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

Yutong Dai
Lehigh University

Guanyi Wang
National University of Singapore

Frank E. Curtis
Lehigh University

Daniel P. Robinson
Lehigh University

Abstract

This paper introduces a new proximal stochastic gradient method with variance reduction and stabilization for minimizing the sum of a convex stochastic function and a group sparsity-inducing regularization function. Since the method may be viewed as a stabilized version of the recently proposed algorithm `PStorm`, we call our algorithm `S-PStorm`. Our analysis shows that `S-PStorm` has strong convergence results. In particular, we prove an upper bound on the number of iterations required by `S-PStorm` before its iterates correctly identify (with high probability) an optimal support (i.e., the zero and nonzero structure of an optimal solution). Most algorithms in the literature with such a support identification property use variance reduction techniques that require either periodically evaluating an *exact* gradient or storing a history of stochastic gradients. Unlike these methods, `S-PStorm` achieves variance reduction without requiring either of these, which is advantageous. Moreover, our support-identification result for `S-PStorm` shows that, with high probability, an optimal support will be identified correctly in *all* iterations with index above a threshold. We believe that this type of result is new to the literature since the few existing other results prove that the optimal support is identified with high probability at each iteration with a sufficiently large index (meaning that the optimal support might be identified in some iterations, but not in others). Numerical experiments on regularized logistic loss problems show that `S-PStorm` outperforms existing methods in various metrics that measure how efficiently and robustly iterates of an algorithm identify an optimal support.

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

1 INTRODUCTION

We consider the regularized stochastic learning problem

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + r(x), \tag{1}$$

where $f(x) := \mathbb{E}_{\xi\sim\mathcal{P}}[\ell(x;\xi)]$ with ξ being a random vector following a probability distribution $\mathcal{P}$, $\ell(\cdot,\xi)$ is a smooth convex function almost surely with respect to the distribution of ξ, and r is a sparsity-promoting closed convex function with group separable structure, i.e., $r(x) := \sum_{i=1}^{n_\mathcal{G}} r_i([x]_{g_i})$ for some number of groups $n_\mathcal{G} > 0$ with $g_i \subseteq \{1, 2, \ldots, n\}$ for each $i \in \{1, 2, \ldots, n_\mathcal{G}\}$, $\bigcup_{i=1}^{n_\mathcal{G}} g_i = n$, and $g_i \cap g_j = \emptyset$ for all $i \neq j$. Some commonly used regularization functions have these properties, such as the weighted ℓ_1 norm $\sum_{i=1}^{n} \lambda_i|[x]_i|$ and the weighted non-overlapping Group-ℓ_1 norm $\sum_{i=1}^{n_\mathcal{G}} \lambda_i \|[x]_{g_i}\|$, where $\{\lambda_i\}$ are positive scalars, $[x]_i$ denotes the ith component of x, $[x]_{g_i}$ denotes the subvector of x with entries from g_i, and $\|\cdot\|$ is the ℓ_2 norm. Problem (1) is general enough to cover a broad class of problems of interest. In particular, when data samples (ξ) are available in a streaming manner, problem (1) recovers online convex learning (Hazan et al., 2016), and when $\mathcal{P}$ is a uniform distribution over a finite set $\{1, 2, \cdots, N\}$, problem (1) recovers many regularized finite-sum problems (Tibshirani, 1996; Hastie et al., 2009).

In this work, we are interested in designing an algorithm for solving problem (1) that can identify the support of an optimal solution (i.e., the zero and nonzero group structure of an optimal solution) in a finite number of iterations. This can be useful for applications like variable selection in regression problems (Tibshirani, 1996). It can also be used in combination with higher-order methods to design more efficient algorithms. For example, subspace acceleration methods benefit from F being smooth over the variables in the support of an optimal solution, even though F may be non-differentiable over the entire set of variables. For such methods, once the support is identified, more powerful methods (e.g., truncated Newton's method (Nocedal and Wright, 2006)) can be applied over the variables in the support to accelerate the local convergence rate (Wright, 2012; Chen et al., 2017; Curtis et al., 2022).

1.1 Related Work

The proximal stochastic gradient method (Rosasco et al., 2020) and its variants (Xiao and Zhang, 2014; Defazio et al., 2014; Wang et al., 2019; Pham et al., 2020; Tran-Dinh et al., 2022) are perhaps the most popular methods for solving problem (1). Since there is a large body of work on proximal stochastic gradient methods, we will (in alignment with the contributions of our work) focus on methods that have both a convergence guarantee and support identification property. Support identification is also sometimes referred to as manifold identification (Wright, 2012; Poon et al., 2018; Sun et al., 2019; Lee and Wright, 2012).

