Noname manuscript No.
(will be inserted by the editor)

"Active-set complexity" of proximal gradient

How long does it take to find the sparsity pattern?

Julie Nutini · Mark Schmidt · Warren Hare

Received: date / Accepted: date

Abstract Proximal gradient methods have been found to be highly effective for solving minimization problems with non-negative constraints or ℓ_1-regularization. Under suitable nondegeneracy conditions, it is known that these algorithms identify the optimal sparsity pattern for these types of problems in a finite number of iterations. However, it is not known how many iterations this may take. We introduce the notion of the "active-set complexity", which in these cases is the number of iterations before an algorithm is guaranteed to have identified the final sparsity pattern. We further give a bound on the active-set complexity of proximal gradient methods in the common case of minimizing the sum of a strongly-convex smooth function and a separable convex non-smooth function.

Keywords convex optimization · non-smooth optimization · proximal gradient method · active-set identification · active-set complexity

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [Discovery Grant, reference numbers #355571-2013, #2015-06068].
Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) [Discovery Grant, numéros de référence #355571-2013, #2015-06068].

Julie Nutini
Department of Computer Science, The University of British Columbia
201-2366 Main Mall, Vancouver BC, V6T 1Z4, Canada
E-mail: jnutini@cs.ubc.ca

Mark Schmidt
Department of Computer Science, The University of British Columbia
E-mail: schmidtm@cs.ubc.ca

Warren Hare
Department of Mathematics, The University of British Columbia Okanagan
E-mail: warren.hare@ubc.ca

1 Motivation

We consider the problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x)+g(x), \tag{1}$$

where f is μ-strongly convex and the gradient ∇f is L-Lipschitz continuous. We assume that g is a separable function,

$$g(x)=\sum_{i=1}^{n} g_i(x_i),$$

and each g_i only needs to be a proper convex and lower semi-continuous function (it may be non-smooth or infinite at some x_i). In machine learning, a common choice of f is the squared error $f(x)=\frac{1}{2}\|Ax-b\|^2$ (or an ℓ_2-regularized variant to guarantee strong-convexity). The squared error is often paired with a scaled absolute value function $g_i(x_i)=\lambda|x_i|$ to yield a sparsity-encouraging ℓ_1-regularization term. This is commonly known as the LASSO problem [24]. The g_i can alternatively enforce bound constraints (e.g., the dual problem in support vector machine optimization [9]), such as the x_i must be non-negative, by defining $g_i(x_i)$ to be an indicator function that is zero if the constraints are satisfied and ∞ otherwise.

One of most widely-used methods for minimizing functions of this form is the proximal gradient (PG) method [16,1,20,3], which uses an iteration update given by

$$x^{k+1}=\mathbf{prox}_{\frac{1}{L}g}\left(x^k-\frac{1}{L}\nabla f(x^k)\right),$$

where the proximal operator is defined as

$$\mathbf{prox}_{\frac{1}{L}g}(x)=\underset{y}{\operatorname{argmin}}\,\frac{1}{2}\|y-x\|^2+\frac{1}{L}g(y).$$

When the proximal gradient method is applied with non-negative constraints or ℓ_1-regularization, an interesting property of the method is that the iterations x^k will match the sparsity pattern of the solution x^* for all sufficiently large k (under a mild technical condition). Thus, after a finite number of iterations the algorithm "identifies" the final set of non-zero variables. This is useful if we are only using the algorithm to find the sparsity pattern, since it means we do not need to run the algorithm to convergence. It is also useful in designing faster algorithms (for example, see [14,10,6] for non-negativity constrained problems and [25,23,11]) for ℓ_1-regularized problems). After we have identified the set of non-zero variables we could switch to a more sophisticated solver like Newton's method applied to the non-zero variables. In any case, we should expect the algorithm to converge faster after identifying the final sparsity pattern, since it will effectively be optimizing over a lower-dimensional subspace.

