricted secant inequality – RSI(ν))**.** *A function $f(x) : \mathbb{R}^n \to \mathbb{R}$ satisfies the restricted secant inequality (RSI) with constant $\nu > 0$ if it is differentiable and obeys*

$$\langle \nabla f(x) - \nabla f(x_{\text{prj}}), x - x_{\text{prj}} \rangle \ge \nu \|x - x_{\text{prj}}\|^2, \tag{7}$$

where $x_{\text{prj}} = \text{Proj}_{\mathcal{X}^*}(x)$ is the projection of x onto the solution set $\mathcal{X}^*$. Such f is called an RSI function.

Note that $\nabla f(x_{\text{prj}}) = 0$ by definition. Constant ν can be viewed as a lower bound of the average curvature of f between x and x_{prj}. Since the goal of minimization is to reach the solution set $\mathcal{X}^*$, in order to have linear convergence, it turns out only the "average minimum curvature" between the current x and its projection x_{prj} matters. Using RSI, we introduce restricted strongly convex (RSC) functions.

Definition 3 (Restricted strong convexity – RSC(ν))**.** *A function $f(x) : \mathbb{R}^n \to \mathbb{R}$ is restricted strongly convex with constant $\nu > 0$ if it is convex, has a finite minimizer, and satisfies RSI(ν).*

RSC is weaker than strong convexity as (7) is a relaxation to inequality (4). Some of our convergence results will be given for the following new classes of functions.

Definition 4 (New function classes)**.** *Let $R, \nu > 0$. Define function classes*

$$\begin{aligned}\mathcal{L}_R(\mathbb{R}^n) &:= \{f : \mathbb{R}^n \to \mathbb{R} \mid f \text{ is convex and } RLG(R)\},\\ \mathcal{R}_{R,\nu}(\mathbb{R}^n) &:= \{f \in \mathcal{L}_R(\mathbb{R}^n) \mid f \text{ is } RSC(\nu)\},\\ \hat{\mathcal{R}}_{L,\nu}(\mathbb{R}^n) &:= \{f \in \mathcal{F}_L(\mathbb{R}^n) \mid f \text{ is } RSC(\nu)\}.\end{aligned}$$

By definition, if $\mu \geq \nu$ and $L = R$, then we have

$$\begin{array}{ccccc}\mathcal{S}_{L,\mu}(\mathbb{R}^n) & \subset & \mathcal{R}_{R,\nu}(\mathbb{R}^n) & \subset & \\ & \subset & \hat{\mathcal{R}}_{L,\nu}(\mathbb{R}^n) \subset \mathcal{F}_L(\mathbb{R}^n) & \subset & \mathcal{L}_R(\mathbb{R}^n).\end{array}$$

Definition 3 is different from another recent definition of restricted strong convexity from [5].

Definition 5 (Restricted strong convexity of [5])**.** *A function $f(x) : \mathbb{R}^n \to \mathbb{R}$ satisfies the restricted strong convexity at x_0 with constants $\kappa_1, \kappa_2 > 0$ and tolerance function $r(x)$ if it is differentiable and*

$$f(x_0 + \delta) - f(x_0) - \langle f'(x_0), \delta\rangle \geq \kappa_1\|\delta\|^2 - \kappa_2(r(x_0))^2, \tag{8}$$

for all $\delta \in \mathcal{C}$, where $\mathcal{C}$ is a certain point set.

Definition 5 is a local and weakened version of strong convexity. With $r(x) = 0$ and $\mathbb{C} = \mathbb{R}^n$, it reduces to the standard strong convexity.

Many of the recent algorithms for sparse optimization are observed to converge quickly, at least on problems that are not severely "ill-conditioned"; however, their underlying objective functions are not strongly convex – a property commonly used to ensure global linear convergence. When A has more columns than rows, a function in the form of $g(Ax - b)$, even with a strongly convex function g, is "flat" along many directions. Gradients along these directions are small, so minimization can progress very slowly. However, in problems with certain types of A and an additional regularization function $r(x)$ such as the ℓ_1-norm, moving along these directions will significantly change $r(x)$. We believe this has the definition of restricted strong convexity in [1], which extends the ordinary definition by including the relaxation term involving $r(x)$. That paper argues that, with high probability for problems with A that is random or satisfies certain restricted eigenvalue properties, Definition 5 is satisfied by $f(x) = g(Ax - b) + r(x)$, and as a result, the prox-linear or gradient-projection iteration has a (nearly-)linear convergence behavior, specifically,

$$\|x^{(k+1)} - x^*\|^2 \leq c^k\|x^{(0)} - x^*\|^2 + o(\|x^* - x^o\|^2),$$

where $c < 1$, x^* and x^o are the minimizer and underlying true signal, respectively, and $x^{(k)}$ stands for the kth iterate. Our paper focuses on the minimization of convex differentiable functions in the general setting and establishes unmodified sublinear and linear convergence without a probabilistic argument.

2.1 Properties

This subsection gives the core lemmas for establishing the main convergence results.

Lemma 1. *Let $\mathcal{X}^*$ be the nonempty solution set of (1). If $f \in \mathcal{L}_R(\mathbb{R}^n)$ with $R > 0$, then we have*

1) For any $(x, y) \in \Omega$ given in (6), it holds

$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{R}{2}\|x - y\|^2; \tag{9}$$

2) For any $y \in \mathcal{X}^*$, it holds

$$\frac{1}{2R}\|\nabla f(x)\|^2 \leq \langle \nabla f(x), x - y\rangle. \tag{10}$$

Proof. For any $(x, y) \in \Omega$, (9) follows from

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \langle \nabla f(x + \tau(y - x)), y - x\rangle d\tau \\ &= f(x) + \langle \nabla f(x), y - x\rangle + \int_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x\rangle d\tau \\ &\leq f(x) + \langle \nabla f(x), y - x\rangle + \int_0^1 \|\nabla f(x + \tau(y - x)) - \nabla f(x)\|\|y - x\| d\tau \\ &\leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{R}{2}\|x - y\|^2, \end{aligned}$$

where the first inequality follows from the Cauchy-Schwartz inequality and the second one follows from the definition of RLG. For part 2), for any $y \in \mathcal{X}^*$ we have

$$\begin{aligned} f^* = f(y) &\leq f(x - R^{-1}\nabla f(x)) \\ &\leq f(x) + \langle \nabla f(x), (x - R^{-1}\nabla f(x)) - x\rangle + \frac{R}{2}\|(x - R^{-1}\nabla f(x)) - x\|^2 \\ &= f(x) - (2R)^{-1}\|\nabla f(x)\|^2, \end{aligned}$$

where the second inequality follows from part 1). Therefore, we have

$$\frac{1}{2R}\|\nabla f(x)\|^2 \leq f(x) - f(y) \leq \langle \nabla f(x), x - y\rangle,$$

where the second inequality utilizes the convexity of f. □

Note that for general y, the inequality (10) does not hold. For example, setting $y = x - \eta\nabla f(x)$ with $0 < \eta < \frac{1}{2R}$ and assuming $\nabla f(x) \neq 0$ give $\langle \nabla f(x), x - y\rangle = \eta \cdot \|\nabla f(x)\|^2 < \frac{1}{2R}\|\nabla f(x)\|^2$.

Lemma 2. *Let $\mathcal{X}^*$ be the nonempty solution set of (1). If $f \in \mathcal{R}_{R,\nu}(\mathbb{R}^n)$ with $R > 0$ and $\nu > 0$, then for every $\theta \in [0, 1]$ the following holds:*

$$\langle \nabla f(x) - \nabla f(x_{\text{prj}}), x - x_{\text{prj}}\rangle \geq \frac{\theta}{2R}\|\nabla f(x) - \nabla f(x_{\text{prj}})\|^2 + (1 - \theta)\nu\|x - x_{\text{prj}}\|^2, \tag{11}$$

where x_{prj} is the projection of x onto the solution set $\mathcal{X}^*$.

Proof. Obviously, $x_{\text{prj}} \in \mathcal{X}^*$ and $\nabla f(x_{\text{prj}}) = 0$. Thus, from part 2) of Lemma 1, we have

$$\langle \nabla f(x) - \nabla f(x_{\text{prj}}), x - x_{\text{prj}}\rangle \geq \frac{1}{2R}\|\nabla f(x) - \nabla f(x_{\text{prj}})\|^2. \tag{12}$$

On the other hand, from the definition of RSC(ν), we obtain

$$\langle \nabla f(x) - \nabla f(x_{\text{prj}}), x - x_{\text{prj}}\rangle \geq \nu\|x - x_{\text{prj}}\|^2. \tag{13}$$

Inequality (11) follows from (12) and (13). □

Parameter θ in (11) will be optimized to obtain a convergence bound.