Proximal stochastic gradient-type methods are based on iterations that take the form

$$y_{k+1} \leftarrow \operatorname{prox}_{\alpha_k r}(x_k - \alpha_k d_k) \text{ with } \alpha_k > 0, \tag{2}$$

where $\operatorname{prox}_{\alpha_k r}(\cdot)$ is the proximal operator (Beck, 2017, Definition 6.1) associated with r and step size $\alpha_k > 0$ and d_k is an estimator of $\nabla f(x_k)$. If $d_k = \nabla \ell(x_k; \xi_k)$ for some realization ξ_k of the random variable ξ and $x_{k+1} = y_{k+1}$, then (2) recovers the proximal stochastic gradient method.

As observed by Poon et al. (2018) and Sun et al. (2019), the proximal stochastic gradient method does not have a support identification property because the error in the stochastic gradient estimator $\epsilon_k = d_k - \nabla f(x_k)$ does not vanish as k goes to infinity. One way of overcoming this deficiency is to employ variance reduction techniques. When r is the weighted ℓ_1-norm, Sun et al. (2019) considers the variance reduction properties of `ProxSVRG`, `SAGA`, and `RDA` (i.e, they consider whether $\mathbb{E}[\|\epsilon_k\|] \to 0$),[1] and establishes an active-set identification property (in expectation) for these three methods. Specifically, for a given sufficiently large k, they show that the zero groups of x_k agree with the zero groups of the optimal solution (in expectation). Moreover, when F is strongly convex so that a unique minimizer x^* exists, by knowing the rates at which $\{\mathbb{E}[\|x_k - x^*\|]\}$ and $\{\mathbb{E}[\|\epsilon_k\|]\}$ converge to zero, Sun et al. (2019, Theorem 4) establishes an upper bound, that holds in expectation, on the number of iterations before the zero variables are identified. When r is strongly convex, Lee and Wright (2012) establishes for `RDA` that, for any given sufficiently large k, the support of x_k matches that of x^* with high probability. (Observe that this means that the supports can match in some such iterations while not in other such iterations.) Later, Huang and Lee (2022) extends this result for `RDA` to the non-convex setting by making additional assumptions on the rate of convergence of the iterates and the step sizes.

A drawback of `ProxSVRG` and `SAGA` is that they are only applicable when problem (1) has a finite-sum structure, i.e., $\mathcal{P}$ is a uniform distribution over a finite set $\{1, 2, \ldots, N\}$. In particular, `ProxSVRG` requires an extra *exact* evaluation of ∇f every epoch, and `SAGA` requires one *exact* evaluation of ∇f in the first iteration and stores a history of stochastic gradients in a matrix of size $N \times n$, where N is the size of the data set and n is the number of optimization variables. Thus, `ProxSVRG` and `SAGA` are not practical for applications involving streaming data or large N.

The recent work by Cutkosky and Orabona (2019) and its extension by Xu and Xu (2020) consider a new stochastic gradient estimator called `Storm`. When `Storm` is combined with a proper step size selection strategy, it has a variance reduction property, and yet never requires an exact evaluation of ∇f. Our method `S-PStorm` draws inspiration from their work and introduces an iterate stabilization update to achieve a support identification property without having to store a history of stochastic gradients or to compute an exact evaluation of ∇f. The above results are summarized in Table 1.

Table 1: The first column gives the algorithm name. The second column shows the convergence rate of the iterates with $\rho_{\texttt{ProxSVRG}} > 0$ and $\rho_{\texttt{SAGA}} > 0$. The third column shows the support identification complexity where Δ^* and δ^* are positive constants (see (6) and (7)). (The Δ^* appearing in the result for our method `S-PStorm` is a consequence of our accounting for both zero and nonzero groups, whereas the other results are derived based on when only the zero groups are identified.) The result for `RDA` is valid when f and r are both strongly convex whereas the result for `S-PStorm` only assumes strong convexity of f. The fourth column indicates how often a method evaluates an exact gradient, and the fifth column gives the storage costs. The results for `ProxSVRG` and `SAGA` hold only when problem (1) has a finite-sum structure.

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	$\mathcal{O}(n)$

1.2 Contributions

This paper makes three main contributions.

1. We establish the variance reduction property (with high probability) of the `Storm` stochastic gradient estimator (Theorem 3.1), which is missing in Xu and Xu (2020). This is achieved by introducing a simple stabilization step in line 12 of Algorithm 1, which we show allows for a constant step size to be employed. This result is interesting in its own right, and the fact that our method allows for a constant step size to be used is a crucial property that we leverage in proving a support identification result.

[1] These results for `ProxSVRG`, `SAGA`, and `RDA` can be found in Table 2, Appendix C.3, and Appendix C.4 of (Sun et al., 2019).

2. To the best of our knowledge, RDA and our proposed S-PStorm are the only methods with a support identification property that neither require an exact gradient evaluation nor incur excessive storage costs. Compared with RDA, S-PStorm has a stronger notion of support identification (formalized in Definition 1.3). In particular, we show that, with high probability, *all* sufficiently large iterates in S-PStorm will correctly identify the support of the optimal solution. In contrast, RDA proves that each iterate with sufficiently large index will identify the support of the optimal solution with high probability (meaning that the support might be identified correctly in some iterations and not in others). We are able to obtain this stronger result as a consequence of the construction of the Storm stochastic gradient estimator and the added stabilization step, which allow for a sharp union bound (see Remark 3.6 for additional details).

