A Variance-Reduced and Stabilized Proximal Stochastic Gradient Method with Support Identification Guarantees for Structured Optimization

Yutong Dai[1] Guanyi Wang [2] Frank E. Curtis[1] Daniel P. Robinson[1]

Lehigh University[1] National University of Singapore[2]

paper github

Introduction

Motivation

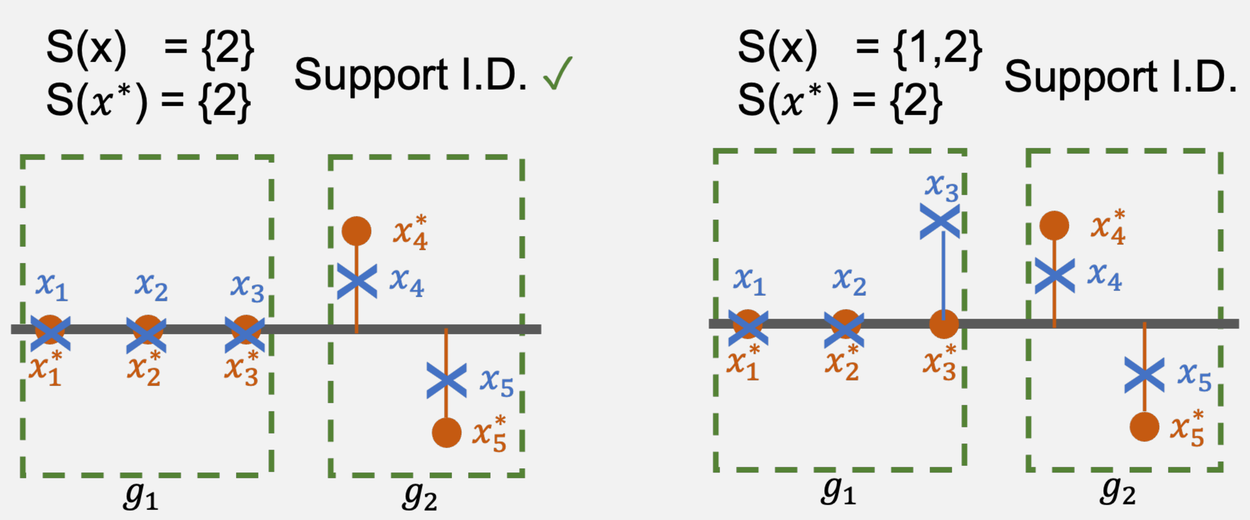
- Regularized learning problems are ubiquitous in machine learning and sparse solutions are often preferred and obtained via nonsmooth regularizers.
- Full gradient evaluation in large-scale problems or online-learning problems are prohibitive, hence the mainstream uses stochastic gradient-type methods with variance reduction. Yet, most variance reduction techniques require at least one full gradient evaluation or the storage of a stochastic gradient table.

Problem Setting

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + r(x).$$

- $f(x) := \mathbb{E}_{\xi\sim\mathcal{P}}[\ell(x;\xi)]$ population loss with $\xi \sim \mathcal{P}$
- $\ell(\cdot,\xi)$ is a smooth convex function almost surely w.r.t the distribution of ξ
- r is a sparsity-promoting convex function with a group separable structure

Support Identification and Consistent Support Identification



- Group Structure: $\bigcup_{i=1}^{n_\mathcal{G}} g_i = n$ and $g_i \cap g_j = \emptyset$ for all $i \in [n_\mathcal{G}]$.
- Support of x: $\mathcal{S}(x) = \{i \in [n_\mathcal{G}] \mid [x]_{g_i} \neq 0\}$.
- Support Identification Property: For any sufficiently large k, $\mathcal{S}(x_k) = \mathcal{S}(x^*)$ holds with high probability (w.h.p.), i.e., $\mathbb{P}\{\mathcal{S}(x_k) = \mathcal{S}(x^*)\} \geq p$.
- Consistent Support Identification Property: For all sufficiently large k, $\mathcal{S}(x_k) = \mathcal{S}(x^*)$ holds w.h.p., i.e., $\mathbb{P}\{\cap_{k\geq K}\{\mathcal{S}(x_k) = \mathcal{S}(x^*)\}\} \geq p$.

Contributions

- Propose variance reduction method `S-PStorm` with neither any exact gradient evaluation nor storage of a stochastic gradient table.
- Establish the consistent support identification property of `S-PStorm`, which is stronger than the support identification property of `RDA`.
- Show better performances of `S-PStorm` over `RDA` on a class of test problems.