Lemma 3. *Let $f(x)$ satisfy RSI(ν), $\nu > 0$, and $\mathcal{X}^*$ be the nonempty solution set. For $\forall x \in \mathbb{R}^m$ we have*

$$f(x) - f(x_{\text{prj}}) \geq \frac{\nu}{2}\|x - x_{\text{prj}}\|^2, \tag{14}$$

where x_{prj} is the projection of x onto the solution set $\mathcal{X}^*$.

Proof. Since for any $\tau \in [0,1]$ point $y_\tau = x_{\text{prj}} + \tau(x - x_{\text{prj}}) \in \lfloor x, x_{\text{prj}} \rfloor$ projects to $\mathcal{X}^*$ at x_{prj}, we have

$$f(x) = f(x_{\text{prj}}) + \int_0^1 \langle \nabla f(x_{\text{prj}} + \tau(x - x_{\text{prj}})), x - x_{\text{prj}} \rangle d\tau \tag{15a}$$

$$= f(x_{\text{prj}}) + \int_0^1 \frac{1}{\tau} \langle \nabla f(x_{\text{prj}} + \tau(x - x_{\text{prj}})) - \nabla f(x_{\text{prj}}), \tau(x - x_{\text{prj}}) \rangle d\tau \tag{15b}$$

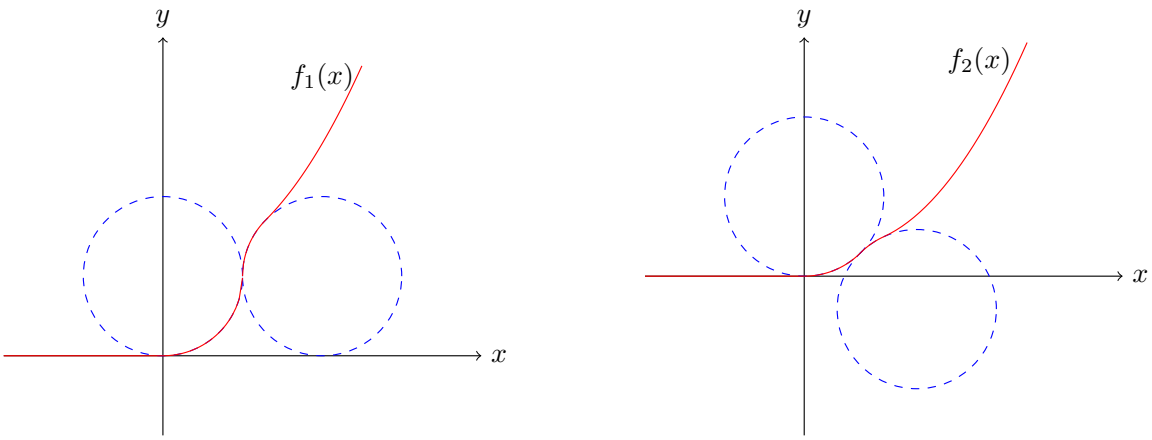
$$\geq f(x_{\text{prj}}) + \int_0^1 \frac{1}{\tau} \nu \tau^2 \|x - x_{\text{prj}}\|^2 d\tau \tag{15c}$$

$$= f(x_{\text{prj}}) + \frac{\nu}{2} \|x - x_{\text{prj}}\|^2 \tag{15d}$$

where (15b) follows from $\nabla f(x_{\text{prj}}) = 0$ and (15c) from RSI(ν). □

It is worth noting that since x_{prj} is restricted, inequality (14) does not mean that f grows *everywhere* quicker than the quadratic function $q(x) = \frac{\nu}{2}\|x - x_{\text{prj}}\|^2$.

2.2 Examples of RSI and RSC functions



(a) (b)

Figure 1: Non-convex functions satisfying RSI

Examples 1 and 2 below are non-convex and probably of no practical use. However, they illustrate that RSI inequality (7) imposes a "minimum average curvature" of f between x and x_{prj}, and unlike (4), it alone does *not* guarantee convexity. Hence, the RSC definition must explicitly include convexity.

Example 1 (Figure 1(a), RSI and non-convex)**.**

$$f_1(x) = \begin{cases} 0, & x \leq 0, \\ 1 - \sqrt{1 - x^2}, & 0 \leq x \leq 1, \\ 1 + \sqrt{1 - (x-2)^2}, & 1 \leq x \leq 2 - \frac{\sqrt{2}}{2}, \\ \frac{1}{2}(x - 1 + \frac{\sqrt{2}}{2})^2 + \frac{1+\sqrt{2}}{2}, & x \geq 2 - \frac{\sqrt{2}}{2}. \end{cases} \tag{16}$$

f_1 is non-convex, and its minimizer set is $(-\infty, 0]$. Since $f_1'(x) \to +\infty$ as $x \to 1$, f_1' is not Lipschitz continuous. f_1 satisfies RSI(ν) with $\nu = \frac{2}{4-\sqrt{2}} = \min_{x \geq 0} f_1'(x)/x$.

Example 2 (Figure 1(b), RSI and non-convex)**.**

$$f_2(x) = \begin{cases} 0, & x \le 0, \\ 1 - \sqrt{1 - x^2}, & 0 \le x \le \frac{\sqrt{2}}{2}, \\ \sqrt{1 - (x - \sqrt{2})^2} - \sqrt{2} + 1, & \frac{\sqrt{2}}{2} \le x \le 1, \\ \frac{1}{2}(x - 1 + \sqrt{\frac{\sqrt{2}-1}{2}})^2 + \sqrt{2\sqrt{2} - 2} + \frac{5-5\sqrt{2}}{4}, & x \ge 1. \end{cases} \tag{17}$$

f_2 is non-convex, and its minimizer set is $(-\infty, 0]$. Unlike f_1, $\max_{x \ge 0} \frac{\nabla f_2(x)}{x}$ is finite and thus f_2 has a Lipschitz continuous gradient. f_2 satisfies RSI(ν) with $\nu = \sqrt{\frac{\sqrt{2}-1}{2}} = \min_{x \ge 0} f_2'(x)/x$.

Examples 3 and 4 below explain that RSC and strict convexity do not contain each other, and strong convexity is strictly included in their intersection. Recall that a function f is strictly convex if $f(\alpha x + (1 - \alpha)y) < \alpha f(x) + (1 - \alpha) f(y)$ for any $x \ne y$ and $\alpha \in (0, 1)$.

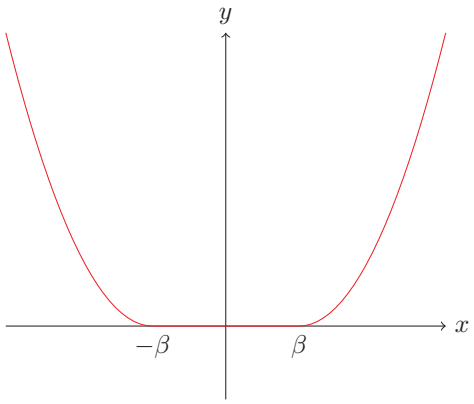


Figure 2: RSC but not strictly convex

Example 3 (Figure 2, RSC but not strictly convex)**.** *Let $x \in \mathbb{R}, \beta > 0$ and define*

$$\begin{aligned} \text{shrink}_\beta(x) &= \text{sign}(x) \max\{|x| - \beta, 0\}, \\ f_3(x) &= \frac{1}{2} \| \text{shrink}_\beta(x) \|^2. \end{aligned} \tag{18}$$

f is not strictly convex since $f_3(x) = 0$ for $x \in \mathcal{X}^* = [-\beta, \beta]$, which is its minimizer set. On the other hand, $f_3(x) = (1/2)\|x - \beta\|^2$ for $x \ge \beta$ and $f_3(x) = (1/2)\|x + \beta\|^2$ for $x \le \beta$, so f_3 is RSC(ν) with $\nu = 1$.

Example 4 (Strictly convex, but not RSC)**.** *Functions $f(x) = x^4$ and $f(x) = e^x$ are strictly convex but* not *RSC. In particular, $f(x) = e^x$ does not have a minimizer though it is lower bounded by 0.*

Motivated by the above examples, we can divide convex differentiable functions into subclasses of RSC, strictly convex, and strongly convex functions depicted in Figure 3. Strictly and strongly convex functions do not need to be differentiable. Although our definition of RSC can be generalized for non-differentiable functions through their subdifferentials, we keep it simple as is.