3. Our numerical experiments on regularized logistic loss functions with weighted group ℓ_1-norm regularization show that S-PStorm outperforms popular methods in metrics that measure how efficiently and robustly iterates of an algorithm identify an optimal support, and in the final objective value achieved.

1.3 Notation and Preliminaries

Throughout the paper we use the following notation. We use $\|\cdot\|$ to represent the ℓ_2 norm, $|\mathcal{S}|$ to denote the cardinality of a set $\mathcal{S}$, and $\mathbb{N}_+$ and $\mathbb{R}_+$ to be the sets of positive integers and positive real numbers, respectively. For $N \in \mathbb{N}_+$, we define $[N] := \{1, 2, \cdots, N\}$. For $x \in \mathbb{R}^n$ and index set $\mathcal{I} \subseteq [n]$, we use $[x]_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ to denote the subvector of x that corresponds to the elements of $\mathcal{I}$. For two sequences of non-negative real numbers $\{\phi_k\}_{k\geq 1}$ and $\{\psi_k\}_{k\geq 1}$, we say $\phi_k = \mathcal{O}(\psi_k)$ if and only if there exist constants $k_0 \in \mathbb{N}_+$ and $M \in \mathbb{R}_+$ such that $\phi_k \leq M\psi_k$ for all $k \geq k_0$.

Let us now formally define what we mean by the support, a support identification property, and a consistent support identification property for a randomized algorithm.

Definition 1.1 (support)**.** The support of a point $x \in \mathbb{R}^n$ is denoted by $\mathcal{S}(x)$ and defined as

$$\mathcal{S}(x) := \{i \in [n_{\mathcal{G}}] \mid [x]_{g_i} \neq 0\},$$

where $\{g_i\}_{i=1}^{n_{\mathcal{G}}}$ forms a non-overlapping partition of $[n]$. We say that $x \in \mathbb{R}^n$ has optimal support if and only if $\mathcal{S}(x) = \mathcal{S}(x^*)$ for some solution $x^* \in \mathbb{R}^n$ to problem (1).

Definition 1.2 (support identification property)**.** A randomized algorithm is said to have the *support identification property* if and only if there exists $K \in \mathbb{N}_+$ and $p \in (0, 1]$ such that, when the algorithm generates a sequence of vectors $\{y_k\}_{k=1}^{\infty}$, one finds for each $k \geq K$ that the event $\{\mathcal{S}(y_k) = \mathcal{S}(x^*)\}$ occurs with probability at least p.

Definition 1.3 (consistent support identification property)**.** A randomized algorithm has the *consistent support identification property* if and only if there exist $K \in \mathbb{N}_+$ and $p \in (0, 1]$ so that, when the algorithm generates a sequence of vectors $\{y_k\}_{k=1}^{\infty}$, one finds that the event $\mathcal{E}_{\text{id}} := \bigcap_{k\geq K}^{\infty} \{\mathcal{S}(y_k) = \mathcal{S}(x^*)\}$ occurs with probability at least p.

While Lee and Wright (2012) and Sun et al. (2019) prove the support identification property of their algorithms (see Definition 1.2), we prove the stronger consistent support identification property (see Definition 1.3) for S-PStorm.

We next introduce some properties related to the proximal operator. For any $\alpha > 0$ and convex function r, the proximal operator $\text{prox}_{\alpha r}(\cdot)$ is single-valued. We define

$$\chi(x; \alpha) := \tfrac{1}{\alpha} \left\|\text{prox}_{\alpha r}(x - \alpha \nabla f(x)) - x\right\|, \quad (3)$$

which is the norm of the so-called gradient mapping, and is known to serve as an optimality measure for problem (1) (Beck, 2017, Theorem 10.7 (b)).

2 ALGORITHM

In this section, we present S-PStorm as Algorithm 1 for solving problem (1). At the beginning of iteration k, a mini-batch of independently and identically distributed (i.i.d) data samples $\{\xi_{k,i}\}_{i=1}^{m}$ are drawn according to the distribution $\mathcal{P}$, and two stochastic gradients v_k and u_k are formed at the current iterate x_k and the previous iterate x_{k-1} in (4)–(5). Then, the Storm stochastic gradient estimator is constructed in line 9. After performing the proximal stochastic gradient update to obtain y_k, a stabilization step is performed in line 12. As shown in the proof of Theorem 3.1, the stabilization step is critical because it allows for a constant step size strategy to be employed (i.e., $\alpha_k \equiv \underline{\alpha} > 0$ for all k), which in turn allows us to prove a consistent support identification result for S-PStorm.

3 ANALYSIS

We begin this section by introducing the assumptions under which our convergence analysis is performed.

