Accelerated Stochastic Greedy Coordinate Descent

Chaobing Song, Shaobo Cui, Yong Jiang and Shu-Tao Xia

Tsinghua University

Objectives

In this paper, we propose two algorithms called SOTOPO and ASGCD to solve the well-known ℓ_1-regularized problem, which has the following properties.

- The first attempt to accelerate greedy coordinate descent (GCD) by Nesterov's acceleration mechanism
- Reducing the greedy selection complexity of GCD by a factor of sample size
- The iterative solution is sparse from the beginning of the iteration
- The optimal convergence rate $O(1/\sqrt{\epsilon})$
- Better performance for dense problems with sparse solutions

Introduction

Given the dataset $\{(a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n)\}$ with $a_i \in \mathbb{R}^d, b_i \in \mathbb{R}$. We solve the ℓ_1-regularized problem

$$\frac{1}{n}\sum_{i=1}^{n} f_i(a_i^T x, b_i) + \lambda\|x\|_1. \tag{1}$$

- **One iteration level**: The Motivation of SOTOPO
  Assume $x_0 = 0$.
  - Proximal Gradient Descent

    $$x_k = \arg\min_{x\in\mathbb{R}^d}\{\langle\nabla f(x_{k-1}), x - x_{k-1}\rangle + \frac{1}{2\eta}\|x - x_{k-1}\|_2^2 + \lambda\|x\|_1\} \tag{2}$$

    Pros: Combined with Nesterov's acceleration, the optimal convergence rate can be obtained
    Cons: x_k will be not sparse until performing enough iterations
  - Greedy coordinate descent such as GS-q rule

    $$i_{\max} = \arg\min_i \left\{\min_d\{f(x_k) + \nabla_i f(x_k)d + \frac{1}{2\eta}d^2 + g_i(x_{k,i} + d) - g_i(x_{k,i})\}\right\}$$
    $$x_{k+1} = x_k - \eta d\mathbf{e}_{i_{\max}}$$

    Pros: x_k will be sparse at the beginning of the iterations
    Cons: Nonconvex, therefore cannot use Nesterov's acceleration to obtain $O(1/\sqrt{\epsilon})$.
  - SOTOPO
    Pros:
    - x_k will be sparse in all the iterations
    - Convex and therefore use Nesterov's acceleration to obtain $O(1/\sqrt{\epsilon})$

    Cons: Nontrivial to solve
- **Optimization framework level**: The Motivation of ASGCD
  Variance reduction and Nesterov's acceleration have been applied to
  - gradient descent
  - gandomized coordinate descent

  but not been applied to
  - greedy coordinate descent

  yet, which is important to sparse optimization.

The problem to solve: ℓ_1-regularized ℓ_1-norm square approximation

Set $x_0 = 0$. In each iteration, we need to solve

$$x_k = \arg\min_{x\in\mathbb{R}^d}\left\{\langle\nabla f(x_{k-1}), x - x_{k-1}\rangle + \frac{1}{2\eta}\|x - x_{k-1}\|_1^2 + \lambda\|x\|_1\right\}$$

Properties:

- By $\|x - x_{k-1}\|_1^2$, x_k will be updated on only one or several coordinates;
- By $\lambda\|x\|_1$, x_k will be sparse even after enough iterations
- If $\lambda = 0$, it is equivalent to greedy coordinate descent
- It is a non-separable and double non-smooth problem
- However it is convex!

Solving the problem: SOft ThreshOlding PrOjection (SOTOPO)

1. Unconstrained, non-separable and non-smooth $\Longrightarrow$ Simplex constrained, separable and smooth

   A key variational equality: $\|x\|_1^2 = \inf_{\theta:\sum_{i=1}^d \theta_i=1, \theta_i\geq 0} \sum_{i=1}^{d} \frac{x_i^2}{\theta_i}$
2. Optimize θ first, x later $\Longrightarrow$ Optimize x first, θ later.

   A key observation: A soft thresholding form can be obtained for x with respect to θ
3. Unconstrained nonsmooth optimization of $x \Longrightarrow$ Projecting θ to a simplex to optimize a smooth loss function $J(\theta)$
4. Using the property of $J(\theta)$ to perform complexity reduction (A novel contribution)

Main Result for SOTOPO

The SOTOPO algorithm can get the exact minimizer in each iteration of the ℓ_1-regularized ℓ_1-norm square approximation problem.