Example 5 (Dual objective of augmented ℓ_1 model)**.** *Let $A \in \mathbb{R}^{m \times n}$. The Lagrange dual problem to*

$$\min \left\{ \|x\|_1 + \frac{1}{2\alpha}\|x\|^2 : Ax = b \right\} \tag{19}$$

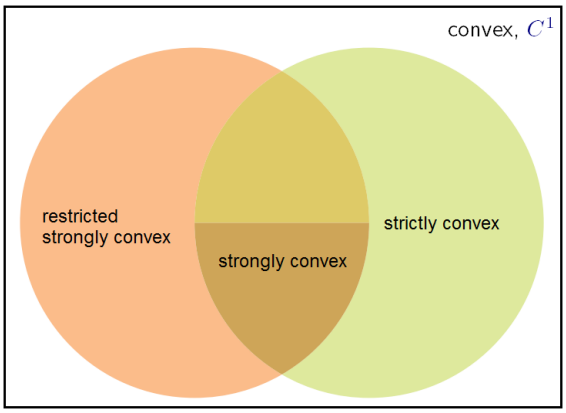


Figure 3: Classes of convex differentiable functions

is

$$\max_y f(y) = b^T y - \frac{\alpha}{2}\|\operatorname{shrink}_1(A^T y)\|^2, \tag{20}$$

where $\operatorname{shrink}_1(z)$ *is given in* (18). *Provided that* $Ax = b$ *is consistent,* [4] *shows that* $-f$ *is RSC*(ν) *with* $\nu > 0$. *(See Lemma 7 of* [4] *for an explicit lower bound of* ν*).*

Admittedly, establishing RSC and deriving a bound for ν are not straightforward as they typically involve projection to the minimizer set $\mathcal{X}^*$, which may not be easy to analytically derive. On the other hand, we have to live with RSC as we will show later that it is both sufficient and necessary. Next we present a useful result for certain composite functions.

Theorem 1 (Linear composition)**.** *Let* $g \in \hat{\mathcal{R}}_{L,\nu}(\mathbb{R}^m)$. *If* g *has a unique minimizer* y^* *and matrix* $A \in \mathbb{R}^{m\times n}$ $(m \le n)$ *has full row-rank, then function* $f(x) = g(Ax)$ *is RSC. Specifically,*

$$f(x) \in \hat{\mathcal{R}}_{\bar{L},\bar{\nu}}(\mathbb{R}^m), \tag{21}$$

where $\bar{L} = L\|A\|^2$ *and* $\bar{\nu} = \nu\lambda_{\min}(AA^T)$.

Applying this theorem, any strongly convex function g with Lipschitz continuous gradient satisfies the condition of Theorem 1 and thus $f(x) = g(Ax)$ is RSC if A has full row-rank though f is generally not strongly convex. (f will be strongly convex if A has full column-rank, following a standard argument). $f(x) = g(Ax)$ arises in various applications including examples in convex quadratic minimization, statistical regression, routing problems in data networks, and many others.

Proof of Theorem 1. For any $x, y \in \mathbb{R}^n$, we have

$$\|\nabla f(x) - \nabla f(y)\| = \|A^T\nabla g(Ax) - A^T\nabla g(Ay)\| \le L\|A\|\|A(x-y)\| \le (L\|A\|^2)\|x-y\|,$$

which means $f \in \mathcal{F}_{\bar{L}}$. By definition, the minimizer set of f is

$$\mathcal{X}^* = \{x \in \mathbb{R}^n : Ax = y^*\},$$

which is nonempty since A has full row-rank. The projection of any $x \in \mathbb{R}^n$ to $\mathcal{X}^*$ is

$$x_{\text{prj}} = x + A^T(AA^T)^{-1}(y^* - Ax).$$

Since $\nabla f(x) = A^T \nabla g(Ax)$, we

$$\langle \nabla f(x), x - x_{\rm prj} \rangle = \langle \nabla g(Ax) - \nabla g(Ax_{\rm prj}), Ax - Ax_{\rm prj} \rangle \geq \nu \|A(x - x_{\rm prj})\|^2 \geq (\nu \lambda_{\min}(AA^T)) \|x - x_{\rm prj}\|^2.$$

where the first inequality follows from $g \in \mathcal{R}_{L,\nu}$ and the second one from $x - x_{\rm prj} \in \text{Range}(A^T)$. □

2.3 Convex conjugacy

The conjugate of convex function f is

$$f^*(y) := \sup_x \{\langle y, x \rangle - f(x)\}. \tag{22}$$

A duality relation can be obtained between RLG and RSC, in analogy to the well-known results that a convex function f is differentiable and ∇f is Lipschitz-continuous with constant L if and only if f^* is strongly convex with constant $1/L$. In this subsection, we consider non-differentiable functions to present our result (while we restrict ourselves to differentiable functions in other sections).

Definition 6. *Let f be a convex function. We say that f has restricted Lipschitz* subgradients *if there exists $L > 0$ such that for any $x \neq 0$,*

$$L\langle p - q, x \rangle \geq \|p - q\|^2, \quad \forall p \in \partial f(x),\ q = \text{Proj}_{\partial f(0)}(p).$$

Definition 6 applies to non-differentiable functions while the usual Lipschitz continuity of gradient of course requires differentiability. In Example 5, the primal objective (19) is non-differentiable but satisfies Definition 6 with $L = \alpha^{-1}$.

Theorem 2. *Let f be a strictly convex function and $0 \in \text{dom} f$. f has restricted Lipschitz subgradients with constant $L > 0$ if and only if f^* is RSC with constant $L^{-1} > 0$.*

Proof. Due to the strict convexity of f, the sup-problem in (22) has a unique solution, denoted by $x(y)$, which satisfies

$$0 \in y - \partial f(x(y)).$$

Also, f^* is differentiable since f is strictly convex, and $\nabla f^*(y) = x(y)$.

Consider problem $\min f^*(y)$, which has solution set $\mathcal{Y}^* = \{y : \nabla f^*(y) = 0\} = \{y : x(y) = 0\} = \partial f(0)$.

"$\Longrightarrow$" Pick $y \notin \mathcal{Y}^*$ and let $y_{\rm prj} = \text{Proj}_{\mathcal{Y}^*}(y) = \text{Proj}_{\partial f(0)}(y) \in \mathcal{Y}^*$. From $y \in \partial f(x(y))$,

$$\langle \nabla f^*(y) - \nabla f^*(y_{\rm prj}), y - y_{\rm prj} \rangle = \langle x(y), y - y_{\rm prj} \rangle \geq L^{-1} \|y - y_{prj}\|^2,$$

where the last inequality follows from Definition 6.

"$\Longleftarrow$" Pick any $x \neq 0$ and $p \in \partial f(x)$. Let $y = p$ and $y_{\rm prj} = q = \text{Proj}_{\partial f(0)}(p)$. Then, $\nabla f^*(y) = x$ and $\nabla f^*(y_{\rm prj}) = 0$. Then,

$$L\langle p - q, x \rangle = L\langle y - y_{\rm prj}, \nabla f^*(y) - \nabla f^*(y_{\rm prj}) \rangle \geq \|y - y_{prj}\|^2 = \|p - q\|^2,$$

where the inequality follows from the definition of RSC. □

3 Main results

This section derives the complexity bounds for the ordinary and accelerated gradient methods under RLG and/or RSC conditions; the derived complexities are summarized in Table 2. The bounds are presented for the following error quantities:

1. Objective error: $\Delta_k := f(x^{(k)}) - f^*$, where $f^* = \min_{x \in \mathbb{R}^n} f(x)$;
2. Solution error: $r_k := \|x^{(k)} - x^{(k)}_{\rm prj}\| = \min\{\|x^{(k)} - x^*\| : x^* \in \mathcal{X}^*\}$.

function class	1st-order oracle lower bound	ordinary gradient method	accelerated gradient method
$\mathcal{L}_R(\mathbb{R}^n)$	$O\left(\sqrt{\frac{R}{\epsilon}}\right)$	Theorem 3: $O\left(\frac{R}{\epsilon}\right)$	Theorem 6: $O\left(\sqrt{\frac{R}{\epsilon}}\right)$
$\mathcal{R}_{R,\nu}(\mathbb{R}^n)$	$O\left(\sqrt{\frac{R}{\nu}}\log\frac{1}{\epsilon}\right)$	Theorem 4: $O\left(\frac{R}{\nu}\log\frac{1}{\epsilon}\right)$	Theorem 7: $O\left(\sqrt{\frac{R}{\nu}}\log\frac{1}{\epsilon}\right)$

Table 2: Complexities of the new classes of functions

3.1 Ordinary gradient descent

Algorithm 1 Ordinary gradient descent method

Input: Initialize $x^{(0)} \in \mathbb{R}^n$ and select stepsize $h > 0$.