3.1 Assumptions

Our first assumption concerns strong convexity of f and Lipschitz continuity of the gradient of the loss function ℓ.

Assumption 3.1. The following hold:

1. f is μ_f-strongly convex over $\mathbb{R}^n$ and r_i is convex and closed over $\mathbb{R}^n$ for all $i \in [n_{\mathcal{G}}]$.

2. There exists a constant $L_g > 0$ such that, for any $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ and any $\xi \sim \mathcal{P}$, it holds that

$$\|\nabla \ell(x, \xi) - \nabla \ell(y, \xi)\| \leq L_g \|x - y\|,$$

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
$$v_k \leftarrow \frac{1}{m}\sum_{i=1}^{m} \nabla \ell(x_k; \xi_{ki}). \tag{4}$$
  if $k = 1$ **then**
    Set $d_k \leftarrow v_k$.
  else
    Set
$$u_k \leftarrow \frac{1}{m}\sum_{i=1}^{m} \nabla \ell(x_{k-1}; \xi_{ki}). \tag{5}$$
    Set $d_k \leftarrow v_k + (1-\beta_k)(d_{k-1} - u_k)$.
  end if
  Compute $y_k \leftarrow \text{prox}_{\alpha_k r}(x_k - \alpha_k d_k)$.
  Set $x_{k+1} \leftarrow x_k + \zeta\beta_k(y_k - x_k)$.
end for

i.e., ∇f is L_g-Lipschitz continuous.

The strong convexity assumption on f is for deriving a complexity result for consistent support identification. This assumption can be relaxed to f being convex if, similar to Sun et al. (2019), we instead assume that there exists a decreasing sequence $\{\nu_k\}$ such that $\mathbb{P}(\{\|x_k - x^*\| \leq \nu_k\}) = 1$. Under this assumption, we can also prove a consistent support identification result for `S-PStorm`, although without an explicit upper bound on K in Definition 1.3—whereas under Assumption 3.1 we provide such an upper bound. The smoothness assumption on $\ell(\cdot, \xi)$ is standard (Cutkosky and Orabona, 2019; Xu and Xu, 2020).

For our next assumption, we refer to the filtration—defined by the initial point and sequence of mini-batch stochastic gradients—corresponding to the stochastic process generated by the algorithm. Denoting $\mathcal{F}_1 := \sigma(x_1)$ and, for all $k \geq 2$, denoting $\mathcal{F}_k$ as the σ-algebra generated by the random variables $\{\{\Xi_{1,i}\}_{i=1}^m, \ldots, \{\Xi_{(k-1),i}\}_{i=1}^m\}$ (of which $\{\{\xi_{1,i}\}_{i=1}^m, \ldots, \{\xi_{(k-1),i}\}_{i=1}^m\}$ is a realization), it follows that $\{\mathcal{F}_k\}$ is this filtration of interest. Recall that the distribution $\mathcal{P}$ of ξ is independent of the filtration.

Assumption 3.2. The following hold:

1. For all $k \geq 1$, $\mathbb{E}_{\xi\sim\mathcal{P}}[\nabla \ell(x_k;\xi) \mid \mathcal{F}_k] = \nabla f(x_k)$.
2. There exists $G_r \in \mathbb{R}_+$ such that, for all $k \geq 1$, $\mathbb{P}\{\|g_r\|_2 \leq G_r,\ \forall g_r \in \partial r(x_k)\} = 1$.
3. There exists $\sigma \in \mathbb{R}_+$ such that, for all $k \geq 1$, $\mathbb{P}_{\xi\sim\mathcal{P}}\{\|\nabla \ell(x_k, \xi) - \nabla f(x_k)\| \leq \sigma \mid \mathcal{F}_k\} = 1$.
4. There exists $G_d \in \mathbb{R}_+$ such that, for all $k \geq 1$, $\mathbb{P}_{\xi\sim\mathcal{P}}\{\|d_k\| \leq G_d \mid \mathcal{F}_k\} = 1$.

Assumption 3.2(1) ensures that the stochastic gradient $\nabla \ell(x;\xi)$ is an unbiased estimator of the gradient $\nabla f(x)$ for all $x \in \mathbb{R}^n$. Assumption 3.2(2) provides a constant upper bound on the norm of an element of $\partial r(x)$ for all $x \in \mathbb{R}^n$, which exists when r is the weighted ℓ_1-norm or weighted group ℓ_1-norm, for example. Assumption 3.2(3) guarantees (almost surely) a bound on the difference between $\nabla \ell(x_k;\xi)$ and $\nabla f(x_k)$ for all $k \in \mathbb{N}_+$. This assumption is implied by the uniform bound assumption on $\nabla \ell(x;\xi)$ used in (Liu et al., 2022, Assumption 4). It may be possible to relax Assumption 3.2(3) by assuming that the stochastic gradient error has a sub-exponential tail, e.g., Na et al. (2022), which we leave as future work. Assumption 3.2(4) is implied by the following two, perhaps more natural, assumptions: (i) There exists a constant $c_e > 0$ such that, for all k, it holds that $\mathbb{P}\{\|d_k - \nabla f(x_k)\| \leq c_e | \mathcal{F}_k\} = 1$, i.e., the error in the stochastic gradient estimator d_k is almost surely bounded; and (ii) There exists a constant c_α such that, for a given $\alpha > 0$ and all $k \geq 1$, it holds that $\mathbb{P}\{\chi(x_k;\alpha) \leq c_\alpha | \mathcal{F}_k\} = 1$ (also see (3)), i.e., the optimality measure is almost surely bounded. Note that Assumption 3.2(4) is slightly weaker than a bounded iterates assumption, which is also made in `RDA` (Lee and Wright, 2012). A proof that Assumption 3.2(4) follows from (i) and (ii) can be found in Appendix A.4.