Main Result for ASGCD

ASGCD achieves an ϵ-additive error (*i.e.*, $\mathbb{E}[F(\tilde{x}^S)] - F(x^*) \leq \epsilon$) using at most $O\left(\frac{\sqrt{L_1}\|x^*\|_1 \log d}{\sqrt{\epsilon}}\right)$ iterations.

The problem to solve: the ℓ_1-regularized problem

- Assumption: for each $f_i(a_i^T x, b_i)$, instead of assume

  $$\|\nabla f_i(a_i^T x, b_i) - \nabla f_i(a_i^T y, b_i)\|_2 \leq L_2\|x - y\|_1, \tag{3}$$

  we assume

  $$\|\nabla f_i(a_i^T x, b_i) - \nabla f_i(a_i^T y, b_i)\|_\infty \leq L_1\|x - y\|_1 \tag{4}$$
- Observation: for dense samples, it is possible that $L_1 \ll L_2$.
- Goal:
  - The iterative solutions are sparse for all the iterations
  - Obtaining the optimal convergence rate $O(1/\sqrt{\epsilon})$
  - Reducing the iterative complexity by stochastic optimization

Solving the problem: accelerated stochastic greedy coordinate descent (ASGCD)

- Method 1: linear coupling of
  - SOTOPO
  - ℓ_q-norm based Composite Mirror Descent [1]

  to obtain the $O(1/\sqrt{\epsilon})$ rate and sparse iterative solution
- Method 2: Katyusha framework to reduce the complexity of greedy selection

Table 1: Convergence rate on ℓ_1-regularized empirical risk minimization problems. (For GCD, the convergence rate is applied for $\lambda = 0$.)

Algorithm Type	Paper	Rate
Non-Acc, Primal, ℓ_2-norm	SAGA [2]	$O\left(\frac{L_2\|x^*\|_2^2}{\epsilon}\right)$
ACC, Primal, ℓ_2-norm	Katyusha [3]	$O\left(\frac{\sqrt{L_2}\|x^*\|_2}{\sqrt{\epsilon}}\right)$
Acc, Dual, ℓ_2-norm	Acc-SDCA [4]	$O\left(\frac{\sqrt{L_2}\|x^*\|_2}{\sqrt{\epsilon}}\log(\frac{1}{\epsilon})\right)$
Non-Acc, Primal, ℓ_1-norm	GCD [5]	$O\left(\frac{L_1\|x^*\|_1^2}{\epsilon}\right)$
Acc, Primal, ℓ_1-norm	ASGCD	$O\left(\frac{\sqrt{L_1}\|x^*\|_1 \log d}{\sqrt{\epsilon}}\right)$

Experiment

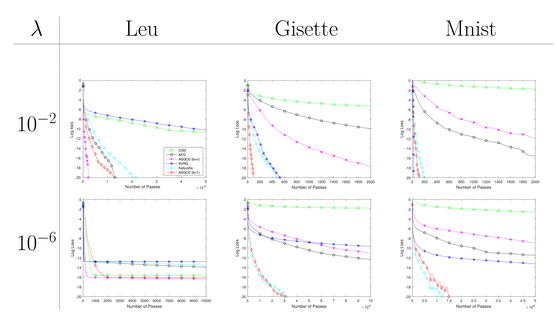

Figure 1: Comparing AGCD ($b = 1$) and ASGCD ($b = n$) with CGD, SVRG, AFG and Katyusha on Lasso.

References

[1] John C Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent.
In *COLT*, pages 14–26, 2010.

[2] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives.
In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[3] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods.
ArXiv e-prints, abs/1603.05953, 2016.

[4] Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization.
In *ICML*, pages 64–72, 2014.

[5] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent.
ArXiv e-prints, abs/1407.1537, July 2014.

Acknowledgements

The authors thank for Yuchen Zhang from Stanford and the anonymous reviewers.

Contact Information

- Email: songcb16@mails.tsinghua.edu.cn
- Wechat: word_sword