The idea of finitely identifying the set of non-zero variables dates back at least 40 years to the work of Bertsekas [2] who showed that the projected

gradient method identifies the sparsity pattern in a finite number of iterations when using non-negative constraints (and suggests we could then switch to a superlinearly convergent unconstrained optimizer). Subsequent works have shown that finite identification occurs in much more general settings including cases where g is non-separable, where f may not be convex, and even where the constraints may not be convex [7,26,13,12]. The active-set identification property has also been shown for other algorithms like certain coordinate descent and stochastic gradient methods [18,27,15].

Although these prior works show that the active-set identification must happen after some finite number of iterations, they only show that this happens asymptotically. In this work, we introduce the notion of the "active-set complexity" of an algorithm, which we define as the number of iterations required before an algorithm is guaranteed to have reached the active-set. We further give bounds, under the assumptions above and the standard nondegeneracy condition, on the active-set complexity of the proximal gradient method. We are only aware of one previous work giving such bounds, the work of Liang et al. who included a bound on the active-set complexity of the proximal gradient method [17, Proposition 3.6]. Unlike this work, their result does not evoke strong-convexity. Instead, their work applies an inclusion condition on the local subdifferential of the regularization term that ours does not require. By focusing on the strongly-convex case (which is common in machine learning due to the use of regularization), we obtain a simpler analysis and a much tighter bound than in this previous work. Specifically, both rates depend on the "distance to the subdifferential boundary", but in our analysis this term only appears inside of a logarithm rather than outside of it.

2 Notation and assumptions

We assume that f is μ-strongly convex so that for some $\mu > 0$, we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2, \quad \text{for all } x, y \in \mathbb{R}^n.$$

Further, we assume that its gradient ∇f is L-Lipschitz continuous, meaning that

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|, \quad \text{for all } x, y \in \mathbb{R}^n. \tag{2}$$

By our separability assumption on g, the subdifferential of g is simply the concatenation of the subdifferentials of each g_i. Further, the subdifferential of each individual g_i at any $x_i \in \mathbb{R}$ is defined by

$$\partial g_i(x_i) = \{v \in \mathbb{R} : g_i(y) \geq g_i(x_i) + v \cdot (y - x_i), \text{ for all } y \in \mathbf{dom}\ g_i\},$$

which implies that the subdifferential of each g_i is just an interval on the real line. In particular, the interior of the subdifferential of each g_i at a non-differentiable point x_i can be written as an open interval,

$$\operatorname{int} \partial g_i(x_i) \equiv (l_i, u_i), \tag{3}$$

where $l_i \in \mathbb{R} \cup \{-\infty\}$ and $u_i \in \mathbb{R} \cup \{\infty\}$ (the ∞ values occur if x_i is at its lower or upper bound, respectively).

As in existing literature on active-set identification [13], we require the *nondegeneracy* condition that $-\nabla f(x^*)$ must be in the "relative interior" of the subdifferential of g at the solution x^*. For simplicity, we present the non-degeneracy condition for the special case of (1).

Assumption 1 *We assume that x^* is a nondegenerate solution for problem* (1)*, where x^* is* nondegenerate *if and only if*

$$\begin{cases} -\nabla_i f(x^*) = \nabla_i g(x_i^*) & \text{if } \partial g_i(x_i^*) \text{ is a singleton } (g_i \text{ smooth at } x_i^*) \\ -\nabla_i f(x^*) \in \text{int}\, \partial g_i(x_i^*) & \text{if } \partial g_i(x_i^*) \text{ is not a singleton } (g_i \text{ non-smooth at } x_i^*). \end{cases}$$

Under this assumption, we ensure that $-\nabla f(x^*)$ is in the "relative interior" (see [5, Section 2.1.3]) of the subdifferential of g at the solution x^*. In the case of non-negative constraints, this requires that $\nabla_i f(x^*) > 0$ for all variables i that are zero at the solution ($x_i^* = 0$). For ℓ_1-regularization, this requires that $|\nabla_i f(x^*)| < \lambda$ for all variables i that are zero at the solution, which is again a strict complementarity condition [11].[1]