Our last assumption is on the parameters of Algorithm 1.

Assumption 3.3. The sequences $\{\beta_k\}$ and $\{\alpha_k\}$ in Algorithm 1 are chosen, with $c > 1$ and $\underline{\alpha} \in (0,\infty)$, to satisfy $\beta_k = \min\{1/2, c/(k+1)\}$ and $\alpha_k \equiv \underline{\alpha}$ for all $k \geq 1$.

The constant $1/2$ appearing in the definition of β_k in Assumption 3.3 can be replaced by any constant between zero and one; the choice of $1/2$ is to simplify expressions appearing throughout our analysis.

3.2 Convergence Analysis

The first result establishes the variance reduction property of the `Storm` stochastic gradient estimator.

Theorem 3.1. *Let Assumption 3.1–Assumption 3.3 hold, let $\epsilon_k = d_k - \nabla f(x_k)$ for all $k \in \mathbb{N}_+$, and define $\underline{k} = \lceil (2c) - 1 \rceil$. Then, for any $k \geq \underline{k}$ and any $\eta_k \in (0,1)$, the event $\mathcal{E}_k := \{\|\epsilon_k\| \leq U(k)\}$ holds with probability at least $1 - \eta_k$, where for some constant $C \in \mathbb{R}_+$ independent of k, one defines*

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

The proof of Theorem 3.1 is presented in Appendix A.1.

Remark 3.1. The upper bound $U(k)$ in Theorem 3.1 is independent of the mini-batch size m. This is due to the bound in Assumption 3.2(3) that holds almost surely.

Remark 3.2. By setting $\eta_k = \frac{\eta_0}{k^2}$ for all $k \in \mathbb{N}_+$ with constant $\eta_0 \in (0, 6/\pi^2)$, one obtains $U(k) = \mathcal{O}(\max\{\sqrt{\log k}/k^c, \sqrt{\log k/k}\})$ so that $\{\|\epsilon_k\|\} \to 0$ with high probability. This is formalized in the next result.

Corollary 3.1. *Let Assumption 3.1–Assumption 3.3 hold, $\eta_k = \frac{\eta_0}{k^2}$ for all $k \geq 1$ with $\eta_0 \in (0, 6/\pi^2)$, and $\mathcal{E}_k$ be defined as in Theorem 3.1. Then, the event $\mathcal{E} := \bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k$ happens with probability at least $1 - \frac{\eta_0 \pi^2}{6}$.*

The proof of Corollary 3.1 can be found in Appendix A.2.

Next, we establish the rate of convergence of the iterate sequence $\{x_k\}$ with high probability (for small η_0).

Theorem 3.2. *Let $\underline{\alpha} = \mu_f / L_g^2$, $\zeta \in (0, 2)$, $\theta \geq 2$, $c = (2\theta L_g^2)/(\zeta \mu_f^2) > 2$, and $\underline{k} = \lceil 2c - 1 \rceil$. Set $\eta_k = \eta_0 / k^2$ for all $k \geq 1$ with $\eta_0 \in (0, 6/\pi^2)$. Then, under Assumption 3.1–Assumption 3.3, there exists a constant $C_3 \in \mathbb{R}_+$ independent of k, such that the event $\mathcal{E}_k^x := \left\{ \|x_k - x^*\|^2 \leq \bar{c}_1 \frac{\|x_{\underline{k}} - x^*\|^2}{k^\theta} + \bar{c}_2 \cdot \frac{\log \frac{2k}{\eta_0}}{k} \right\}$ with $\bar{c}_1 := (\underline{k} + 2)^\theta$ and $\bar{c}_2 := C_3 \zeta \left(\frac{\mu_f^2}{L_g^4} + \frac{2}{L_g^2} \left(1 + \frac{\mu_f}{L_g}\right)^2 \right) (\sigma + L_g(G_r + G_d)\zeta\underline{\alpha})^2$ satisfies*

$$\mathbb{P}\left[\bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k^x \right] \geq 1 - \eta_0 \pi^2/6 > 0.$$

The proof of Theorem 3.2 is presented in Appendix A.3.