1: **for** $k = 0, 1, \cdots$, **do**

2: $\quad x^{(k+1)} = x^{(k)} - h\nabla f(x^{(k)})$;

3: **end for**

Theorem 3 (Sublinear convergence for $\mathcal{L}_R(\mathbb{R}^n)$)**.** *Assume that in problem (1), $f \in \mathcal{L}_R(\mathbb{R}^n)$ with $R > 0$. Then Algorithm 1 with stepsize $h \in (0, 1/R]$ converges sublinearly with*

$$\Delta_k = O(\frac{R\, r_0^2}{k}),$$

where $r_0 = \|x^{(0)} - x^{(0)}_{\text{prj}}\|$. It reaches ϵ-accuracy (i.e., $\Delta_k < \epsilon$) in $O(\frac{R}{\epsilon})$ iterations.

Proof. Firstly, we prove that r_k is non-increasing and thus uniformly bounded by r_0. From part 2) of Lemma 1 and $h = \alpha/R$, where $\alpha \in (0, 1]$, we have

$$h^2\|\nabla f(x^{(k)})\|^2 = 2\alpha h \cdot \frac{1}{2R}\|\nabla f(x^{(k)})\|^2 \le 2\alpha h\langle \nabla f(x^{(k)}), x^{(k)} - x^{(k)}_{\text{prj}}\rangle \le 2h\langle \nabla f(x^{(k)}), x^{(k)} - x^{(k)}_{\text{prj}}\rangle,$$

so in turn we get from $x^{(k+1)} = x^{(k)} - h\nabla f(x^{(k)})$ that

$$\begin{aligned} r_{k+1}^2 = \|x^{(k+1)} - x^{(k+1)}_{\text{prj}}\|^2 &\le \|x^{(k+1)} - x^{(k)}_{\text{prj}}\|^2 && \text{(23a)}\\ &= \|x^{(k)} - x^{(k)}_{\text{prj}} - h\nabla f(x^{(k)})\|^2 && \text{(23b)}\\ &= \|x^{(k)} - x^{(k)}_{\text{prj}}\|^2 - 2h\langle \nabla f(x^{(k)}), x^{(k)} - x^{(k)}_{\text{prj}}\rangle + h^2\|\nabla f(x^{(k)})\|^2 \le r_k^2 && \text{(23c)} \end{aligned}$$

and $r_k \le r_0$, $\forall k$.

Next, by the convexity of f, $\langle \nabla f(x^{(k)}), x^{(k)} - x^*\rangle \ge f(x^k) - f^* \ge 0$. Since $r_k \le r_0$, we have the bound

$$\|\nabla f(x^{(k)})\| \ge \frac{r_k}{r_0}\|\nabla f(x^{(k)})\| \ge \frac{|\langle \nabla f(x^{(k)}), x^{(k)} - x^*\rangle|}{r_0} \ge \frac{\Delta_k}{r_0}.$$

By part 1) of Lemma 1, we have

$$\begin{aligned} \Delta_{k+1} &\le \Delta_k + \langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)}\rangle + \frac{R}{2}\|x^{(k+1)} - x^{(k)}\|^2\\ &= \Delta_k - h(1 - \frac{hR}{2})\|\nabla f(x^{(k)})\|^2\\ &\le \Delta_k - \frac{h}{r_0^2}(1 - \frac{hR}{2})\Delta_k^2. \end{aligned}$$

For $h = \alpha/R$, where $0 < \alpha \leq 1$, $\frac{h}{r_0^2}(1 - \frac{hR}{2}) = \frac{\alpha(2-\alpha)}{2(Rr_0^2)} = O(\frac{1}{Rr_0^2})$. Dividing the both sides of $\Delta_{k+1} \leq \Delta_k - O(\frac{1}{Rr_0^2})\Delta_k^2$ by $\Delta_k\Delta_{k+1}$, we get $(1/\Delta_{k+1}) \geq (1/\Delta_k) + O(\frac{1}{Rr_0^2})$. Therefore, $\Delta_k = O(Rr_0^2/k)$, following from which $\Delta_k < \epsilon$ is guaranteed in $O(Rr_0^2/\epsilon) = O(R/\epsilon)$ iterations. □

(Restricted) Lipschitz continuity of ∇f alone cannot provide a decay rate for r_k. In fact, r_k can decay arbitrarily slowly as function f becomes arbitrarily close to being flat near its minimizer. With the addtional RSC assumptions, the theorems below give geometrically-decaying bounds for both r_k and Δ_k.

Theorem 4 (linear convergence for $\mathcal{R}_{R,\nu}$)**.** *Assume that in problem (1), $f \in \mathcal{R}_{R,\nu}(\mathbb{R}^n)$ with some $R, \nu > 0$. Then Algorithm 1 with stepsize $h = \frac{1}{2R}$ converges linearly with*

$$r_{k+1} \leq (1 - \frac{\nu}{2R})^{1/2} \cdot r_k,$$
$$\Delta_k \leq \frac{R}{2} r_0^2 (1 - \frac{\nu}{2R})^k.$$

It reaches ϵ-accuracy in $O\left(\frac{R}{\nu} \log \frac{1}{\epsilon}\right)$ iterations.

Conversely, assuming that f has the unique solution x^* and Algorithm starts from arbitrary $x^{(0)}$ has a finite stepsize h, linear convergence in the form of $\|x^{(k+1)} - x^*\|^2 \leq (1-\delta)\|x^{(k)} - x^*\|^2$ for some $0 < \delta < 1$ requires f to be RSC(ν) for some $\nu > 0$.

Proof. Recall that $x_{\text{prj}}^{(k)}$ is the projection of $x^{(k)}$ onto the solution set $\mathcal{X}^*$ and $r_k = \|x^{(k)} - x_{\text{prj}}^{(k)}\|$. Thus, $\nabla f(x_{\text{prj}}^{(k)}) = 0$. For every $\theta \in [0, 1]$ we have

$$\|x^{(k+1)} - x_{\text{prj}}^{(k+1)}\|^2 \leq \|x^{(k)} - x_{\text{prj}}^{(k)}\|^2 - 2h\langle \nabla f(x^{(k)}), x^{(k)} - x_{\text{prj}}^{(k)} \rangle + h^2 \|\nabla f(x^{(k)}) - \nabla f(x_{\text{prj}}^{(k)})\|^2 \tag{24a}$$
$$\leq \|x^{(k)} - x_{\text{prj}}^{(k)}\|^2 - 2h(\frac{\theta}{2R}\|\nabla f(x^{(k)}) - \nabla f(x_{\text{prj}}^{(k)})\|^2 + (1-\theta)\nu\|x^{(k)} - x_{\text{prj}}^{(k)}\|^2) \tag{24b}$$
$$+ h^2\|\nabla f(x^{(k)}) - \nabla f(x_{\text{prj}}^{(k)})\|^2$$
$$= (1 - 2(1-\theta)\nu h)\|x^{(k)} - x_{\text{prj}}^{(k)}\|^2 + (h^2 - \frac{\theta h}{R})\|\nabla f(x^{(k)}) - \nabla f(x_{\text{prj}}^{(k)})\|^2, \tag{24c}$$

where inequality (24a) follows from (23) and inequality (24b) utilizes (11). We minimize (24c) over θ and h and obtain $\theta = \frac{1}{2}$ and $h = \frac{1}{2R}$; the details can be found in Appendix. Then from (24c) we get

$$\|x^{(k+1)} - x_{\text{prj}}^{(k+1)}\|^2 \leq (1 - \frac{\nu}{2R})\|x^{(k)} - x_{\text{prj}}^{(k)}\|^2, \tag{25}$$

i.e., $r_{k+1} \leq (1 - \frac{\nu}{2R})^{1/2} \cdot r_k$.

By part 1) of Lemma 1, $\nabla f(x_{\text{prj}}^{(k)}) = 0$, and $r_{k+1} \leq (1 - \nu/L)^{1/2} \cdot r_k$, we derive that

$$\Delta_k = f(x^{(k)}) - f^* \leq \frac{R}{2}\|x^{(k)} - x_{\text{prj}}^{(k)}\|^2 = \frac{R}{2} r_k^2 \leq \frac{R}{2} r_0^2 (1 - \frac{\nu}{2R})^k, \tag{26}$$

which shows $\Delta_k \leq \frac{R}{2} r_0^2 (1 - \frac{\nu}{2R})^k$, following from which $\Delta_k < \epsilon$ is guaranteed in $O\left(\frac{R}{\nu} \log \frac{1}{\epsilon}\right)$ iterations.