Definition 1 The *active-set* $\mathcal{Z}$ for a separable g is defined as

$$\mathcal{Z} = \{i : \partial g_i(x_i^*) \text{ is not a singleton}\}.$$

By the above definition and recalling the interior of the subdifferential of g_i as defined in (3), the set $\mathcal{Z}$ includes indices i where x_i^* is equal to the lower bound on x_i, is equal to the upper bound on x_i, or occurs at a non-smooth value of g_i. In the case of non-negative constraints and ℓ_1-regularization under Assumption 1, $\mathcal{Z}$ is the set of non-zero variables at the solution. Formally, the *active-set identification property* for this problem is that for all sufficiently large k we have that $x_i^k = x_i^*$ for all $i \in \mathcal{Z}$. An important quantity in our analysis is the minimum distance to the nearest boundary of the subdifferential (3) among indices $i \in \mathcal{Z}$. This quantity is given by

$$\delta = \min_{i \in \mathcal{Z}} \left\{ \min\{-\nabla_i f(x^*) - l_i, u_i + \nabla_i f(x^*)\} \right\}. \tag{4}$$

3 Finite-time active-set identification

In this section we show that the PG method identifies the active-set of (1) in a finite number of iterations. Although this result follows from the more general results in the literature, by focusing on (1) and the case of strong-convexity we give a substantially simpler proof that will allow us to easily bound the active-set iteration complexity of the method.

Before proceeding to our main contributions, we state the linear convergence rate of the proximal gradient method to the (unique) solution x^*.

[1] Note that $|\nabla_i f(x^*)| \leq \lambda$ for all i with $x_i^* = 0$ follows from the optimality conditions, so this assumption simply rules out the case where $|\nabla_i f(x_i^*)| = \lambda$.

Theorem 1 *[22, Prop. 3] Consider problem* (1)*, where f is μ-strongly convex with L-Lipschitz continuous gradient, and the g_i are proper convex and lower semi-continuous. Then for every iteration $k \geq 1$ of the proximal gradient method, we have*

$$\|x^k - x^*\| \leq \left(1 - \frac{1}{\kappa}\right)^k \|x^0 - x^*\|, \tag{5}$$

where $\kappa := L/\mu$ is the condition number of f.

Next, we state the finite active-set identification result. Our argument essentially states that $\|x^k - x^*\|$ is eventually always less then $\delta/2L$, where δ is defined as in (4), and at this point the algorithm always sets x_i^k to x_i^* for all $i \in \mathcal{Z}$.

Lemma 1 *Consider problem* (1)*, where f is μ-strongly convex with L-Lipschitz continuous gradient, and the g_i are proper convex and lower semi-continuous. Let Assumption 1 hold for the solution x^*. Then for any proximal gradient method with a step-size of $1/L$, there exists a $\bar{k}$ such that for all $k > \bar{k}$ we have $x_i^k = x_i^*$ for all $i \in \mathcal{Z}$.*

Proof By the definition of the proximal gradient step and the separability of g, for all i we have

$$x_i^{k+1} \in \operatorname*{argmin}_y \left\{ \frac{1}{2} \left| y - \left(x_i^k - \frac{1}{L} \nabla_i f(x^k) \right) \right|^2 + \frac{1}{L} g_i(y) \right\}.$$

This problem is strongly-convex, and its unique solution satisfies

$$0 \in y - x_i^k + \frac{1}{L} \nabla_i f(x^k) + \frac{1}{L} \partial g_i(y),$$

or equivalently that

$$L(x_i^k - y) - \nabla_i f(x^k) \in \partial g_i(y). \tag{6}$$

By Theorem 1, there exists a minimum finite iterate $\bar{k}$ such that $\|x^{\bar{k}} - x^*\| \leq \delta/2L$. Since $|x_i^k - x_i^*| \leq \|x^k - x^*\|$, this implies that for all $k \geq \bar{k}$ we have

$$-\delta/2L \leq x_i^k - x_i^* \leq \delta/2L, \quad \text{for all } i. \tag{7}$$