Remark 3.3. Theorem 3.2 provides a $\mathcal{O}(\sqrt{\log k/k})$ convergence rate for $\|x_k - x^*\|$ for all $k \geq \underline{k}$ with high probability. It is worth noting that the constant $\underline{k}$ depends on the square of the condition number L_g/μ_f. We also note that the first term $\bar{c}_1 \|x_{\underline{k}} - x^*\|^2 / k^\theta$ can be made to converge to zero arbitrarily fast by choosing θ as large as desired, although this results in larger $\underline{k}$. It is the second term $\bar{c}_2 \log(\frac{2k}{\eta_0})/k$ that dictates the overall convergence rate of the iterates. This rate of convergence is obtained by using the rate at which the error in the `Storm` stochastic gradient estimator converges to zero (see Remark 3.2).

Remark 3.4. Theorem 3.2 establishes a sub-linear rate of convergence for the iterates with high probability for strongly convex loss functions. However, it remains unknown whether there exists a method that has a linear convergence rate for strongly convex functions and avoids huge storage and exact gradient evaluations.

3.3 Support Identification

In this section, we restrict our attention to r being the weighted non-overlapping group ℓ_1 regularizer, i.e., $r(x) = \sum_{i=1}^{n_\mathcal{G}} \lambda_i \|[x]_{g_i}\|$ with $n_\mathcal{G} > 0$, $\{g_i\} \subseteq [n]$ for each $i \in [n_\mathcal{G}]$, $\bigcup_{i=1}^{n_\mathcal{G}} g_i = [n]$, $g_i \cap g_j = \emptyset$ for all $i \neq j$, and $\{\lambda_i\}_{i=1}^{n_\mathcal{G}}$ strictly positive group weights.

Let us now introduce quantities that are crucial for establishing our support identification result. Specifically, let x^* be the unique solution to problem (1). Define

$$\Delta := \begin{cases} \min_{i \in \mathcal{S}(x^*)} \|[x^*]_{g_i}\| & \text{if } \mathcal{S}(x^*) \neq \emptyset, \\ 1 & \text{if } \mathcal{S}(x^*) = \emptyset, \end{cases}$$

$$\Delta^* := \min\{1, \Delta\}, \tag{6}$$

$$\delta_{\min} := \begin{cases} \min_{i \notin \mathcal{S}(x^*)} \{\lambda_i - \|\nabla_{g_i} f(x^*)\|\} & \text{if } \mathcal{S}(x^*) \subsetneqq [n_\mathcal{G}], \\ 1 & \text{if } \mathcal{S}(x^*) = [n_\mathcal{G}], \end{cases}$$

$$\delta^* := \min\{\delta_{\min}, 1\}. \tag{7}$$

Geometrically, Δ captures the minimum ℓ_2-norm of the groups that are non-zero at x^*, taking into account the possibility that $\mathcal{S}(x^*)$ is empty. The definition of $\delta_{\min}$ measures the minimum distance between λ_i and the corresponding optimal dual variables (see (9)) for groups not in $\mathcal{S}(x^*)$. To see this, without loss of generality, suppose that $\mathcal{S}(x^*) \subsetneqq [n_\mathcal{G}]$. For any $\alpha > 0$ define $z^* := x^* - \alpha \nabla f(x^*)$ and then consider the proximal problem

$$\min_{x \in \mathbb{R}^n} \phi_p(x; x^*, \alpha) := \tfrac{1}{2\alpha} \|x - z^*\|^2 + r(x) \tag{8}$$

and its dual problem

$$\max_{\omega \in \mathbb{R}^n} \phi_d(\omega; x^*, \alpha) := -\left(\tfrac{\alpha}{2}\|\omega\|_2^2 + \omega^T z^*\right) \text{ s.t. } r_*(\omega) \leq 1 \tag{9}$$

where $r_*(\omega) = \max_{i \in [n_\mathcal{G}]} \frac{\|[\omega]_{g_i}\|}{\lambda_i}$ is the dual norm of the weighted group ℓ_1 norm. It can be seen that x^* is the optimal solution to the primal problem (8) (Beck, 2017, Theorem 10.7). Denoting ω^* as the optimal solution to the dual problem (9), it follows that

$$[\omega^*]_{g_i} = -\min\left\{\tfrac{1}{\alpha}, \tfrac{\lambda_i}{\|[z^*]_{g_i}\|}\right\}[z^*]_{g_i} \quad \text{for all } i \in [n_\mathcal{G}]. \tag{10}$$

Then, by the Fenchel-Young inequality (Rockafellar, 1970, Theorem 31.1), it follows that

$$x^* = \alpha \omega^* + z^*. \tag{11}$$

Combining the definition of z^* and (11), one establishes that $\omega^* = \nabla f(x^*)$. Therefore, $\delta_{\min}$ measures the minimum distance from $[\omega^*]_{g_i}$ to the boundary of the ball centered at origin with distance λ_i for all $i \notin \mathcal{S}(x^*)$.

