“Active-set complexity” of proximal-gradient

How long does it take to find the sparsity pattern?

Julie Nutini, Mark Schmidt, Warren Hare

University of British Columbia

Motivation: L1-Regularized Optimization with Proximal-Gradient Method

- Optimization with L1-regularization is widely-studied in various fields,

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

where in this talk we'll assume that ∇f is Lipschitz and f is strongly-convex.

- Key advantage over classic L2-regularization: solution x^* is sparse.
  - It tends to have many values x_i^* equal to exactly 0.

- Proximal-gradient methods are among most widely-used solvers.

$$x^{k+1} = \text{prox}\left(x^k - \alpha_k \nabla f(x^k)\right),$$

where the proximal operator is given by

$$\text{prox}(x) = \operatorname*{argmin}_y \frac{1}{2}\|y - x\|^2 + \alpha_k \lambda \|x\|_1.$$

Active-Set Identification

- With mild assumptions: proximal-gradient "identifies" active set in finite time:
  - For all sufficiently large k, sparsity pattern of x^k matches sparsity pattern of x^*.

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

- Useful if we are only interested in finding the sparsity pattern.
- Convergence rate will be faster once this happens (optimizing over subspace).
  - You could also apply Newton-like methods on the non-zero variables.

Related Work and More-General Results

- Idea of finitely identifying non-zeroes dates back (at least) to Bertskeas [1976].
  - For projected-gradient applied to smooth functions with non-negative constraints.
- Has been shown for a variety of convex/non-convex problems.
  - Burke & Moré [1988], Wright [1993], Hare & Lewis [2004], Hare [2011].
- Has been shown for a variety of other algorithms.
  - Includes certain coordinate descent and stochastic gradient methods.
  - Mifflin & Sagastizábal [2002], Wright [2012], Lee & Wright [2012]

Active-Set Complexity: How long does it take to find the sparsity pattern?

- These prior works only show that identification happens asymptotically.
  - For some finite but unknown k.

- In this work we introduce the notion of "active-set complexity" of an algorithm:
  - The number of iterations before it is guaranteed to have reached the active set.

- We bound active-set complexity of proximal-gradient with separable regularizers.
  - Under the standard non-degeneracy condition required for identification to happen.

- We are only aware of one previous work giving such bounds, Liang et al. [2017].
  - We make stronger assumptions on f (strong-convexity which gives a faster rate).
  - But weaker assumptions on regularizer (no inclusion condition on subdifferential).

Special Case: Optimizing with Non-Negative Constraints

- We will first consider optimization with non-negative constraints,

$$\operatorname*{argmin}_{x \geq 0} f(x),$$

using the projected-gradient method with a step-size of $1/L$,

$$x^{k+1} = \left[x^k - \frac{1}{L}\nabla f(x^k)\right]^+.$$

- This also leads to sparsity, and we use $\mathcal{Z}$ as the indices i where $x_i^* = 0$.

- We'll assume:
  1. Gradient ∇f is L-Lipschitz continuous.
  2. Function f is μ-strongly convex.
  3. Non-degeneracy condition: for all $i \in \mathcal{Z}$ we have $\nabla f(x_i^*) \geq \delta$ for some $\delta > 0$.
     - "You can't have $\nabla_i f(x^*) = 0$ for variables i that are supposed to be zero."
     - This condition is standard: prevents reaching solution through interior.

Active-Set Identification for Non-Negative Constraints

- Let's show that we set $i \in \mathcal{Z}$ to zero when we're "close" to the solution.

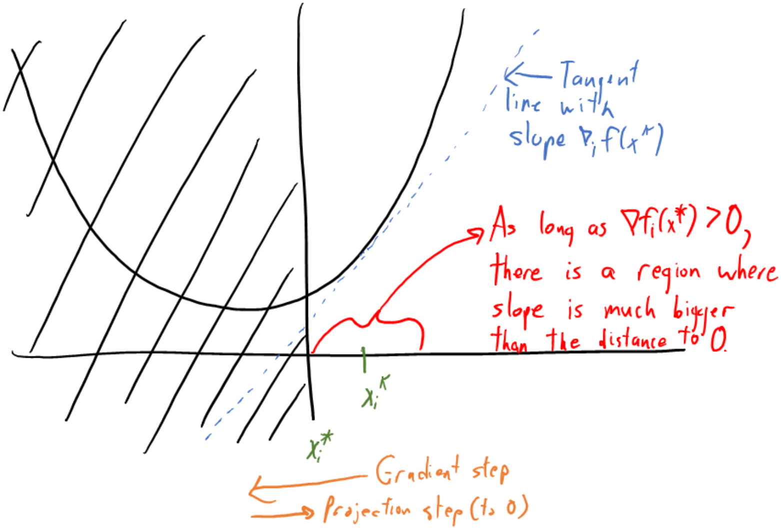
- Consider an iteration k where we have $\|x^k - x^*\| \leq \frac{\delta}{2L}$.

