An Accelerated Doubly Stochastic Gradient Method with Faster Explicit Model Identification

Runxue Bao[1], Bin Gu[2], Heng Huang[1]

[1]Electrical and Computer Engineering, University of Pittsburgh, PA 15213, United States
[2]Mohamed bin Zayed University of Artificial Intelligence, Masdar City, Abu Dhabi, UAE

CIKM 2022

Background

We consider the composite optimization problem involving a data fitting function $\mathcal{F}(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(a_i^\top x)$ plus a block-separable sparsity-inducing regularizer $\Omega(x) = \sum_{j=1}^{q} \Omega_j(x_{\mathcal{G}_j})$ as:

$$\min_{x \in \Re^p} \mathcal{P}(x) := \mathcal{F}(x) + \lambda \Omega(x). \tag{1}$$

where $A = [a_1, \cdots, a_n]^\top \in \Re^{n \times p}$ is the design matrix, $\mathcal{G}$ is the partition, λ is the regularization parameter and $x \in \Re^p$ is the model coefficients.

Definitions and Assumptions

Equicorrelation Set(see [4]): Suppose θ^* is the dual optimal, the equicorrelation set is defined as

$$\mathcal{S}^* := \{j \in \{1, 2, \ldots, q\} : \frac{1}{n}\Omega_j^D(A_j^\top \theta^*) = \lambda\}. \tag{2}$$

Assumption 1: Given the partition $\{\mathcal{G}_1, \ldots, \mathcal{G}_q\}$, all $\nabla_{\mathcal{G}_j} f_i(x) = [\nabla f_i(x)]_{\mathcal{G}_j}$ are block-wise Lipschitz continuous with constant L_i, which means that for any x and x', there exists a constant $L = \max_i L_i$, we have

$$\|\nabla_{\mathcal{G}_j} f_i(x) - \nabla_{\mathcal{G}_j} f_i(x')\| \leq L\|x_{\mathcal{G}_j} - x'_{\mathcal{G}_j}\|. \tag{3}$$

Assumption 2: $\mathcal{F}(x)$ and $\Omega(x)$ are proper, convex and lower-semicontinuous.

Motivation and Challenges

Motivation:

- Doubly stochastic gradient Method [7, 5] suffers huge computational costs in the practical high-dimensional setting,
- Proximal gradient method with screening can simultaneously achieve enjoys the implicit identification and explicit identification.

Challenges:

- Existing safe screening algorithms are limited to the deterministic setting.
- Existing works [1, 2, 3] fail to show how fast we can achieve explicit model identification.

Proposed Method

Algorithm 1 The ADSGD method

Input: $\hat{x}_0$.

for $k = 1, 2, \ldots$ **do**
  $\tilde{x}_{k-1} = \hat{x}_{k-1}, \tilde{\mu}_{k-1} = \nabla \mathcal{F}(\tilde{x}_{k-1}), x_{k-1}^0 = \tilde{x}_{k-1}$.
  Compute θ_{k-1} by (4).
  $r^{k-1} = \sqrt{2T \operatorname{Gap}(\tilde{x}_{k-1}, \theta_{k-1})}$.
  Update $\mathcal{S}_k \subset \mathcal{S}_{k-1}$ by (5).
  Update $A_{\mathcal{S}_k}, x_k^0, \tilde{x}_k, \tilde{\mu}_k$ with $\mathcal{S}_k$.
  for $t = 1, 2, \ldots, mq_k/q$ **do**
    Randomly pick $\mathcal{I} \subset \{1, 2, \ldots, n\}$ and j from $\mathcal{S}_k$.
    $\mu_k = \nabla_{\mathcal{G}_j} \mathcal{F}_{\mathcal{I}}(x_k^{t-1}) - \nabla_{\mathcal{G}_j} \mathcal{F}_{\mathcal{I}}(\tilde{x}_k) + \tilde{\mu}_{\mathcal{G}_j, k}$.
    $x_{k,\mathcal{G}_j}^t = \operatorname{prox}_{\eta,\lambda}^j(x_{\mathcal{G}_j}^{t-1} - \eta \mu_k)$.
  end for
  $\hat{x}_k = \frac{1}{m_k} \sum_{t=1}^{m_k} x_k^t$
end for

Output: Coefficient $\hat{x}_k$.