Algorithm	$\|x_k - x^*\|^2$	Support Identification	# Exact ∇f	Storage
`ProxSVRG`	$\mathcal{O}\left(\rho^k_{\texttt{ProxSVRG}}\right)$	$\mathcal{O}(\log(1/\delta^*))$	every epoch	$\mathcal{O}(n)$
`SAGA`	$\mathcal{O}\left(\rho^k_{\texttt{SAGA}}\right)$	$\mathcal{O}(\log(1/\delta^*))$	once	$\mathcal{O}(Nn)$
`RDA`	$\mathcal{O}(\log k/k)$	$\mathcal{O}\left(\frac{1}{(\delta^*)^4}\right)$	never	$\mathcal{O}(n)$
`S-PStorm`	$\mathcal{O}(\log k/k)$	$\mathcal{O}\left(\max\left\{\frac{1}{(\delta^*)^4}, \frac{1}{(\Delta^*)^4}\right\}\right)$	never	$\mathcal{O}(n)$

Algorithm

Algorithm 1 `S-PStorm`

Inputs: Initial point $x_0 = x_1 \in \mathbb{R}^n$, size of mini-batch $m \in \mathbb{N}_+$, weight sequence $\{\beta_k\}_{k\geq 2} \subset (0,1)$, stepsize sequence $\{\alpha_k\} \subset (0,\infty)$, and parameter $\zeta \in (0,\infty)$.
for $k = 1, 2, \ldots,$ **do**
  Draw m i.i.d samples $\{\xi_{k1}, \cdots, \xi_{km}\}$ w.r.t. $\mathcal{P}$.
  Set $v_k \leftarrow \frac{1}{m}\sum_{i=1}^m \nabla\ell(x_k;\xi_{ki})$.
  if $k = 1$ **then**
    Set $d_k \leftarrow v_k$.
  else
    Set $u_k \leftarrow \frac{1}{m}\sum_{i=1}^m \nabla\ell(x_{k-1};\xi_{ki})$.
    Set $d_k \leftarrow v_k + (1-\beta_k)(d_{k-1} - u_k)$. `Storm` estimator
  end if
  Compute $y_k \leftarrow \mathrm{prox}_{\alpha_k r}(x_k - \alpha_k d_k)$. support inexact **prox** operator evaluation
  Set $x_{k+1} \leftarrow x_k + \zeta\beta_k(y_k - x_k)$. stabilization step
end for

Assumptions

Filtration: A random process $\mathcal{F}_k$ up to time k over the stochastic gradient sampling procedure.

1. Unbiased Stochastic Gradient: $\mathbb{E}_{\xi\sim\mathcal{P}}[\nabla\ell(x_k;\xi) \mid \mathcal{F}_k] = \nabla f(x_k)$.
2. Bounded subdifferential: There exists $G_r > 0$ such that, $\mathbb{P}\{\|g_r\|_2 \leq G_r,\ \forall g_r \in \partial r(x_k)\} = 1$.
3. Bounded errors: There exists $\sigma > 0$ such that $\mathbb{P}_{\xi\sim\mathcal{P}}\{\|\nabla\ell(x_k,\xi) - \nabla f(x_k)\| \leq \sigma \mid \mathcal{F}_k\} = 1$.
4. Bounded steps: There exists $G_d > 0$ such that $\mathbb{P}_{\xi\sim\mathcal{P}}\{\|d_k\| \leq G_d \mid \mathcal{F}_k\} = 1$.
5. Convexity: f is μ_f-strongly convex and r_i is convex and closed for all $i \in [n_\mathcal{G}]$.
6. Smooth loss: $\nabla\ell(\cdot,\cdot)$ is Lipschitz continuous with respect to the first argument.
7. Algorithmic choices: $\beta_k = \min\{1/2, c/(k+1)\}$ and $\alpha_k \equiv \underline{\alpha}$ with $c > 1$ and $\underline{\alpha} \in (0,\infty)$.

Variance Reduction

- Define the error in the gradient estimator as $\epsilon_k = d_k - \nabla f(x_k)$. With $c > 0$, $\eta_k > 0$,

$$U(k) = C\Big(\sigma + L_g(G_r + G_d)\zeta\underline{\alpha}\Big) \cdot \max\left\{\left(\frac{\underline{k}+1}{k+2}\right)^c, \frac{c}{\sqrt{k+2}}\right\}\sqrt{\log\frac{2}{\eta_k}},$$

  then $\mathbb{P}\left[\|\epsilon_k\| \leq U(k)\right] \geq 1 - \eta_k$ for all $k \geq \underline{k} = \lceil(2c) - 1\rceil$.
- If $\eta_k = \eta_0/k^2$ for all $k \geq 1$, then error ϵ_k vanishes at the rate of $\mathcal{O}(\sqrt{\log k/k})$ w.h.p..