Now, we show the converse result. Since f has the unique solution x^*, we have $x_{\text{prj}}^{(k+1)} = x_{\text{prj}}^{(k)} = x^*$. Noticing $x^{(k+1)} = x^{(k)} - h\nabla f(x^{(k)})$, we get

$$\|x^{(k+1)} - x^*\|^2 = \|x^{(k)} - x^*\|^2 - 2h\langle \nabla f(x^{(k)}), x^{(k)} - x^* \rangle + h^2\|\nabla f(x^{(k)}) - \nabla f(x^*)\|^2.$$

From $\|x^{(k+1)} - x^*\|^2 \leq (1-\delta)\|x^{(k)} - x^*\|^2$ for some $0 < \delta < 1$, we have

$$h^2\|\nabla f(x^{(k)}) - \nabla f(x^*)\|^2 - 2h\langle \nabla f(x^{(k)}), x^{(k)} - x^* \rangle \leq -\delta\|x^{(k)} - x^*\|^2,$$

and consequently $\langle \nabla f(x^{(k)}), x^{(k)} - x^* \rangle \geq \frac{\delta}{2h}\|x^{(k)} - x^*\|^2$ after dropping $h^2\|\nabla f(x^{(k)}) - \nabla f(x^*)\|^2 \geq 0$. As $x^{(0)}$ is arbitrary, f is RSC(ν) with $\nu = \frac{\delta}{2h} > 0$. □

If RLG is strengthened to global Lipschitz continuity, we can take a *possibly* larger stepsize $1/L$ instead of $1/(2R)$ and have *possibly* better constants in the bound as follows.

Theorem 5 (Linear convergence for $\hat{\mathcal{R}}_{L,\nu}$)**.** *Assume that in problem (1), ∇f is L-Lipschitz continuous and f is RSC(ν) with $L, \nu > 0$. Then Algorithm 1 with stepsize $h = 1/L$ converges linearly with*

$$r_{k+1} \leq (1 - \nu/L)^{1/2} \cdot r_k,$$
$$\Delta_k \leq \frac{L}{2} r_0^2 (1 - \nu/L)^k.$$

It reaches ϵ-accuracy in $O\left(\frac{L}{\nu} \log \frac{1}{\epsilon}\right)$ iterations.

Proof. By replacing Lemma 1 with the following two Lemmas and repeating the arguments in Theorem 4, the desired linear convergence rates can be derived. □

Lemma 4 ([7] Theorem 2.1.5)**.** *If $f(x) \in \mathcal{F}_L(\mathbb{R}^n)$, it obeys*

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n; \tag{27}$$
$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2, \quad \forall x, y \in \mathbb{R}^n. \tag{28}$$

Lemma 5. *Let $\mathcal{X}^*$ be the nonempty solution set of (1). If ∇f is L-Lipschitz continuous and f is RSC(ν) with $L, \nu > 0$, then for every $\theta \in [0, 1]$ the following holds:*

$$\langle \nabla f(x) - \nabla f(x_{\text{prj}}), x - x_{\text{prj}} \rangle \geq \frac{\theta}{L}\|\nabla f(x) - \nabla f(x_{\text{prj}})\|^2 + (1 - \theta)\nu\|x - x_{\text{prj}}\|^2, \tag{29}$$

where x_{prj} denotes the projection of x onto the solution set $\mathcal{X}^*$.

Proof. Inequality (29) follows from inequalities (7) and (28). □

3.2 Accelerated gradient descent

Algorithm 2 Nesterov's accelerated gradient method

Input: Initialization $y^{(0)} \in \mathbb{R}^n, \theta_0 = 1$, and $h > 0$.

1: **for** $k = 0, 1, \cdots,$ **do**
2: $\quad x^{(k+1)} = y^{(k)} - h\nabla f(y^{(k)})$; (negative gradient step)
3: $\quad \beta_{k+1} = (1 - \theta_k)(\sqrt{\theta_k^2 + 4} - \theta_k)/2$; (extrapolation weight)
4: $\quad y^{(k+1)} = x^{(k+1)} + \beta_{k+1}(x^{(k+1)} - x^{(k)})$; (extrapolation)
5: $\quad \theta_{k+1} = \theta_k(\sqrt{\theta_k^2 + 4} - \theta_k)/2$; (dampening of acceleration parameter)
6: **end for**

Algorithm 2 is equivalent to Constant Step Scheme II on Page 80 of [7] (their $\alpha_k \equiv \theta_k$, their $q = 0$) and FISTA on Page 193 of [2] without the nonsmooth regularization function g (their $t_k \equiv 1/\theta_k$[1]).

Theorem 6. *Assume that in problem (1), $f \in \mathcal{L}_R(\mathbb{R}^n)$ with $R > 0$. Then Algorithm 2 with $h = 1/R$ converges sublinearly with*

$$\Delta_k \leq \frac{4R \cdot \|x^{(1)} - x^{(1)}_{\text{prj}}\|^2}{(k + 1)^2}. \tag{30}$$

It reaches ϵ-accuracy in $O(\sqrt{\frac{R}{\epsilon}})$ iterations.

[1] Step 5 of Algorithm 2 satisfies $\theta_{k+1}^2 = (1 - \theta_{k+1})\theta_k^2$; plugging $\theta_k = 1/t_k$ and $\theta_{k+1} = 1/t_{k+1}$, we obtain $t_{k+1}^{-2} = (1 - t_{k+1}^{-1})t_k^{-2}$, which gives step (4.2) in [2]. Also, β_{k+1} equals $\frac{t_k - 1}{t_{k+1}}$ in (4.3).

The proof below is self-contained and inspired by [12]. Its $O(\sqrt{\frac{R}{\epsilon}})$ is better than $O(\frac{R}{\epsilon})$ of Theorem 3.

Proof. Sequences $\{\theta_k\}$ and $\{\beta_k\}$ obey the following recursive relationships:

$$\frac{1}{\theta_k^2} = \frac{1-\theta_{k+1}}{\theta_{k+1}^2} \quad \text{and} \quad \beta_{k+1} = \theta_k(1-\theta_k)/(\theta_k^2+\theta_{k+1}) = \theta_{k+1}(\frac{1}{\theta_k}-1).$$

Defining $x^{(0)}=0$ and $v^{(k+1)} = x^{(k)} + \frac{1}{\theta_k}(x^{(k+1)}-x^{(k)})$, we can rewrite $y^{(k+1)} = \theta_{k+1}v^{(k+1)} + (1-\theta_{k+1})x^{(k+1)}$. From part 1) of Lemma 1 and the convexity of f, for any $z \in \mathbb{R}^n$ we have

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

Setting $z = \theta_k x^* + (1-\theta_k)x^{(k)}$, where $x^* \in \mathcal{X}^*$, and using the convexity of f, we get

$$f(x^{(k+1)}) \le \theta_k f^* + (1-\theta_k)f(x^{(k)}) + R\langle x^{(k+1)} - y^{(k)}, \theta_k x^* + (1-\theta_k)x^{(k)} - x^{(k+1)} \rangle + \frac{R}{2}\|x^{(k+1)} - y^{(k)}\|^2. \quad (32)$$

Since $\theta_k x^* + (1-\theta_k)x^{(k)} - x^{(k+1)} = \theta_k(x^* - v^{(k+1)})$ and $x^{(k+1)} - y^{(k)} = \theta_k(v^{(k+1)} - v^{(k)})$, we have

$$\begin{aligned}
R\langle x^{(k+1)} - y^{(k)}, \theta_k x^* + (1-\theta_k)x^{(k)} - x^{(k+1)} \rangle &= R\theta_k^2 \langle v^{(k+1)} - v^{(k)}, x^* - v^{(k+1)} \rangle \\
&= R\theta_k^2 \langle v^{(k+1)} - x^*, v^{(k)} - x^* \rangle - R\theta_k^2 \|v^{(k+1)} - x^*\|^2
\end{aligned}$$

and

$$\frac{R}{2}\|x^{(k+1)} - y^{(k)}\|^2 = \frac{R\theta_k^2}{2}(\|v^{(k+1)} - x^*\|^2 + \|v^{(k)} - x^*\|^2 - 2\langle v^{(k+1)} - x^*, v^{(k)} - x^* \rangle). \quad (33)$$