Further, the Lipschitz continuity of ∇f in (2) implies that we also have

$$\begin{aligned} |\nabla_i f(x^k) - \nabla_i f(x^*)| &\leq \|\nabla f(x^k) - \nabla f(x^*)\| \\ &\leq L\|x^k - x^*\| \\ &\leq \delta/2, \end{aligned}$$

which implies that

$$-\delta/2 - \nabla_i f(x^*) \leq -\nabla_i f(x^k) \leq \delta/2 - \nabla_i f(x^*). \tag{8}$$

To complete the proof it is sufficient to show that for any $k \geq \bar{k}$ and $i \in \mathcal{Z}$ that $y = x_i^*$ satisfies (6). Since the solution to (6) is unique, this will imply the desired result. We first show that the left-side is less than the upper limit u_i of the interval $\partial g_i(x_i^*)$,

$$\begin{aligned} L(x_i^k - x_i^*) - \nabla_i f(x^k) &\leq \delta/2 - \nabla_i f(x^k) && \text{(right-side of (7))} \\ &\leq \delta - \nabla_i f(x^*) && \text{(right-side of (8))} \\ &\leq (u_i + \nabla_i f(x^*)) - \nabla_i f(x^*) && \text{(definition of } \delta\text{, (4))} \\ &\leq u_i. \end{aligned}$$

We can use the left-sides of (7) and (8) and an analogous sequence of inequalities to show that $L(x_i^k - x_i^*) - \nabla_i f(x^k) \geq l_i$, implying that x_i^* solves (6). □

4 Active-set complexity

The active-set identification property shown in the previous section could also be shown using the more sophisticated tools used in related works [7,13]. However, an appealing aspect of the simple argument above is that it is clear how to bound the active-set complexity of the method. We formalize this in the following result.

Corollary 1 *Consider problem* (1)*, where f is μ-strongly convex with L-Lipschitz continuous gradient, and the g_i are proper convex and lower semi-continuous. Let Assumption 1 hold for the solution x^*. Then the proximal gradient method with a step-size of $1/L$ identifies the active-set after at most $\kappa \log(2L\|x^0 - x^*\|/\delta)$ iterations.*

Proof Using Theorem 1 and $(1 - 1/\kappa)^k \leq \exp(-k/\kappa)$, we have

$$\|x^k - x^*\| \leq \exp(-k/\kappa)\|x^0 - x^*\|.$$

The proof of Lemma 1 shows that the active-set identification occurs whenever the inequality $\|x^k - x^*\| \leq \delta/2L$ is satisfied. For this to be satisfied, it is sufficient to have

$$\exp(-k/\kappa)\|x^0 - x^*\| \leq \frac{\delta}{2L}.$$

Taking the log of both sides and solving for k gives the result. □

It is interesting to note that this bound only depends logarithmically on $1/\delta$, and that if δ is quite large we can expect to identify the active-set very quickly. This $O(\log(1/\delta))$ dependence is in contrast to the previous result of Liang et al. who give a bound of the form $O(1/\sum_{i=1}^n \delta_i^2)$ where δ_i is the distance of $\nabla_i f$ to the boundary of the subdifferential ∂g_i at x^* [17, Proposition 3.6]. Thus, our bound will typically be tighter as it only depends logarithmically on the single smallest δ_i (though we make the extra assumption of strong-convexity). In Section 1, we considered two specific cases of problem (1), for which we can define δ:

1. If the g_i enforce non-negativity constraints, then $\delta = \min_{i \in \mathcal{Z}} \nabla_i f(x^*)$.
2. If g is a scaled ℓ_1-regularizer, then $\delta = \lambda - \max_{i \in \mathcal{Z}} |\nabla_i f(x^*)|$.

