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

Julie Nutini (UBC), Mark Schmidt (UBC) and Warren Hare (UBC Okanagan)

OVERVIEW: Asymptotic finite identification of active-set

Motivation:

- Idea of active-set identification dates back at least 40 years to the work of Bertsekas.
  - → **Faster** if only want to find sparsity pattern (do not have to run to convergence).
  - → **Faster** if switch to solver like Newton's method on non-zero variables (superlinear).
- Prior works show active-set identification happens after some finite number of iterations.
  - → **Question:** When exactly is this guaranteed?

This Work:

- ⋆ Give new simple analysis for active-set identification of proximal gradient methods.
- ⋆ Introduce the notion of the *active-set complexity* of an algorithm.
  - → Number of iterations required before algorithm is guaranteed to have reached active-set.
- ⋆ Derive explicit bounds on active-set complexity of proximal gradient methods.

Example

- Consider applying proximal gradient methods on an L1-regularized problem,

$$\operatorname*{argmin}_{x\in\mathbb{R}^n} f(x) + \lambda\|x\|_1.$$

- Under mild assumptions, x^k matches sparsity pattern of x^* for all sufficiently large k.

$$x^0 = \begin{pmatrix} x_1^0 \\ x_2^0 \\ x_3^0 \\ x_4^0 \\ x_5^0 \end{pmatrix} \quad \xrightarrow{\text{after finite } k \text{ iterations}} \quad x^k = \begin{pmatrix} x_1^k \\ 0 \\ 0 \\ x_4^k \\ 0 \end{pmatrix}, \quad \text{where} \quad x^* = \begin{pmatrix} x_1^* \\ 0 \\ 0 \\ x_4^* \\ 0 \end{pmatrix}$$

Algorithm

- We consider the general optimization problem

$$\operatorname*{argmin}_{x\in\mathbb{R}^n} \quad f(x) + \sum_{i=1}^n g_i(x_i), \tag{1}$$

where each g_i is separable, convex and lower semi-continuous (not necessarily smooth).

- In machine learning, common examples are:
  - → f is an (L2-regularized) quadratic, $f(x) = \frac{1}{2}\|Ax-b\|^2 \;\; (\; + \; \|x\|^2 \;)$
  - → (Non-negative constraints) g_i is the indicator function on the non-negative orthant,

$$\delta_{\geq 0}(x_i) = \begin{cases} 0 & \text{if } x_i \geq 0, \\ \infty & \text{if } x_i < 0. \end{cases}$$

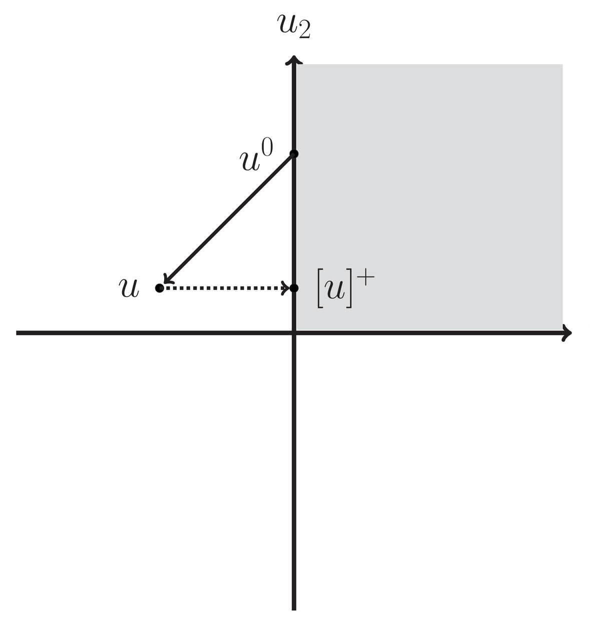
  - → (L1-regularized problem) $g_i = \lambda|x_i|$ and encourages sparsity in the solution.
- The proximal gradient (PG) method uses an iteration update given by