Substituting these equations into the last two terms of (32), we get

$$f(x^{(k+1)}) \le \theta_k f^* + (1-\theta_k)f(x^{(k)}) - \frac{R\theta_k^2}{2}\|v^{(k+1)} - x^*\|^2 + \frac{R\theta_k^2}{2}\|v^{(k)} - x^*\|^2. \quad (34)$$

Reordering the terms and dividing by θ_k^2 and then recursively deducing, we have

$$\begin{aligned}
\frac{1}{\theta_k^2}(f(x^{(k+1)}) - f^*) + \frac{R}{2}\|v^{(k+1)} - x^*\|^2 &\le \frac{1-\theta_k}{\theta_k^2}(f(x^{(k)}) - f^*) + \frac{R}{2}\|v^{(k)} - x^*\|^2 \quad (35a) \\
&= \frac{1}{\theta_{k-1}^2}(f(x^{(k)}) - f^*) + \frac{R}{2}\|v^{(k)} - x^*\|^2 \quad (35b) \\
&\le \cdots \le f(x^{(1)}) - f^* + \frac{R}{2}\|v^{(1)} - x^*\|^2 \quad (35c)
\end{aligned}$$

where the last inequality follows from $\theta_0 = 1$. Since $v^{(1)} = x^{(1)}$ and $f(x^{(1)}) - f^* \le \frac{R}{2}\|x^{(1)} - x^*\|^2$ from part 1) of Lemma 1, we finally obtain

$$f(x^{(k+1)}) - f^* \le R\theta_k^2\|x^{(1)} - x^*\|^2 \le R\theta_k^2\|x^{(1)} - x_{\text{prj}}^{(1)}\|^2. \quad (36)$$

Finally, we derive $\theta_k < \frac{2}{k+2}$ for $k = 0, 1, 2, \ldots$ from which the sublinear convergence rate (30) and its corresponding complexity will follow. From $\theta_0 = 1$ and Step 5 of Algorithm 2, we have $\theta_k > 0$. From $\sqrt{\theta_k^2 + 4} > 2$ and Step 5 again, we have $\frac{\theta_{k+1}}{\theta_k} > \frac{2-\theta_k}{2}$ and thus $\frac{1}{\theta_{k+1}} - 1 = \frac{\theta_{k+1}}{\theta_k^2} > \frac{1}{\theta_k} - \frac{1}{2} = (\frac{1}{\theta_k} - 1) + \frac{1}{2}$. Hence, for all $k \ge 0$, we have $\frac{1}{\theta_k} - 1 > \frac{k}{2}$ or $\theta_k < \frac{2}{k+2}$. □

Algorithm 3 Algorithm 2 with restarts

Input: Initialization $y^{(0,0)} \in \mathbb{R}^n, \theta_0 = 1$, restart interval K.

1: **for** $j = 0, 1, \cdots$, **do**

2: obtain $x^{(j,K)}$ by running Algorithm 2 for K iterations;

3: set $x^{(j+1,0)} = x^{(j,K)}$, $y^{(j+1,0)} = x^{(j,K)}$ and $\theta_0 = 1$;

4: **end for**

Theorem 7. *Assume that in problem (1), $f \in \mathcal{R}_{R,\nu}(\mathbb{R}^n)$ with some $R > 0, \nu > 0$. Then Algorithm 3 with $h = 1/R$ and $K = \sqrt{8eR/\nu}$ reaches ϵ-accuracy in $\mathcal{O}(\sqrt{\frac{R}{\nu}} \log \frac{1}{\epsilon})$ iterations.*

Proof. At iteration j of Algorithm 3, we have

$$f(x^{(j+1,0)}) - f^* = f(x^{(j,K)}) - f^* \leq \frac{4R \cdot \|x^{(j,0)} - x^{(j,0)}_{\text{prj}}\|_2^2}{K^2} \leq \frac{8R}{\nu K^2}(f(x^{j,0}) - f^*) \tag{37}$$

where the first inequality follows from the convergence guarantee (30) of Algorithm 2 and the second from Lemma 3. After jK iterations, by the setting of $K = \sqrt{8eR/\nu}$ we have

$$f(x^{(j,0)}) - f^* \leq (\frac{8R}{\nu K^2})^j (f(x^{0,0}) - f^*) = (\frac{1}{e})^j (f(x^{0,0}) - f^*) \tag{38}$$

Thus, to obtain an ϵ-solution, we only need to take $j = \mathcal{O}(\log(1/\epsilon))$ and hence the total number of iterations $jK = \mathcal{O}(\sqrt{\frac{R}{\nu}} \log \frac{1}{\epsilon})$, which completes the proof. □

The above result and proof were motivated by [9]. Compared to [9] and [10], we use weaker conditions.

4 Application to augmented ℓ_1 minimization

4.1 An improved convergence rate

The augmented ℓ_1 model (19) returns an exact solution to

$$\min_x \{\|x\|_1 : \ Ax = b\} \tag{39}$$

provided that α in (19) is large enough. For most problems where a sparse solution x^* is expected from (39), such as those arising in compressive sensing, paper [4] argues that $\alpha = 10\|x^*\|_\infty$ is sufficient. The Lagrange dual of (19), which is problem (20), has an unconstrained and differentiable objective function. By Example 5, the negative of the dual objective function, $-f(y)$, satisfies RSC. In addition, f has an L-Lipschitz continuous gradient ∇f with $L = \alpha\|A\|^2$. Therefore, we can apply Theorems 5 and 7 to the ordinary and accelerated gradient iterations for (20).

The gradient ascent iteration for (20) is known as the linearized Bregman algorithm (LBreg):

$$x^{(k+1)} \leftarrow \alpha \, \text{shrink}(A^T y^{(k)}), \tag{40a}$$

$$y^{(k+1)} \leftarrow y^{(k)} + h(b - Ax^{(k+1)}), \tag{40b}$$

where $x^{(k)}$ and $y^{(k)}$ are the primal and dual variables at iteration k and $h > 0$ is the step size. One can verify that $(b - Ax^{(k+1)})$ is the gradient to the objective of (20). The solution set is given by

$$\mathcal{Y}^* = \{y \in \mathbb{R}^m : b - \alpha A \, \text{shrink}(A^T y) = 0\} = \{y \in \mathbb{R}^m : \alpha \, \text{shrink}(A^T y) = x^*\} \tag{41}$$

where x^* is assumed to be the unique solution to (19); the derivation can be found in [4].

Paper [4] shows

$$\|y^{(k)} - y^{(k)}_{\text{prj}}\| \le \sqrt{1 - \left(\frac{\nu}{L}\right)^2}\, \|y^{(k-1)} - y^{(k-1)}_{\text{prj}}\|.$$

Applying Theorem 5, we obtain a tighter convergence bound:

Theorem 8. *In problem (20), assume that $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are nonzero and $Ax = b$ are consistent. Let f^* be the optimal objective value of (20). The linearized Bregman iteration (40) starting from any $y^{(0)} \in \mathbb{R}^n$ with step size $h_k = \frac{1}{L}$ generates a Q-linearly converging sequence $\{y^{(k)}\}$*

$$\|y^{(k)} - y^{(k)}_{\text{prj}}\| \le \sqrt{1 - \frac{\nu}{L}}\|y^{(k-1)} - y^{(k-1)}_{\text{prj}}\|, \quad \forall k \ge 1. \tag{42}$$

The objective value converges R-linearly as

$$f^* - f(y^{(k)}) \le \frac{L}{2}\|y^{(0)} - y^{(0)}_{\text{prj}}\|^2(1 - \frac{\nu}{L})^k, \quad \forall k \ge 1. \tag{43}$$

Furthermore, $x^{(k)}$ converges R-linearly as

$$\|x^{(k+1)} - x^*\| \le L\|y^{(0)} - y^{(0)}_{\text{prj}}\|(1 - \frac{\nu}{L})^{k/2}, \quad \forall k \ge 1, \tag{44}$$

where x^* is the solution to (19). The results are in the global sense.