In the first case we identify the non-zero variables after $\kappa \log(2L\|x^0 - x^*\|/\min_{i \in \mathcal{Z}} \nabla_i f(x^*))$ iterations. If the minimum gradient over the active-set at the solution δ is zero, then we may approach the active-set through the interior of the constraint and the active-set may never be identified (this is the purpose of the nondegeneracy condition). Similarly, for ℓ_1-regularization this result also gives an upper bound on how long it takes to identify the sparsity pattern.

Above we have bounded the number of iterations before $x_i^k = x_i^*$ for all $i \in \mathcal{Z}$. However, in the non-negative and L1-regularized applications we might also be interested in the number of iterations before we always have $x_i^k \neq 0$ for all $i \notin \mathcal{Z}$. More generally, the number of iterations before x_i^k for $i \notin \mathcal{Z}$ are not located at non-smooth or boundary values. It is straightforward to bound this quantity. Let $\Delta = \min_{i \notin Z}\{|x_i^n - x_i^*|\}$ where x_i^n is the nearest non-smooth or boundary value along dimension i. Since (5) shows that the proximal-gradient method contracts the distance to x^*, it cannot set values x_i^k for $i \notin \mathcal{Z}$ to non-smooth or boundary values once $\|x^k - x^*\| \leq \Delta$. It follows from (5) that $\kappa \log(\|x^0 - x^*\|/\Delta)$ iterations are needed for the values $i \notin \mathcal{Z}$ to only occur at smooth/non-boundary values.

5 General step-size

The previous sections considered a step-size of $1/L$. In this section we extend our results to handle general constant step-sizes, which leads to a smaller active-set complexity if we use a larger step-size depending on μ. To do this, we require the following result, which states the generalized convergence rate bound for the proximal gradient method. This result matches the known rate of the gradient method with a constant step-size for solving strictly-convex quadratic problems [4, §1.3], and the rate of the projected-gradient algorithm with a constant step-size for minimizing strictly-convex quadratic functions over convex sets [4, §2.3].

Theorem 2 *Consider problem* (1)*, where f is μ-strongly convex with L-Lipschitz continuous gradient, and g is proper convex and lower semi-continuous. Then for every iteration $k \geq 1$ of the proximal gradient method with a constant step-size $\alpha > 0$, we have*

$$\|x^k - x^*\| \leq Q(\alpha)^k \|x^0 - x^*\|, \tag{9}$$

where $Q(\alpha) := \max\{|1 - \alpha L|, |1 - \alpha\mu|\}$.

We give the proof in the Appendix. Theorem 1 is a special case of Theorem 2 since $Q(1/L) = 1 - \mu/L$. Further, Theorem 2 gives a faster rate if we minimize Q in terms of α to give $\alpha = 2/(L + \mu)$, which yields a faster rate of

$$Q\left(\frac{2}{L+\mu}\right) = 1 - \frac{2\mu}{L+\mu} = \frac{L-\mu}{L+\mu}.$$

This faster convergence rate for the proximal gradient method may be of independent interest in other settings, and we note that this result does not require g to be separable. We also note that, although the theorem is true for any positive α, it is only interesting for $\alpha < 2/L$ since for $\alpha \geq 2/L$ it does not imply convergence.

Lemma 2 *Consider problem* (1)*, where f is μ-strongly convex with L-Lipschitz continuous gradient, and the g_i are proper convex and lower semi-continuous. Let Assumption 1 hold for the solution x^*. Then for any proximal gradient method with a constant step-size $0 < \alpha < 2/L$, there exists a $\bar{k}$ such that for all $k > \bar{k}$ we have $x_i^k = x_i^*$ for all $i \in \mathcal{Z}$.*

We give the proof Lemma 2 in the Appendix, which shows that we identify the active-set when $\|x^k - x^*\| \leq \delta\alpha/3$ is satisfied. Using this result, we prove the following active-set complexity result for proximal gradient methods when using a general fixed step-size (the proof is once again found in the Appendix).