The discussion above leads to a non-degeneracy assumption: For groups of variables not in $\mathcal{S}(x^*)$, their corresponding dual variables are strictly feasible, i.e., $\|[\omega^*]_{g_i}\| < \lambda_i$ for all $i \notin \mathcal{S}(x^*)$. Let us formally state this non-degeneracy assumption using $\omega^* = \nabla f(x^*)$ to make it consistent with the literature (Lee and Wright, 2012; Poon et al., 2018; Sun et al., 2019; Curtis et al., 2022).

Assumption 3.4. The scalar δ^* in (7) satisfies $\delta^* > 0$.

With the non-degeneracy assumption in hand, we may now give a sufficient condition for support identification.

Theorem 3.3. *Let Assumption 3.4 hold. Given $\alpha > 0$, $d \in \mathbb{R}^n$, and the optimal solution x^* to problem (1), let us define $z = x - \alpha d$ and $y = \operatorname{prox}_{\alpha r}(z)$. If*

$$\left\| \frac{[z - x^*]_{g_i}}{\alpha} + \nabla_{g_i} f(x^*) \right\| < \delta^* \text{ for all } i \notin \mathcal{S}(x^*),$$

then $\mathcal{S}(y) \subseteq \mathcal{S}(x^*)$. Furthermore, if $\|y - x^*\| < \Delta^*$, then $\mathcal{S}(x^*) \subseteq \mathcal{S}(y)$ so that, in fact, $\mathcal{S}(y) = \mathcal{S}(x^*)$.

The proof of Theorem 3.3 is presented in Appendix A.4.

Remark 3.5. Theorem 3.3 extends the result in (Sun et al., 2019, Lemma 1) from the ℓ_1 regularizer to the group ℓ_1 regularizer considered here. Also, our result slightly strengthens theirs since they only discuss the result $\mathcal{S}(y) \subseteq \mathcal{S}(x^*)$.

Using the sufficient conditions for support identification from Theorem 3.3, the result of consistent support identification (Definition 1.3) can now be established.

Theorem 3.4. *Let Assumption 3.1–Assumption 3.4 hold, $\zeta \in (0, 2)$, $\theta \geq 2$, $c = (2\theta L_g^2)/(\zeta \mu_f^2) > 2$, and $\underline{k} = \lceil 2c - 1 \rceil$. Consider the sequence $\{y_k\}$ of Algorithm 1 and define the event $\mathcal{E}_k^{id} = \{\mathcal{S}(y_k) = \mathcal{S}(x^*)\}$ for all $k \geq 1$. Then, there exists constants $\{C_{41}, C_{42}\} \subseteq \mathbb{R}_+^n$ that are independent of k, $k_{\delta^*} = (C_{41}/\delta^*)^4$ and $k_{\Delta^*} = (C_{42}/\Delta^*)^4$ such that, with $K := \max\{k_{\delta^*}, k_{\Delta^*}, \underline{k}\}$, it follows that*

$$\mathbb{P}\left[\bigcap_{k \geq K}^{\infty} \mathcal{E}_k^{id}\right] \geq 1 - \frac{\eta_0 \pi^2}{6} > 0.$$

The proof of Theorem 3.4 is presented in Appendix A.5.

Remark 3.6. Using Theorem 3.4 and results from Xiao (2009) and Lee and Wright (2012), we can also derive a high probability support identification complexity bound for `RDA` for any given iterate x_k, which is different from the result in Sun et al. (2019, Theorem 5). To do so, we need extra assumptions on the function r that do not hold for the weighted group ℓ_1-norm, and boundedness of $\{\nabla \ell(x_k; \xi_k)\}$ generated by `RDA`[2]. Specifically, we consider the update of `RDA` as $x_{k+1} = \operatorname{prox}_{\alpha_k r}(-\alpha_k d_k)$ with $\alpha_k = \frac{\sqrt{k}}{\alpha}$, where $d_k = \frac{1}{k}\sum_{i=1}^k \nabla \ell(x_i; \xi_i)$[3]. It follows from Lemma A.5(3) that

$$\mathbb{P}\left[\mathcal{S}(x_{k+1}) = \mathcal{S}(x^*)\right] \geq 1 - \eta_k^{\texttt{RDA}}$$

where

$$\eta_k^{\texttt{RDA}} = \max\left\{\mathcal{O}\left(\frac{1}{\delta^* \cdot k^{1/4}}\right), \mathcal{O}\left(\frac{1}{\Delta^* \cdot k^{1/4}}\right)\right\}.$$

[2]See Lemma A.5(2) for precise details of the assumptions.

[3]See Lemma A.5(1) to see how this form of the update is equivalent to the `RDA` update presented in Xiao (2009).