Proof. Due to (40a), (41), the expression $\nabla f(y) = b - \alpha\,\text{shrink}(A^T y)$, and the Lipschitz property (3) of $\nabla f(y)$, we have

$$\|x^{(k+1)} - x^*\| = \|\alpha\,\text{shrink}(A^T y^{(k)}) - \alpha\,\text{shrink}(A^T y^{(k)}_{\text{prj}})\|, \tag{45a}$$

$$= \|\nabla f(y^{(k)}) - \nabla f(y^{(k)}_{\text{prj}})\|, \tag{45b}$$

$$\le L\|y^{(k)} - y^{(k)}_{prj}\|. \tag{45c}$$

which gives (44). The remained results follow from Theorem 5 applied to $-f$. □

4.2 Numerical simulation

To demonstrate the convergence results, we compared the following algorithms for problem (20):

1. fixed-step gradient ascent (Algorithm 1);
2. gradient ascent with Nesterov's acceleration (Algorithm 2, [3]);
3. Nesterov's acceleration with *restart* (Algorithm 4 with *restart*);
4. Nesterov's acceleration with *skip* (Algorithm 4 with *skip*).

Although for (20) we can compute $K = \sqrt{8eL/\nu}$ using the lower bound of ν given in Example 5 and thus run Algorithm 3 with restart every K iterations, such K was found too large. Instead, we ran Algorithm 4, which uses the following scheme to trigger *restart* as suggested in [10] (the inequality is given in the opposite directions for concave maximization):

$$\text{Gradient scheme: } \nabla f(y^{(k-1)})^T(y^{(k)} - y^{(k-1)}) < 0.$$

We also introduce the *skip* heuristic: set $\beta_{k+1} = 0$ (and make *no* change to θ_k).

Algorithm 4 Nesterov's accelerated gradient method with *reset*

Input: Initialization $y^{(0)} \in \mathbb{R}^n, \theta_0 = 1$, and $h > 0$.

for $k = 0, 1, \cdots,$ **do**

  $x^{(k+1)} = y^{(k)} - h\nabla f(y^{(k)})$; (negative gradient step)

  If *restart* then

    $\theta_k = 1$ and $\beta_{k+1} = 0$;

  elseif *skip* then

    $\beta_{k+1} = 0$;

  else

    $\beta_{k+1} = (1 - \theta_k)(\sqrt{\theta_k^2 + 4} - \theta_k)/2$; (extrapolation weight)

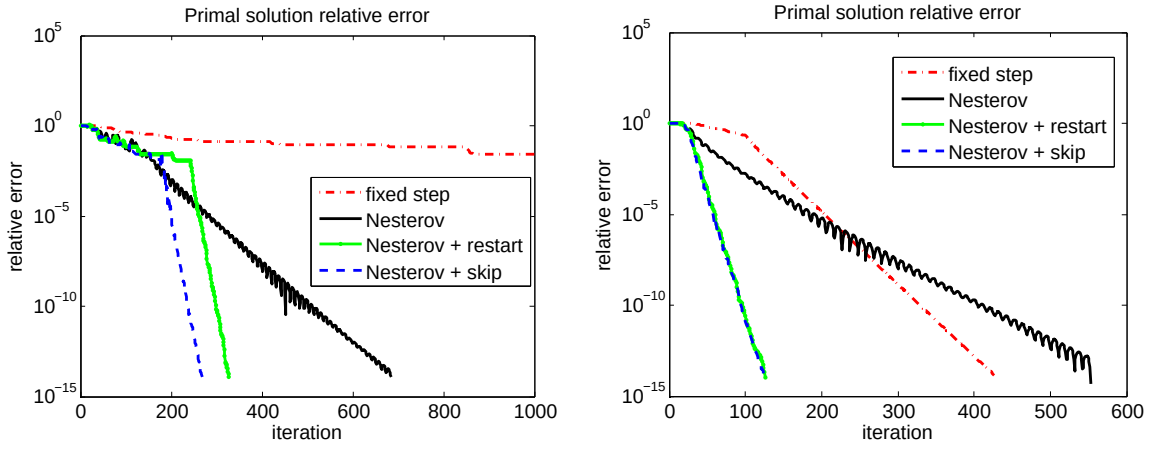
  End if

  $y^{(k+1)} = x^{(k+1)} + \beta_{k+1}(x^{(k+1)} - x^{(k)})$; (extrapolation)

  $\theta_{k+1} = \theta_k(\sqrt{\theta_k^2 + 4} - \theta_k)/2$; (dampening of acceleration parameter)

end for

The comparisons use two examples. Each had sparse signals x^o with 512 entries, out of which 25 were nonzero entries sampled independently from the standard Gaussian distribution (Test 1, Figure 4(a)) or set to ± 1 uniformly randomly (Test 2, Figure 4(b)). Both examples have the same sensing matrix A with 256 rows and entries sampled independently from the standard Gaussian distribution. We used the following parameters: $b = Ax^o$, $\alpha = 10\|x^o\|_\infty$, and $h = \frac{1}{L} = \frac{1}{\alpha\|A\|^2}$. All iterations were stopped upon $\|Ax^{(k)} - b\| < 10^{-14}\|b\|$. Figure 4 depicts the relative error $\frac{\|x^{(k)} - x^0\|}{\|x^0\|}$ versus iteration k.



(a) Test 1: Gaussian sparse vector recovery

(b) Test 2: Bernoulli sparse vector recovery

Figure 4: Relative error of primal variable $x^{(k)}$

The fixed-step gradient iteration converged very slowly in Test 1, much slower than in Test 2; this can be explained by a smaller ν in Test 1 (see Lemma 7 of [4] for an explicit lower bound of ν). The fixed-step iteration exhibited a linear-convergence behavior in Test 2 though we cannot tell the same from Test 1,.

The accelerated gradient method performed similarly in both tests. Its performance was significantly improved in the second phase by *restart* and *skip*. In Test 1, *skip* was more effective. The two schemes did not appear to make much difference in these tests. It is interesting to note that in Test 2, both *restart* and *skip* had faster rates of convergence than the fixed-step gradient iteration; this deserves further tests and

perhaps theoretical investigation.

As the focus of this paper is not numerical simulation, we do not present more numerical results. For the interested reader, the source code can be found on the second author's homepage.

5 Conclusions

The convergence behavior of gradient methods on convex differentiable functions is one of the core questions in convex optimization. It is known to many researchers that global Lipschitz continuity of ∇f is more than sufficient for sublinear convergence and asking f to be strongly convex is also too much for linear convergence. For the ordinary and accelerated gradient methods, this paper shows using rather straightforward steps that these conditions restricted to certain line segments are sufficient for the existing convergence results to hold. In addition, it shows that strong convexity restricted to between current point x and its projection to the solution set is also necessary for the geometric decay of solution error.

For the accelerated gradient method to achieve the best worst-case bound $O\left(\sqrt{\frac{R}{\nu}}\log\frac{1}{\epsilon}\right)$ on (restricted) strongly convex functions, the modulus ν of the objective function must be given. This is not practical. It is an open question to design a method with this bound but not requiring the knowledge of ν. On the other hand, the *restart* and *skip heuristics* appear to improve the performance of the accelerated method.

Acknowledgements

We want to thank Profs. Q. Ling, S. Ma, Z. Wen, and Y. Zhang and graduate students Z. Peng and Y. Xu for discussions and corrections. The work of H. Zhang is supported by China Scholarship Council during his visit to Rice University, and in part by Graduate School of NUDT under Funding of Innovation B110202, Hunan Provincial Innovation Foundation For Postgraduate CX2011B008, and National Science Foundation of China under Grants No. 61271014 and No.61072118. H. Zhang thanks Rice University, CAAM Department, for hosting him. The work of W. Yin is supported in part by NSF grants DMS-0748839 and ECCS-1028790, and ONR Grant N00014-08-1-1101.