Corollary 2 *Consider problem* (1)*, where f is μ-strongly convex with L-Lipschitz continuous gradient (for $\mu < L$), and the g_i are proper convex and lower semi-continuous. Let Assumption 1 hold for the solution x^*. Then for any proximal gradient method with a constant step-size α, such that $0 < \alpha < 2/L$, the active-set will be identified after at most $\frac{1}{\log(1/Q(\alpha))} \log(3||x^0 - x^*||/(\delta\alpha))$ iterations.*

Finally, we note that as part of a subsequent work we have analyzed the active-set complexity of block coordinate descent methods [21]. The argument in that case is similar to the argument presented here. The main modification needed to handle coordinate-wise updates is that we must use a coordinate selection strategy that guarantees that we eventually select all $i \in \mathcal{Z}$ that are not at their optimal values for some finite $k \geq \bar{k}$.

Acknowledgements The authors would like to express their thanks to the anonymous referees for their valuable feedback.

Appendix

Proof of Theorem 2. For any $\alpha > 0$, by the non-expansiveness of the proximal operator [8, Lem 2.4] and the fact that x^* is a fixed point of the proximal gradient update for any $\alpha > 0$, we have

$$\begin{aligned}
&\|x^{k+1} - x^*\|^2 \\
&\quad = \|\mathbf{prox}_{\alpha g}(x^k - \alpha\nabla f(x^k)) - \mathbf{prox}_{\alpha g}(x^k - \alpha\nabla f(x^*))\|^2 \\
&\quad \leq \|(x^k - \alpha\nabla f(x^k)) - (x^* - \alpha\nabla f(x^*))\|^2 \\
&\quad = \|x^k - x^* - \alpha(\nabla f(x^k) - \nabla f(x^*))\|^2 \\
&\quad = \|x^k - x^*\|^2 - 2\alpha\langle\nabla f(x^k) - \nabla f(x^*), x^k - x^*\rangle + \alpha^2\|\nabla f(x^k) - \nabla f(x^*)\|^2.
\end{aligned}$$

By the L-Lipschitz continuity of ∇f and the μ-strong convexity of f, we have [19, Thm 2.1.12]

$$\langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle \geq \frac{1}{L+\mu} \|\nabla f(x^k) - \nabla f(x^*)\|^2 + \frac{L\mu}{L+\mu} \|x^k - x^*\|^2,$$

which yields

$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \frac{2\alpha L\mu}{L+\mu}\right) \|x^k - x^*\|^2 + \alpha \left(\alpha - \frac{2}{L+\mu}\right) \|\nabla f(x^k) - \nabla f(x^*)\|^2.$$

Further, by the μ-strong convexity of f, we have for any $x, y \in \mathbb{R}^n$ [19, Thm 2.1.17],

$$\langle \nabla f(x) - \nabla f(y) \rangle \geq \mu \|x - y\|^2,$$

which by Cauchy-Schwartz gives

$$\|\nabla f(x) - \nabla f(y)\| \geq \mu \|x - y\|.$$

Combining this with the L-Lipschitz continuity condition in (2) shows that $\mu \leq L$. Therefore, for any $\beta \in \mathbb{R}$ (positive or negative) we have

$$\beta \|\nabla f(x) - \nabla f(y)\|^2 \leq \max\{\beta L^2, \beta \mu^2\} \|x - y\|^2.$$

Thus, for $\beta := \left(\alpha - \frac{2}{L+\mu}\right)$, we have

$$\begin{aligned}
&\|x^{k+1} - x^*\|^2 \\
&\quad \leq \left(1 - \frac{2\alpha L\mu}{L+\mu}\right) \|x^k - x^*\|^2 + \alpha \max\left\{L^2\beta, \mu^2\beta\right\} \|x^k - x^*\|^2 \\
&\quad = \max\left\{\left(1 - \frac{2\alpha L\mu}{L+\mu}\right) + \alpha L^2\beta, \left(1 - \frac{2\alpha L\mu}{L+\mu}\right) + \alpha\mu^2\beta\right\} \|x^k - x^*\|^2 \\
&\quad = \max\left\{1 - \frac{2\alpha L(L+\mu)}{L+\mu} + \alpha^2 L^2, 1 - \frac{2\alpha\mu(L+\mu)}{L+\mu} + \alpha^2\mu^2\right\} \|x^k - x^*\|^2 \\
&\quad = \max\left\{(1 - \alpha L)^2, (1 - \alpha\mu)^2\right\} \|x^k - x^*\|^2 \\
&\quad = Q(\alpha)^2 \|x^k - x^*\|^2.
\end{aligned}$$

