"Active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern?

Julie Nutini jnutini@cs.ubc.ca
University of British Columbia
Mark Schmidt schmidtm@cs.ubc.ca
University of British Columbia
Warren Hare warren.hare@ubc.ca
University of British Columbia

Abstract

Proximal gradient methods have been found to be highly effective for solving minimization problems with non-negative constraints or L1-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.

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 convex and lower semi-continuous (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 L2-regularized variant to guarantee strong-convexity). A common choice of the g_i is a scaled absolute value function $g_i(x_i) = \lambda|x_i|$, which yields a sparsity-encouraging L1-regularization term. The g_i can also enforce bound constraints, 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, 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

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

When the proximal gradient method is applied with non-negative constraints or L1-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: after we have identified the set of non-zero variables we could switch to a more

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

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.

These ideas date back at least 40 years to the work of Bertsekas [1976] 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 this idea is true 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 [Burke and Moré, 1988, Wright, 1993, Hare and Lewis, 2004, Hare, 2011]. The active-set identification property has also been shown for other algorithms like certain coordinate descent and stochastic gradient methods [Mifflin and Sagastizábal, 2002, Wright, 2012, Lee and Wright, 2012].

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.

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 sub-differentials of each g_i. Further, the subdifferential of each individual g_i at any $x_i \in \mathbb{R}$ is defined by

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

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 x_i can be written as an open interval,

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

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 [Hare and Lewis, 2004], 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 nondegeneracy condition for the special case of (1).

Definition 1. *A solution x^* of the problem (1) is* nondegenerate *if and only if*

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

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 L1-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 [De Santis et al., 2016].[1]

Definition 2. *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. 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 $i \in \mathcal{Z}$.

[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 $|\nabla_i f(x_i^*)| = \lambda$.

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) we give a substantially simpler proof that will allow us to bound the active-set iteration complexity of the method.

Before proceeding to our main contributions, we note that our assumptions imply that the iterates converge to the (unique) solution x^* with a linear rate [Schmidt et al., 2011],

$$\|x^k - x^*\| \leq \left(1 - \frac{1}{\kappa}\right)^k \|x^0 - x^*\|, \tag{4}$$

where κ is the condition number of f. 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

$$\delta = \min_{i \in \mathcal{Z}} \left\{ \min\{-\nabla_i f(x^*) - l_i, u_i + \nabla_i f(x^*)\} \right\}. \tag{5}$$

Our argument essentially states that $\|x^k - x^*\|$ is eventually always less then $\delta/2L$, and at this point the algorithm always sets x_i^k to x_i^* for all $i \in \mathcal{Z}$.

Lemma 1. *Suppose we apply the proximal gradient method with a step-size of $1/L$ to problem (1), under the assumptions discussed in the introduction. If the solution x^* is nondegenerate then 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

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

or equivalently that

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

By the bound on the proximal gradient iterates (4), 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 $\partial g_i(x_i^*)$,

$$\begin{aligned} L(x_i^k - x_i^*) - \nabla_i f(x^k) &\leq \delta/2 - \nabla_i f(x^k) && \text{(right-side of (7))} \\ &\leq \delta - \nabla_i f(x^*) && \text{(right-side of (8))} \\ &\leq (u_i + \nabla_i f(x^*)) - \nabla_i f(x^*) && \text{(definition of } \delta\text{, (5))} \\ &\leq u_i. \end{aligned}$$

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 property shown in the previous section could also be shown using the more sophisticated tools used in related works [Burke and Moré, 1988, Hare and Lewis, 2004]. However, an appealing aspect of the simple argument above is that it is clear how to bound the active-set complexity of the method. In particular, the proof shows that the active-set identification occurs whenever the inequality $\|x^k - x^*\| \leq \delta/2L$ is satisfied. By using that $(1-\kappa)^k \leq \exp(-\kappa k)$, the linear convergence rate (4) implies the following result.

Corollary 1. *The active-set will be identified after at most* $\kappa \log(2L\|x^0 - x^*\|/\delta)$ *iterations.*

Considering our introductory statements, we have two special cases. First,

$$\text{if } g_i \text{ are non-negative constraints, then } \delta = \min_{i \in \mathcal{Z}} \nabla_i f(x^*). \tag{9}$$

Second,

$$\text{if } g \text{ is the L1-regularization function, then } \delta = \lambda - \max_{i \in \mathcal{Z}} |\nabla_i f(x^*)|. \tag{10}$$

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 this minimum 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 L1-regularization this result also gives an upper bound on how long it takes to identify the sparsity pattern.

It is interesting to note that this bound only depends logarithmically on δ, and that if δ is quite large then we can expect to identify the active-set very quickly. We note that the argument can also be modified to use other step-sizes, provided that we can write the algorithm in terms of a step-size that is guaranteed to be bounded from below. The argument can also be modified to analyze coordinate descent methods, provided that we can guarantee that all coordinates $i \in \mathcal{Z}$ that are not already active are eventually selected by the method for some $k \geq \bar{k}$.

References

D. P. Bertsekas. On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Trans. Autom. Control*, 21 (2):174–184, 1976.

J. V. Burke and J. J. Moré. On the identification of active constraints. *SIAM J. Numer. Anal.*, 25(5):1197–1211, 1988.

M. De Santis, S. Lucidi, and F. Rinaldi. A fast active set block coordinate descent algorithm for ℓ_1-regularized least squares. *SIAM J. Optim.*, 26(1):781–809, 2016.

W. L. Hare. Identifying active manifolds in regularization problems. In H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 261–271. Springer New York, New York, NY, 2011.

W. L. Hare and A. S. Lewis. Identifying active constraints via partial smoothness and prox-regularity. *J. Convex Analysis*, 11(2):251–266, 2004.

S. Lee and S. J. Wright. Manifold identification in dual averaging for regularized stochastic online learning. *J. Mach. Learn. Res.*, 13(1):1705–1744, 2012.

R. Mifflin and C. Sagastizábal. Proximal points are on the fast track. *J. Convex Analysis*, 9(2):563–579, 2002.

M. Schmidt, N. L. Roux, and F. R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 1458–1466, Grenada, Spain, 2011.

S. J. Wright. Identifiable surfaces in constrained optimization. *SIAM J. Control Optim.*, 31(4):1063–1079, 1993.

S. J. Wright. Accelerated block-coordinate relaxation for regularized optimization. *SIAM J. Optim.*, 22(1): 159–186, 2012.