- In this region we have two useful properties for all $i \in \mathcal{Z}$:
  1. The value of the variable must be small: $x_i^k \leq \frac{\delta}{2L}$.
     - Since $x_i^* = 0$ and x_i^k is within $\delta/2L$ of x_i.
  2. The value of the gradient must be large: $\nabla_i f(x^k) \geq \delta/2$.
     - Since $\nabla_i f(x^*) \geq \delta$ and ∇f is Lipschitz.

- Plugging these into the projected-gradient update gives for $i \in \mathcal{Z}$ that

$$x_i^{k+1} = \left[x_i^k - \frac{1}{L}\nabla_i f(x^k)\right]^+ \leq \left[\frac{\delta}{2L} - \frac{\delta}{2L}\right]^+ = 0.$$

Active-Set Identification for Non-Negative Constraints



Active-Set Complexity for Non-Negative Constraints

- Under our assumptions it is known that the iterates converge linearly,

$$\|x^k - x^*\| \le (1 - \kappa^{-1})^k \|x^0 - x^*\|,$$

  where the condition number κ is L/μ.

- Thus, for all sufficiently large k we have $\|x^k - x^*\| \le \frac{\delta}{2L}$.
  - For these k the algorithm will have the correct active set.

- Using $(1 - \kappa^{-1})^k \le \exp(-k/\kappa)$ and solving for k gives

$$\kappa \log(2L\|x^0 - x^*\|/\delta),$$

  so we find the sparsity pattern after this many iterations ("active-set complexity").

Active-Set Complexity for Non-Smooth Regularizers

- Paper generalizes argument to lower/upper bounds and a separable regularizer,

$$\operatorname*{argmin}_{l \leq x \leq u} f(x) + \sum_{i=1}^{n} g_i(x_i).$$

- Key differences:
  - The set $\mathcal{Z}$ will be variables occuring at bounds or non-smooth points.
    - For L1-regularization this is again the variables with $x_i^* = 0$.
  - The quantity δ will be the "minimum distance to the sub-differential boundary",

$$\delta = \min_{i \in \mathcal{Z}} \{\min\{-\nabla_i f(x^*) - \min\{\partial g_i(x_i^*)\}, \max\{\partial g_i(x_i^*)\} + \nabla_i f(x^*)\}\}.$$

    - For L1-regularization this is $\delta = \lambda - \max_{i \in \mathcal{Z}}\{|\nabla f_i(x^*)|\}$.
  - The non-degeneracy condition is that $\delta > 0$.
    - For L1-regularization we require $|\nabla_i f(x^*)| \neq \lambda$ for $i \in \mathcal{Z}$.
  - Proof needs to bound x_i^k from above and below based on $\partial g_i(x_i^*)$.

Discussion

- Bound only depends logarithmically on δ:
  - If δ is large we can expect to identify the active-set very quickly.
- Our $O(\log(1/\delta))$ bound will tend to be faster than previous $O(1/\sum_{i=1}^{n} \delta_i^2)$.
  - Logarithmic dependence on smallest δ_i, but we assumed strong-convexity.
- In the paper we also analyze a general step-size $\alpha_k < 2/L$.
  - Can give faster rate, and argument is similar but result is a bit uglier.
- In the paper we bound complexity for $i \notin \mathcal{Z}$ to *not* get set to 0.
- Argument easily extends to group-separable regularizers.
- Can be extended to accelerated proximal-gradient and Newton-proximal.
  - Open problem: can we design new algorithms with lower active-set complexity?

Coordinate Descent (is a bit weird for active-set complexity)

- More recent work: active-set complexity of block coordinate descent.
  - This work has made me think about why newer algorithms might be found.

- Key differences when you analyse active-set complexity of coordinate descent:
  - The radius where we identify the active set is larger than $\delta/2L$.
    - Because you can use larger step-sizes in coordinate descent.
  - You don't identify active set immediately when you enter the radius.
    - You still have to select all sub-optimal $i \in \mathcal{Z}$ ("coupon collecting").

- Coupon collecting for different coordinate selection strategies:
  - Cyclic selection: n iterations.
  - Random: $O(n \log n)$ iterations for uniform, can be much higher for non-uniform.
  - Greedy: very-bad theoretically but good in practice.
    - I suspect a simple fix to this is possible.

Superlinear Convergence

- In a typical setting, we might hope that $|\mathcal{Z}^c| << n$.
  - Here we have the potential for faster algorithms by doing Newton steps on $\mathcal{Z}$.
- Some possibilities:
  - At some point, switch from proximal-gradient to Newton on the manifold.
    - Unfortunately, hard to decide when to switch.
  - Each iteration checks progress of proximal-gradient and Newton [Wright, 2012].
  - Two-metric projection [Gafni & Bertsekas, 1984].
    - May require expensive Newton steps before we're on the manifold.
  - Block coordinate descent with proximal-Newton or two-metric projection updates.
    - May be able to keep cost low but eventually get superlinear convergence.
  - There remains some theoretical and experimental work to do here.

Summary

- Proximal-gradient methods identify the sparsity pattern in finite iterations.
- We define "active-set complexity" as the number of iterations needed.
- We bound active-set complexity bound for proximal-gradient.
  - Smooth and strongly-convex f with a separable regularizer.
- We discussed other issues like coordinate descent methods and Newton hybrids.
- Thanks for the invite.