Appendix

We select the parameter θ and step size h in (24c) to minimize the upper bound. Let $r=\frac{\nu}{2R}, h>0$. As we need to deal with the second term in (24c), two cases are studied below depending on the sign of $h^2-\frac{\theta h}{R}$:

Case A: $h^2-\frac{\theta h}{R}\le 0$, i.e., $h\in(0,\frac{\theta}{R}], \theta\in[0,1]$. Applying the Cauchy-Schwartz inequality to RSI, we get

$$\|\nabla f(x^{(k)})-\nabla f(x^{(k)}_{\rm prj})\|^2\ge\nu^2\|x^{(k)}-x^{(k)}_{\rm prj}\|^2. \tag{46}$$

From $h^2-\frac{\theta h}{R}\le 0$ and (24c), we derive that

$$\begin{aligned}\|x^{(k+1)}-x^{(k+1)}_{\rm prj}\|^2 &\le (1-2(1-\theta)\nu h)\|x^{(k)}-x^{(k)}_{\rm prj}\|^2+\nu^2(h^2-\frac{\theta h}{R})\|x^{(k)}-x^{(k)}_{\rm prj}\|^2, &\text{(47a)}\\ &=\left(\nu^2h^2-2\left((1-\theta)\nu+\frac{\theta\nu^2}{2R}\right)h+1\right)\|x^{(k)}-x^{(k)}_{\rm prj}\|^2, &\text{(47b)}\\ &\triangleq f_1(\theta,h)\|x^{(k)}-x^{(k)}_{\rm prj}\|^2. &\text{(47c)}\end{aligned}$$

Let $h_0=\frac{\theta}{2R}+\frac{(1-\theta)}{\nu}$, which is the minimum point of the quadratic function $f_1(\theta,h)$ over variable h for each fixed θ. To determine whether such h_0 is included in the interval $(0,\frac{\theta}{R}]$, we consider $h_0=\frac{\theta}{2R}+\frac{(1-\theta)}{\nu}=\frac{\theta}{R}$

and get $\theta = \frac{1}{1+r}$. Now, we split the interval $[0,1]$ into $[\frac{1}{1+r}, 1]$ and $[0, \frac{1}{1+r})$. If $\theta \in [\frac{1}{1+r}, 1]$, we have $\frac{\theta}{R} \geq h_0$ which means the point $h_0 \in (0, \frac{\theta}{R}]$. Thus,

$$\min_{h \leq \frac{\theta}{R}, \frac{1}{1+r} \leq \theta \leq 1} f_1(\theta, h) = \min_{\frac{1}{1+r} \leq \theta \leq 1} f_1(\theta, h_0) = \min_{\frac{1}{1+r} \leq \theta \leq 1} 1 - (1 - (1+r)\theta)^2 = 1 - r^2,$$

where the minimum value $1 - r^2$ is obtained at $\theta = 1$ and $h = h_0 = \frac{1}{2R}$. If $\theta \in [0, \frac{1}{1+r})$, we have $\frac{\theta}{R} < h_0$ which means the point $h_0 \notin (0, \frac{\theta}{R}]$. By monotone decreasing of $f_1(\theta, h)$ on the interval $h \leq \frac{2\theta}{L}$ for each fixed θ, we have

$$\min_{h \leq \frac{\theta}{R}, 0 \leq \theta < \frac{1}{1+r}} f_1(\theta, h) = \min_{0 \leq \theta < \frac{1}{1+r}} f_1(\theta, \frac{\theta}{R}) = \min_{0 \leq \theta < \frac{1}{1+r}} 1 - 4\theta(1-\theta)r = 1 - r$$

where the minimum value $1 - r$ is obtained at $\theta = \frac{1}{2}$ and $h = \frac{\theta}{R} = \frac{1}{2R}$; note that $\frac{1}{2} \in [0, \frac{1}{1+r})$ since $r < 1$. Therefore, on the intervals $h \in (0, \frac{\theta}{R}]$ and $\theta \in [0,1]$, the minimum value $1 - r$ of $f_1(\theta, h)$ is obtained at $(\theta, h) = (\frac{1}{2}, \frac{1}{2R})$.

Case B: $h^2 - \frac{\theta h}{R} \geq 0$, i.e., $h \in [\frac{\theta}{R}, +\infty), \theta \in [0,1]$. Applying the Cauchy-Schwartz inequality to part 2) of Lemma 1, we get

$$\|\nabla f(x^{(k)}) - \nabla f(x^{(k)}_{\text{prj}})\|^2 \leq 4R^2 \|x^{(k)} - x^{(k)}_{\text{prj}}\|^2. \tag{48}$$

From $h^2 - \frac{\theta h}{R} \geq 0$ and (24c), we derive that

$$\|x^{(k+1)} - x^{(k+1)}_{\text{prj}}\|^2 \leq (1 - 2(1-\theta)\nu h)\|x^{(k)} - x^{(k)}_{\text{prj}}\|^2 + 4R^2(h^2 - \frac{\theta h}{R})\|x^{(k)} - x^{(k)}_{\text{prj}}\|^2, \tag{49a}$$

$$= (4R^2h^2 - 2(2\theta R + (1-\theta)\nu)h + 1)\|x^{(k)} - x^{(k)}_{\text{prj}}\|^2, \tag{49b}$$

$$\triangleq f_2(\theta, h)\|x^{(k)} - x^{(k)}_{\text{prj}}\|^2. \tag{49c}$$

Let $h_1 = \frac{2\theta R + (1-\theta)\nu}{4R^2}$, which is the minimum point of the quadratic function $f_2(\theta, h)$ over variable h for each fixed θ. Similarly, we split the interval $[0,1]$ into $(\frac{r}{1+r}, 1]$ and $[0, \frac{r}{1+r}]$. If $\theta \in (\frac{r}{1+r}, 1]$, we have $\frac{\theta}{R} > h_1$ which means $h_1 \notin [\frac{\theta}{R}, +\infty)$. By monotone increasing of $f_2(\theta, h)$ on the interval $h \geq \frac{\theta}{R}$ for each fixed θ, we have

$$\min_{h \geq \frac{2\theta}{L}, \frac{r}{1+r} < \theta \leq 1} f_2(\theta, h) = \min_{\frac{r}{1+r} < \theta \leq 1} f_2(\theta, \frac{\theta}{R}) = \min_{\frac{r}{1+r} < \theta \leq 1} 1 - 4\theta(1-\theta)r = 1 - r,$$

where the minimum value $1 - r$ is obtained at $\theta = 1/2$ and $h = \frac{\theta}{R} = \frac{1}{2R}$; note that $\frac{1}{2} \in (\frac{r}{1+r}, 1]$ since $r < 1$. If $\theta \in [0, \frac{r}{1+r}]$, we have $\frac{\theta}{R} \leq h_1$ which means $h_1 \in [\frac{\theta}{R}, +\infty)$. Thus,

$$\min_{h \geq \frac{\theta}{R}, 0 \leq \theta \leq \frac{r}{1+r}} f_2(\theta, h) = \min_{0 \leq \theta \leq \frac{r}{1+r}} f_2(\theta, h_1) = \min_{0 \leq \theta \leq \frac{r}{1+r}} 1 - (\frac{2\theta R + (1-\theta)\nu}{2R})^2 = 1 - (\frac{2\nu}{2R+\nu})^2,$$

where the minimum value is obtained at $\theta = \frac{r}{1+r}$ and $h = h_1$. After simple calculations, it holds $r = \frac{\nu}{2R} > (\frac{2\nu}{2R+\nu})^2$ and hence $1 - r < 1 - (\frac{2\nu}{2R+\nu})^2$. Therefore, on the intervals $h \in [\frac{\theta}{R}, +\infty)$ and $\theta \in [0,1]$, the minimum value $1 - r$ of $f_2(\theta, h)$ is obtained at $(\theta, h) = (\frac{1}{2}, \frac{1}{2R})$ as well.

References

[1] A. Agarwal, S. Negahban, and M. J. Wainwright, Fast global convergence of gradient methods for high-dimensional statistical recovery, To appear in Annals of Statistics, 2012.

[2] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sciences, 2 (2009), pp. 183-202.

[3] B. Huang, S. Q. Ma, and D. Goldfarb, Accelerated Linearized Bregman Method. Journal of Scientific Computation, 54(2013), pp. 428-453.

[4] M.J. Lai and W. Yin, Augmented ℓ_1 and nuclear-norm models with a globally linearly convergent algorithm, To appear in SIAM J. Imaging Sciences, 2013.

[5] S. Negahban, P. Ravikumar, M. J. Wainwright and B. Yu, A unified framework for the analysis of regularized M-estimators, Statistical Science, 27(2012), pp. 538-557.

[6] Y. Nesterov, A method of solving a convex programming problem with convergence rate $O(1/k^2)$, Soviet Mathematics Doklady, 27(1983), pp. 372-376.

[7] Y. Nesterov, Introductory lectures on convex optimization: A basic course, Kluwer Academic Publishers, 2004.

[8] Y. Nesterov, Smooth minimization of non-smooth functions, Mathematical programming, Series A, 103(2005), pp. 127-152.

[9] Y. Nesterov, Gradient methods for minimizing composite objective function, CORE discussion paper, 2007.

[10] B. O'Donoghue and E. Candès, Adaptive restart for accelerated gradient schemes, To appear in Foundations of Computational Mathematics, 2012.

[11] K. Scheinberg, D. Goldfarb, and X Bai, Fast first-order methods for composite convex optimization with line search, submitted, 2011.

[12] P. Tseng, On accelerated proximal gradient methods for convex-concave optimization, submitted to SIAM J. Optim., 2008.