Proposed Method

Eliminating Step: We compute θ_{k-1} with the active set $\mathcal{S}_{k-1}$ from the previous iteration as

$$\theta_{k-1} = \frac{-\nabla \mathcal{F}(\tilde{x}_{k-1})}{\max(1, \Omega^D(A_{\mathcal{S}_{k-1}}^\top \nabla \mathcal{F}(\tilde{x}_{k-1}))/\lambda)}. \tag{4}$$

We obtain new active set $\mathcal{S}_k$ from $\mathcal{S}_{k-1}$ by the screening conducted on all $j \in \mathcal{S}_{k-1}$ as

$$\frac{1}{n}\Omega_j^D(A_j^\top \theta_{k-1}) + \frac{1}{n}\Omega_j^D(A_j) r^{k-1} < \lambda \Rightarrow \tilde{x}^*_{\mathcal{G}_j} = 0. \tag{5}$$

Proposed Method

Doubly Stochastic Gradient Update: ADSGD only computes the partial derivative $\nabla_{\mathcal{G}_j}\mathcal{F}_{\mathcal{I}}(x_k^{t-1})$ on one coordinate block with respect to a sample each time. The proximal step is computed as:

$$\text{prox}^j_{\eta,\lambda}(x'_{\mathcal{G}_j}) = \arg\min_{x_{\mathcal{G}_j}} \frac{1}{2\eta}\|x'_{\mathcal{G}_j} - x_{\mathcal{G}_j}\|^2 + \lambda\Omega_j(x_{\mathcal{G}_j}). \tag{6}$$

Variance Reduction on the Selected Blocks: We adjust the partial gradient estimation over the selected block $\mathcal{G}_j$ to reduce the gradient variance as:

$$\mu_k = \nabla_{\mathcal{G}_j}\mathcal{F}_{\mathcal{I}}(x_k^{t-1}) - \nabla_{\mathcal{G}_j}\mathcal{F}_{\mathcal{I}}(\tilde{x}_k) + \tilde{\mu}_{\mathcal{G}_j,k}. \tag{7}$$

Theoretical Analysis

Linear Convergence: Suppose $\hat{x}_k$ be generated from the k-th iteration of the main loop in Algorithm 1 and let $|\mathcal{I}| \geq T/L$ and $\eta < \frac{1}{4L}$, we have

$$\mathbb{E}\mathcal{P}_k(\hat{x}_k) - \mathcal{P}(x^*) \leq \rho^k[\mathcal{P}(\hat{x}) - \mathcal{P}(x^*)]. \tag{8}$$

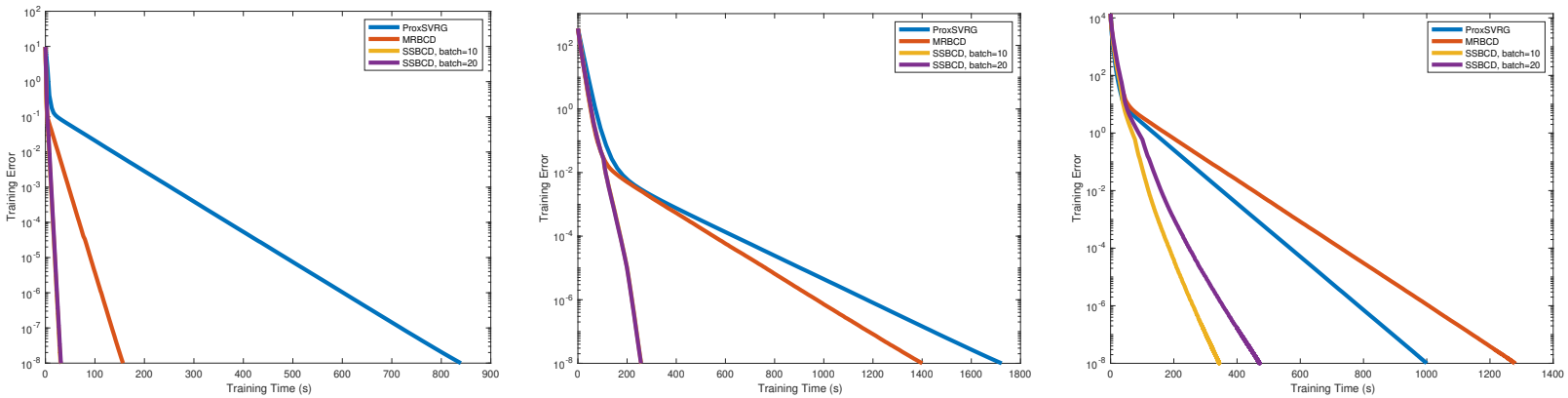
We can choose $|\mathcal{I}| = T/L$, $\eta = \frac{1}{16L}$, and $m = 65qL/\mu$ to make $\rho < 2/3$.