Taking the square root and applying it repeatedly, we obtain our result. □

Proof of Lemma 2. By the definition of the proximal gradient step and the separability of g, for all i we have

$$x_i^{k+1} \in \operatorname*{argmin}_y \left\{\frac{1}{2} \left|y - \left(x_i^k - \alpha \nabla_i f(x^k)\right)\right|^2 + \alpha g_i(y)\right\}.$$

This problem is strongly-convex with a unique solution that satisfies

$$\frac{1}{\alpha}(x_i^k - y) - \nabla_i f(x^k) \in \partial g_i(y). \tag{10}$$

By Theorem 2 and $\alpha < 2/L$, there exists a minimum finite iterate $\bar{k}$ such that $\|x^{\bar{k}} - x^*\| \leq \delta\alpha/3$. Following similar steps as in Lemma 1, this implies that

$$-\delta\alpha/3 \leq x_i^k - x_i^* \leq \delta\alpha/3, \quad \text{for all } i, \tag{11}$$

and by the Lipschitz continuity of ∇f, we also have

$$-\delta\alpha L/3 - \nabla_i f(x^*) \leq -\nabla_i f(x^k) \leq \delta\alpha L/3 - \nabla_i f(x^*). \tag{12}$$

To complete the proof it is sufficient to show that for any $k \geq \bar{k}$ and $i \in \mathcal{Z}$ that $y = x_i^*$ satisfies (10). We first show that the left-side is less than the upper limit u_i of the interval $\partial g_i(x_i^*)$,

$$\begin{aligned}
\frac{1}{\alpha}(x_i^k - x_i^*) - \nabla_i f(x^k) &\leq \delta/3 - \nabla_i f(x^k) && \text{(right-side of (11))} \\
&\leq \delta(1+\alpha L)/3 - \nabla_i f(x^*) && \text{(right-side of (12))} \\
&\leq \delta - \nabla_i f(x^*) && \text{(upper bound on } \alpha\text{)} \\
&\leq (u_i + \nabla_i f(x^*)) - \nabla_i f(x^*) && \text{(definition of } \delta\text{, (4))} \\
&\leq u_i.
\end{aligned}$$

Using the left-sides of (11) and (12), and an analogous sequence of inequalities, we can show that $\frac{1}{\alpha}(x_i^k - x_i^*) - \nabla_i f(x^k) \geq l_i$, implying that x_i^* solves (10). Since the solution to (10) is unique, this implies the desired result. □

Proof of Corollary 2. By Theorem 2, we know that the proximal gradient method achieves the following linear convergence rate,

$$\|x^{k+1} - x^*\| \leq Q(\alpha)^k \|x^0 - x^*\|.$$

The proof of Lemma 2 shows that the active-set identification occurs whenever the inequality $\|x^k - x^*\| \leq \delta\alpha/3$ is satisfied. Thus, we want

$$Q(\alpha)^k \|x^0 - x^*\| \leq \frac{\delta\alpha}{3}.$$

Taking the log of both sides, we obtain

$$k \log\left(Q(\alpha)\right) + \log\left(\|x^0 - x^*\|\right) \leq \log\left(\frac{\delta\alpha}{3}\right).$$

Noting that $0 < Q(\alpha) < 1$ so $\log(Q(\alpha)) < 0$, we can rearrange to obtain

$$k \geq \frac{1}{\log\left(Q(\alpha)\right)} \log\left(\frac{\delta\alpha}{3\|x^0 - x^*\|}\right) = \frac{1}{\log(1/Q(\alpha))} \log\left(\frac{3\|x^0 - x^*\|}{\delta\alpha}\right).$$

