Accelerated Projected Gradient Algorithms for Sparsity Constrained Optimization Problems

Jan Harold Alcantara and Ching-pei Lee
Academia Sinica

NeurIPS 2022

Problem Setup

We consider

$$\min_{w \in A_s} \; f(w) \tag{P}$$

where

- f has L-Lipschitz continuous gradient
- $A_s := \{w \in \mathbb{R}^n : \|w\|_0 \leq s\}$
- f is lower-bounded on A_s
- Known as: feature selection, best subset selection, etc.

Our approach

- Revisit projected gradient (PG) algorithm

$$w^{k+1} \in T_{\mathrm{PG}}^{\lambda}(w^k) := P_{A_s}(w^k - \lambda \nabla f(w^k)) \tag{PG}$$

  with $\lambda \in (0, 1/L)$

- Two acceleration strategies
  1. Acceleration by same-subspace extrapolation
  2. Switching to smooth optimization (Newton) when the right subspace is identified

Strategy 1: Acceleration by same-subspace extrapolation

Decompose A_s as

$$A_s = \bigcup_{J \in \mathcal{J}_s} A_J, \quad A_J := \operatorname{span}\{e_j : j \in J\},$$
$$\mathcal{J}_s := \{J \subseteq \{1, 2, \ldots, n\} : |J| = s\},$$

- Iterates are confined on A_s
- If w^{k-1} and w^k belong to the same A_J for some J, we conduct extrapolation along
$$d^k := w^k - w^{k-1}.$$
Otherwise, skip extrapolation.
- Find suitable $t_k > 0$ and set

$$z^k := w^k + t_k d^k \qquad \text{(Extrapolation)}$$
$$w^{k+1} \in T^{\lambda}_{\mathrm{PG}}(z^k). \qquad \text{(ProjStep)}$$

Global convergence

Definition

The KL condition holds at w^* if there exists neighborhood $U \subset \mathbb{R}^n$ of w^*, $\theta \in [0,1]$, and $\kappa > 0$ such that for every $J \in \mathcal{I}_{w^*}$,

$$(f(w) - f(w^*))^\theta \leq \kappa \|(\nabla f(w))_J\|, \quad \forall w \in A_J \cap U. \tag{KL}$$

Theorem

(**Notation:** n_k denotes the number of successful extrapolation steps in the first k iterations.)

Suppose that there is an accumulation point w^* of the iterates at which (KL) holds. Then $w^k \to w^*$. Moreover, the following rates hold:

(a) If $\theta \in (1/2, 1)$: $f(w^k) - f(w^*) = O((k + n_k)^{-1/(2\theta-1)})$.

(b) If $\theta \in (0, 1/2]$: $f(w^k) - f(w^*) = O(\exp(-(k + n_k)))$.

(c) If $\theta = 0$, or $\theta \in [0, 1/2]$ and f is convex: there is $k_0 \geq 0$ such that $f(w^k) = f(w^*)$ for all $k \geq k_0$.

Strategy 2: Subspace identification and Smooth Optimization

Theorem (Subspace identification)

There exists $N \in \mathbb{N}$ such that

$$\{w^k\}_{k=N}^{\infty} \subseteq \bigcup_{J \in \mathcal{I}_{w^*}} A_J, \qquad \mathcal{I}_{w^*} := \{J \in \mathcal{J}_s : w^* \in A_J\}. \tag{$\star$}$$

whenever $w^k \to w^*$. In particular,

(a) if $T_{\mathrm{PG}}^{\lambda}(w^*)$ is a singleton for an accumulation point w^* of $\{w^k\}$, then w^* is a local minimum, $w^k \to w^*$, and ($\star$) holds.

(b) ($\star$) holds for Algorithm 1 under the hypotheses of the previous theorem.

- We switch to the truncated Newton method after the subspace A_J becomes fixed for multiple iterations
- This strategy provably leads to superlinear or even quadratic convergence

Experiments

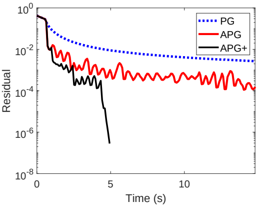
- Loss functions: Least squares and logistic loss
- Datasets: Used public datasets with #instances $\ll$ #features
- Stopping criterion

$$\text{Residual}(w) := \frac{\|w - P_{A_s}(w - \lambda \nabla f(w))\|}{(1 + \|w\| + \lambda \|\nabla f(w)\|)} < \hat{\epsilon} \tag{1}$$

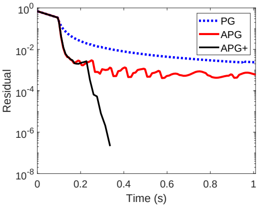
  with $\hat{\epsilon} = 10^{-6}$.
- Compare:
  - PG
  - APG: Our same-subspace extrapolation acceleration
  - APG+: APG plus the smooth Newton part

m = #instances, n = #features, s =#allowed_nonzeros

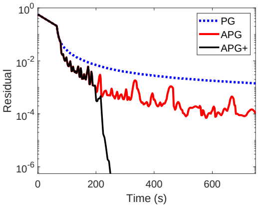
Logistic regression



(a) news20, $m = 15,997$
$n = 1,355,191$, $s = 0.01m$

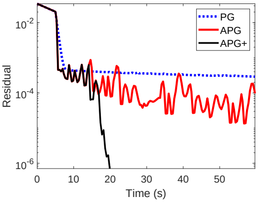


(b) rcv1.binary, $m = 20,242$
$n = 47,236$, $s = 0.01m$

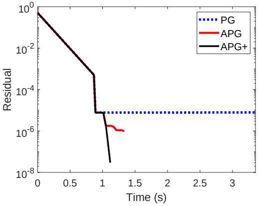


(c) webspam, $m = 280,000$,
$n = 16,609,143$, $s = 0.001m$

Least square



(d) E2006-log1p, $m = 16,087$
$n = 4,272,227$, $s = 0.01m$



(e) E2006-tfidf, $m = 16,087$,
$n = 150,360$, $s = 0.01m$

Thank you for listening!