Explicit Model Identification: Define $\Delta_j \triangleq \frac{n\lambda - \Omega_j^D(A_j^\top \theta^*)}{2\Omega_j^D(A_j)}$, denote σ_A^2 as the spectral norm of A, suppose Ω has a bounded support within a ball of radius M, given any $\gamma \in (0,1)$, any block that $j \notin \mathcal{S}^*$ are correctly identified by ADSGD at iteration $\log_{\frac{1}{\rho}}(1/\epsilon_j)$ with at least probability $1-\gamma$ where $\epsilon_j = \frac{1}{32}\frac{\Delta_j^4 \gamma}{T^3\sigma_A^2 M^2(\mathcal{P}(\hat{x}) - \mathcal{P}(x^*))}$.

Theoretical Analysis

Overall Complexity: Suppose the size of the active features in set $\mathcal{S}_k$ is d_k and d^* is the size of the active features in $\mathcal{S}^*$, given any $\gamma \in (0,1)$, let $K_m = O(\log_{\frac{1}{\rho}}(1/\epsilon_j))$, we have d_k is decreasing and d_{K_m} equals to d^* with at least probability $1-\gamma$. Define $s = \frac{1}{K_c}\sum_{k=1}^{K_c} d_k$ where $K_c = O(\log_{\frac{1}{\rho}}(1/\epsilon))$, the overall complexity of ADSGD is $O((n + T/\mu)s\log(1/\epsilon))$.

Convergence Results



(a) PlantGo (b) Protein (c) Real-sim

Figure: Convergence results of different algorithms for Lasso on different datasets.

We compare the convergence results of ADSGD w.r.t the running time with competitive algorithms ProxSVRG [6] and MRBCD [7, 5].

Conclusion

- We propose a novel accelerated doubly stochastic gradient descent method for generalized sparsity regularized problems with lower overall complexity and faster explicit model identification rate.
- We derive rigorous theoretical analysis for both strongly and nonstrongly convex functions. For strongly convex function, ADSGD can achieve a linear convergence rate and reduce the per-iteration cost with a lower overall complexity $O(s(n + T/\mu)\log(1/\epsilon))$.
- We rigorously prove our ADSGD algorithm can achieve the explicit model identification at a linear rate $O(\log(1/\epsilon_j))$.
- We empirically show that ADSGD can achieve a significant computational gain than existing methods.

Thank You!


O. Fercoq, A. Gramfort, and J. Salmon.
Mind the duality gap: safer rules for the lasso.
In *International Conference on Machine Learning*, pages 333–342, 2015.


E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon.
Gap safe screening rules for sparsity enforcing penalties.
JMLR, 2017.


A. Rakotomamonjy, G. Gasso, and J. Salmon.
Screening rules for lasso with non-convex sparse regularizers.
In *International Conference on Machine Learning*, pages 5341–5350, 2019.


R. J. Tibshirani et al.
The lasso problem and uniqueness.
Electronic Journal of statistics, 7:1456–1490, 2013.


H. Wang and A. Banerjee.
Randomized block coordinate descent for online and stochastic optimization.
arXiv preprint arXiv:1407.0107, 2014.


L. Xiao and T. Zhang.
A proximal stochastic gradient method with progressive variance reduction.
SIAM Journal on Optimization, 24(4):2057–2075, 2014.


T. Zhao, M. Yu, Y. Wang, R. Arora, and H. Liu.
Accelerated mini-batch randomized block coordinate descent method.
NeurIPS, 2014.