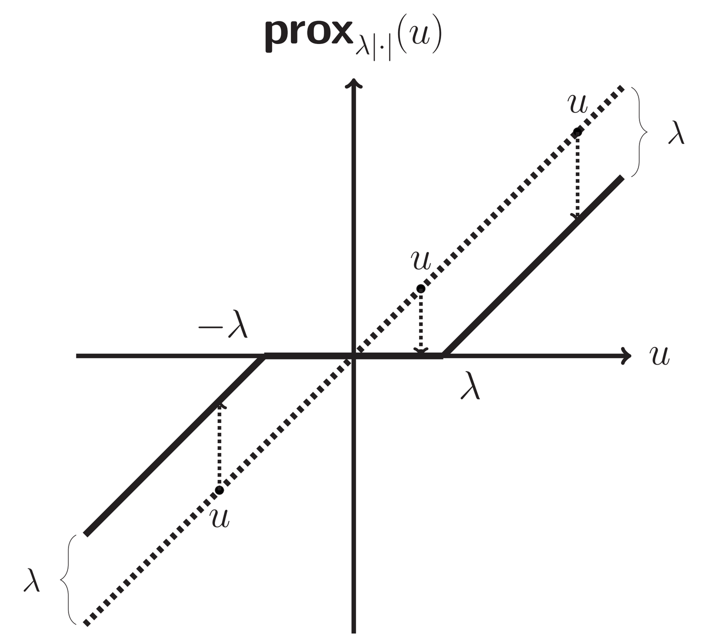
$$x^{k+1} = \mathbf{prox}_{\frac{1}{L}g}\left(x^k - \frac{1}{L}\nabla f(x^k)\right),$$

where the proximal operator is defined as

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

  - → For non-negative constraints: $\mathbf{prox}_{\geq 0}(u) = \mathbf{proj}_{\geq 0}(u) = [u]^+$.
  - → For L1-regularization: $\mathbf{prox}_{\lambda|\cdot|}(u) = \frac{u}{|u|}[|u| - \lambda]^+$.





Assumptions

- We assume the gradient ∇f is L-Lipschitz continuous,

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

and that f is μ-strongly convex,

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

- By the separability of g, the subdifferential of g at any $x \in \mathbb{R}^n$ is given by

$$\partial g(x) = \{(v_1, v_2, \ldots, v_n) \in \mathbb{R}^n : v_i \in \partial g_i(x_i)\},$$

where

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

- The subdifferential of each g_i is an interval on the real line and the interior 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 $u_i \in \mathbb{R}\cup\{\infty\}$.

- We require the nondegeneracy condition for problem (1) to hold at solution x^*:

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

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

  - → For non-negative constraints, requires $\nabla_i f(x^*) > 0$ for all variables i with $x_i^* = 0$.
  - → For L1-regularization, requires $|\nabla_i f(x^*)| < \lambda$ for all variables i with $x_i^* = 0$.

→ By our assumptions, the PG method converges to a unique solution x^* at a linear rate,

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

where κ is the condition number of f.

Active-Set Identification

Definition

The *active-set* for separable g is defined as the set $\mathcal{Z} = \{i : \partial g_i(x_i^*)$ is not a singleton$\}$.

→ By (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.

Definition

The *minimum distance to the nearest boundary of the subdifferential* (3) for all $i \in \mathcal{Z}$,

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

- → For non-negative constraints, we have $\delta = \min_{i\in\mathcal{Z}} \nabla_i f(x^*)$.
- → For L1-regularization, we have $\delta = \lambda - \max_{i\in\mathcal{Z}} |\nabla_i f(x^*)|$.

Definition

The *active-set identification property* for problem (1) is satisfied if for all sufficiently large k, we have that $x_i^k = x_i^*$ for all $i \in \mathcal{Z}$.

→ 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

Suppose we apply the PG method with step-size of $1/L$ to problem (1). 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 PG 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

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

By (4), there exists a minimum finite iterate $\bar{k}$ such that

$$|x_i^{\bar{k}} - x_i^*| \leq \|x^{\bar{k}} - x^*\| \leq \delta/2L,$$

which implies that for all $k \geq \bar{k}$ we have

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

By the Lipschitz continuity of ∇f in (2), we have that

$$|\nabla_i f(x^k) - \nabla_i f(x^*)| \leq \|\nabla f(x^k) - \nabla f(x^*)\| \leq L\|x^k - x^*\| \leq \delta/2,$$

which implies that

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

Finally, it is sufficient to show that for any $k \geq \bar{k}$ and $i \in \mathcal{Z}$ that $y = x_i^*$ satisfies (6). We first show that the left-side is less than the upper limit u_i of the interval $\partial g_i(x_i^*)$,

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

Similar steps using LHS of (7) and (8) shows that $L(x_i^k - x_i^*) - \nabla_i f(x^k) \geq l_i$. □

Active-Set Complexity

Definition

The *active-set complexity* is the number of iterations required before an algorithm is guaranteed to have reached the active-set.

- In our Lemma, we show that active-set identification occurs when $\|x^k - x^*\| \leq \delta/2L$.
- Using $(1-\kappa)^k \leq \exp(-\kappa k)$, the linear convergence rate (4) implies the following result.

Corollary

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

- → Bound only depends logarithmically on δ.
  - if δ is large, then we can expect to identify the active-set very quickly.
- → Can be modified to use other step-sizes and to analyze coordinate descent methods.