Convergence of the Iterates

- Let $\kappa = L_g/\mu_f$, $\underline{\alpha} = 1/(\kappa L_g)$, $\zeta \in (0,2)$, $c = 2\kappa^2/\zeta$, $\underline{k} = \lceil 2c - 1\rceil$, and $\eta_k = \eta_0/k^2$ for all $k \geq 1$ with $\eta_0 \in (0, 6/\pi^2)$. Let $(\bar{c}_1, \bar{c}_2)$ be some positive constants independent of k.
- Let $\mathcal{E}^x_k := \left\{\|x_k - x^*\|^2 \leq \bar{c}_1\frac{\|x_{\underline{k}} - x^*\|^2}{k^\theta} + \bar{c}_2 \cdot \frac{\log\frac{2k}{\eta_0}}{k}\right\}$, then $\mathbb{P}\left[\bigcap_{k\geq\underline{k}}^\infty \mathcal{E}^x_k\right] \geq 1 - \eta_0\pi^2/6 > 0$.
- $\|x_k - x^*\|$ vanishes at the rate of $\mathcal{O}(\sqrt{\log k/k})$ w.h.p. .

Consistent Support Identification

Assume x^* is neither fully dense nor all zero, then

$$\Delta^* = \min_{i\in\mathcal{S}(x^*)}\left\|[x^*]_{g_i}\right\|, \delta^* = \min_{i\notin\mathcal{S}(x^*)}\{\lambda_i - \left\|\nabla_{g_i} f(x^*)\right\|\}.$$

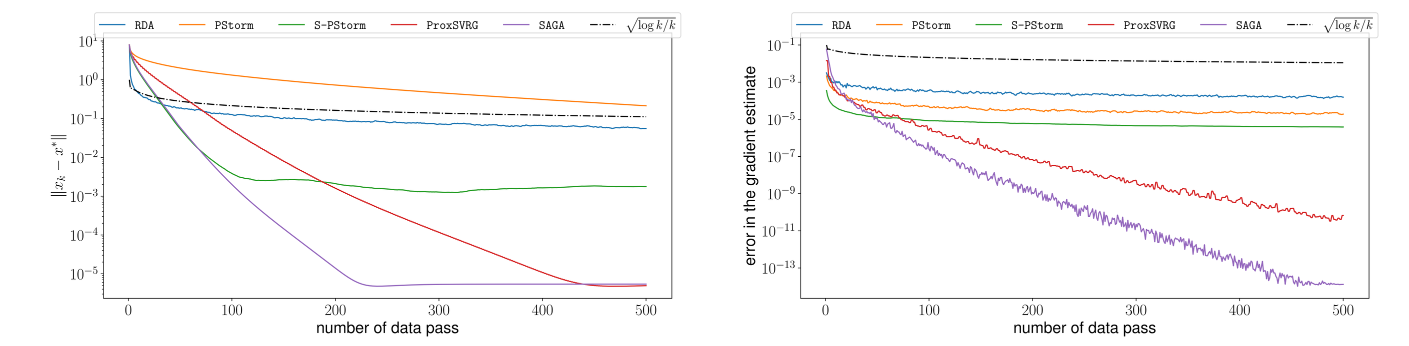
Define $k_{\delta^*} = (C_{41}/\delta^*)^4$, $k_{\Delta^*} = (C_{42}/\Delta^*)^4$, and $K_{\text{id}} = \max\{k_{\delta^*}, k_{\Delta^*}, \underline{k}\}$ with positive constants $\{C_{41}, C_{42}\}$ that are independent of k. Then $\mathbb{P}\left[\bigcap_{k\geq K_{\text{id}}}\{\mathcal{S}(y_k) = \mathcal{S}(x^*)\}\right] \geq 1 - \frac{\eta_0\pi^2}{6} > 0.$

Experiments

Test Problems: 80 test instances derived from 10 datasets from the LIBSVM collection with various solution sparsity levels and group structures.