□

References

1. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
2. Bertsekas, D.P.: On the Goldstein-Levitin-Polyak gradient projection method. IEEE Trans. Autom. Control **21**(2), 174–184 (1976)
3. Bertsekas, D.P.: Convex Optimization Algorithms. Athena Scientific Belmont (2015)
4. Bertsekas, D.P.: Nonlinear Programming, 3rd edn. Athena Scientific (2016)
5. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
6. Buchheim, C., De Santis, M., Lucidi, S., Rinaldi, F., Trieu, L.: A feasible active set method with reoptimization for convex quadratic mixed-integer programming. SIAM J. Optim. **26**(3), 1695–1714 (2016)
7. Burke, J.V., Moré, J.J.: On the identification of active constraints. SIAM J. Numer. Anal. **25**(5), 1197–1211 (1988)
8. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. **4**, 1168–1200 (2005)
9. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**, 273–297 (1995)
10. Curtis, F.E., Han, Z., Robinson, D.P.: A globally convergent primal-dual active-set framework for large-scale convex quadratic optimization. Comput. Optim. Appl. **60**, 311–341 (2015)
11. De Santis, M., Lucidi, S., Rinaldi, F.: A fast active set block coordinate descent algorithm for ℓ_1-regularized least squares. SIAM J. Optim. **26**(1), 781–809 (2016)
12. Hare, W.L.: Identifying active manifolds in regularization problems. In: H.H. Bauschke, R.S. Burachik, P.L. Combettes, V. Elser, D.R. Luke, H. Wolkowicz (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 261–271. Springer New York, New York, NY (2011)
13. Hare, W.L., Lewis, A.S.: Identifying active constraints via partial smoothness and prox-regularity. J. Convex Analysis **11**(2), 251–266 (2004)
14. Krishnan, D., Lin, P., Yip, A.M.: A primal-dual active-set method for non-negativity constrained total variation deblurring problems. IEEE Trans. Image Process. **16**(11), 2766–2777 (2007)
15. Lee, S., Wright, S.J.: Manifold identification in dual averaging for regularized stochastic online learning. J. Mach. Learn. Res. **13**(1), 1705–1744 (2012)
16. Levitin, E.S., Polyak, B.T.: Constrained minimization methods. USSR Comput. Math. Math. Phys. **6**, 1–50 (1966)
17. Liang, J., Fadili, J., Peyré, G.: Activity identification and local linear convergence of forward–backward-type methods. SIAM J. Optim. **27**(1), 408–437 (2017)
18. Mifflin, R., Sagastizábal, C.: Proximal points are on the fast track. J. Convex Analysis **9**(2), 563–579 (2002)
19. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Dordrecht, The Netherlands (2004)
20. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. Ser. B **140**, 125–161 (2013)
21. Nutini, J., Laradji, I., Schmidt, M.: Let's make block coordinate descent go fast: Faster greedy rules, message-passing, active-set complexity, and superlinear convergence. arXiv:1712.08859 (2017)
22. Schmidt, M., Roux, N.L., Bach, F.R.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Proceedings of the 24th International Conference on Neural Information Processing Systems, pp. 1458–1466. Grenada, Spain (2011)
23. Solntsev, S., Nocedal, J., Byrd, R.H.: An algorithm for quadratic ℓ_1-regularized optimization with flexible active-set strategy. Optim. Method Softw. **30**(6), 1213–1237 (2015)
24. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Royal Stat. Soc., Series B **58**(1), 267–288 (1996)
25. Wen, Z., Yin, W., Goldfarb, D., Zhang, Y.: A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation. SIAM J. Sci. Comput. **32**(4), 1832–1857 (2010)

26. Wright, S.J.: Identifiable surfaces in constrained optimization. SIAM J. Control Optim. **31**(4), 1063–1079 (1993)
27. Wright, S.J.: Accelerated block-coordinate relaxation for regularized optimization. SIAM J. Optim. **22**(1), 159–186 (2012)