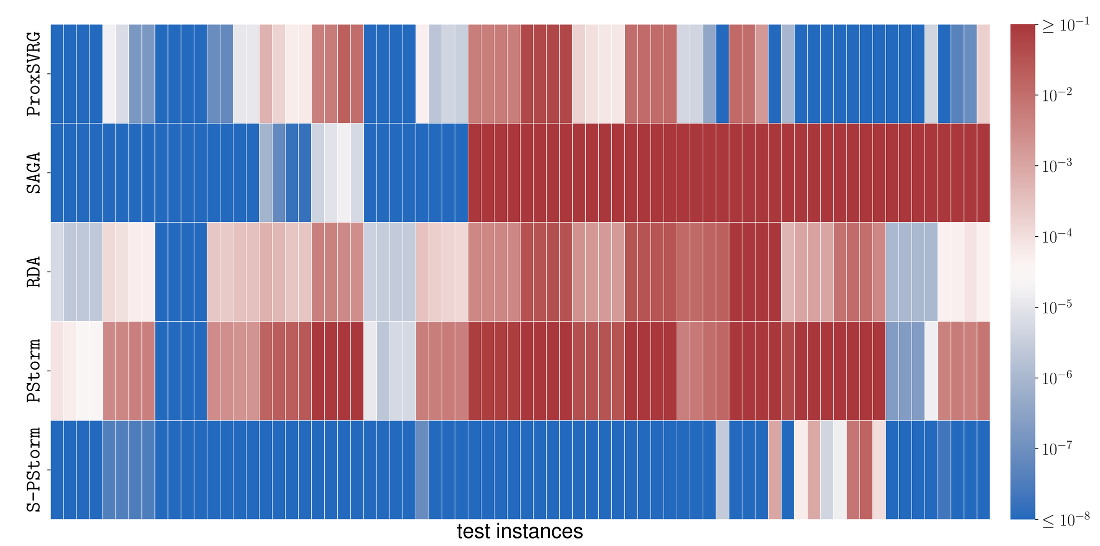
$$\min_{x\in\mathbb{R}^n} \frac{1}{N}\sum_{j=1}^N \log\left(1 + e^{-y_j x^T d_j}\right) + 10^{-5}\|x\|^2 + \sum_{i=1}^{n_\mathcal{G}} \lambda_i \left\|[x]_{g_i}\right\|$$

Convergence of Iterates and Variance Reduction.



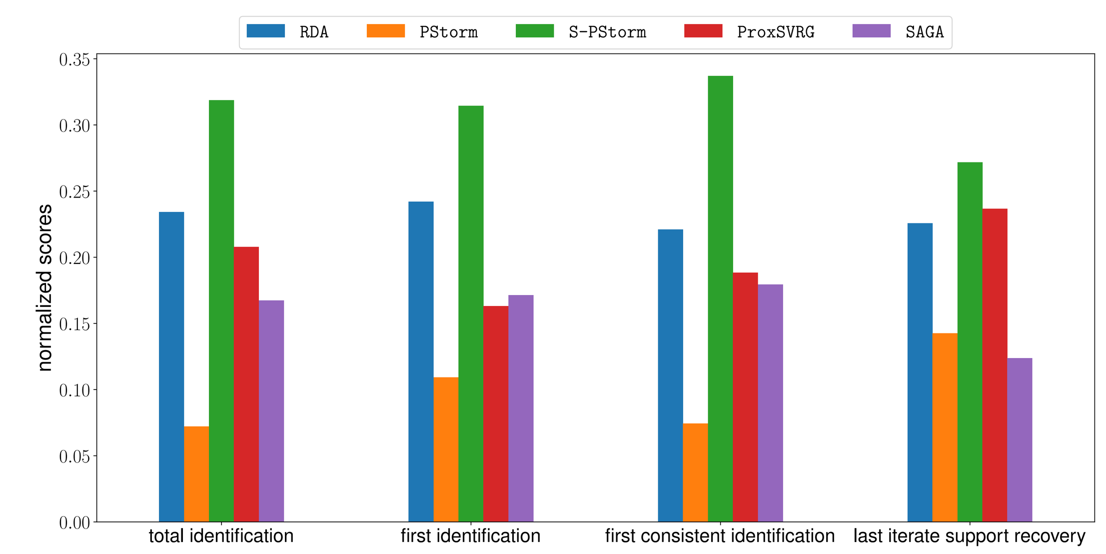
Typical behavior of how fast the $\|x_k - x^*\|$ (left) and gradient error $\|\epsilon_k\|$ (right) converge to 0 for competing methods. `S-PStorm` performs better than `RDA`.

Solution Quality: Measure optimal function value gap.



Each block along x-axis is a test instance.
deep blue better
deep red worse

Support Identification: Rank scores of four metrics measuring performance: 1) total epochs that support identification occurs, 2) the first epoch that y_k identifies the correct support, 3) the first epoch after which y_k consistently identifies the correct support, and 4) the percentage of support $S(x^*)$ recovered by the final y_k.



higher bar better

References

[1] A. Cutkosky and F. Orabona, "Momentum-based variance reduction in non-convex sgd," *Advances in neural information processing systems*, vol. 32, 2019.

[2] Y. Xu and Y. Xu, "Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization," *Journal of Optimization Theory and Applications*, pp. 1–32, 2022.

[3] L. Xiao, "Dual averaging method for regularized stochastic learning and online optimization," *Advances in Neural Information Processing Systems*, vol. 22, 2009.

[4] S. Lee and S. J. Wright, "Manifold identification in dual averaging for regularized stochastic online learning.," *Journal of Machine Learning Research*, vol. 13, no. 6, 2012.

[5] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives," *Advances in neural information processing systems*, vol. 27, 2014.

[6] L. Xiao and T. Zhang, "A proximal stochastic gradient method with progressive variance reduction," *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 2057–2075, 2014.