Since $\sum_{k=1}^{\infty} \eta_k^{\texttt{RDA}}$ diverges, one cannot give a lower bound on $\mathbb{P}\left[\cap_{k \geq K^{\texttt{RDA}}}^{\infty} \{\mathcal{S}(x_k) = \mathcal{S}(x^*)\}\right]$ for some sufficiently large $K^{\texttt{RDA}}$. Instead, for any $\eta_0 \in (0, 1)$, there exists a $\bar{k} = \mathcal{O}\Big(\max\Big\{\left(\frac{1}{\eta_0 \delta^*}\right)^4, \left(\frac{1}{\eta_0 \Delta^*}\right)^4\Big\}\Big)$ such that any given $k \geq \bar{k}$ satisfies $\mathbb{P}[\mathcal{S}(x_{k+1}) = \mathcal{S}(x^*)] \geq 1 - \eta_0$. This establishes the support identification property (see Definition 1.2). However, in Theorem 3.4 we show that `S-PStorm` has a consistent support identification property (see Definition 1.3), which is a stronger result. Lastly, we note that the K value appearing in Theorem 3.4 grows with the condition number L_g/μ_f.

Remark 3.7. Similar to Sun et al. (2019), under additional assumptions, it is possible to extend Theorem 3.4 to the case that f is convex. In particular, if we assume that $\|x_k - x^*\| \leq A_k$ for some optimal solution x^* and a decreasing sequence $\{A_k\}$ with some positive probability (for example, with probability $1 - \eta_k$) for all $k \geq 1$, then we can prove a support identification result, but we no longer have a complexity bound.

4 NUMERICAL EXPERIMENTS

4.1 Problems, Baselines, and Implementation Details

Problems. We consider solving problem (1) with $f(x)$ and $r(x)$ given by the regularized binary logistic loss and group-ℓ_1 regularizer, respectively, resulting in the problem

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

where N is the number of data points, $d_j \in \mathbb{R}^n$ is the jth data point, $y_j \in \{-1, 1\}$ is the class label for the jth data point, and $\lambda_i > 0$ for all $(j, i) \in [N] \times [n_{\mathcal{G}}]$. Data sets for the logistic regression problems were obtained from the LIBSVM repository.[4] We excluded all multi-class (greater than two) classification datasets, datasets with feature less than 50 or samples less than 10000, and all data sets that were too large ($\geq$ 16GB)[5]. Finally, for the adult data (a1a–a9a) and webpage data (w1a–w8a), we used only the largest instances, namely a9a and w8a. This left us with our final subset of 10 data sets that can be found in Table 2. Following Xiao and Zhang (2014), we scaled each data point to have a unit norm, i.e., $\|d_j\| = 1$ for all $j \in [N]$.

For each dataset, we considered four group structures and two different solution sparsity levels, which led to 80 test instances in total. We considered the four different numbers of groups in $\{\lfloor 0.25n \rfloor, \lfloor 0.50n \rfloor, \lfloor 0.75n \rfloor, n\}$, where n

[4]https://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets

[5]Memory usage is counted by a Python object instead of the raw txt file. We also exclude the dataset epsilon since we had an error message indicating a wrong data format in line 33334.

Table 2: Description of the data sets.

data set	N	n
a9a	32561	123
avazu-app.tr	12,642,186	1,000,000
covtype	581,012	54
kdd2010	8,407,752	20,216,830
news20	19,996	1,355,191
phishing	11,055	68
rcv1	20,242	47,236
real-sim	72,309	20,958
url	2,396,130	3,231,961
w8a	49,749	300

is the problem dimension; notice that the last setting recovers ℓ_1-norm regularization. Then, for a given number of groups, the variables were sequentially distributed (as evenly as possible) to the groups; e.g., 10 variables among 3 groups would have been distributed as $g_1 = \{1, 2, 3\}$, $g_2 = \{4, 5, 6\}$, and $g_3 = \{7, 8, 9, 10\}$. We considered two different solution sparsity levels obtained by adjusting the group weights $\{\lambda_i\}$. Specifically, we considered group weights $\lambda_i = \Lambda\sqrt{|g_i|}$ for all $i \in [n_{\mathcal{G}}]$ with $\Lambda = 0.1\Lambda_{\min}$ and $\Lambda = 0.01\Lambda_{\min}$, where $\Lambda_{\min}$ is the minimum positive number such that the solution to the logistic problem with $\lambda_i = \Lambda_{\min}\sqrt{|g_i|}$ is $x = 0$. See Yang and Zou (2015, equation (23)) for the formula to compute $\Lambda_{\min}$.

Baselines. We choose `ProxSVRG` (Xiao and Zhang, 2014), `SAGA` (Defazio et al., 2014), and `RDA` (Xiao, 2009) as baselines since they have theoretical guarantees for identifying the support. We also include `PStorm` (Xu and Xu, 2020) to demonstrate the empirical importance of the modification we made in `S-PStorm` (i.e., the stabilization step in Line 12). We use `FaRSA-Group` (Curtis et al., 2022), a deterministic second-order method, to find a highly accurate estimate to the optimal solution x^* for each test instance by solving the problem to high accuracy (10^{-8}), as measured by the norm of the gradient mapping in (3).

Implementation Details We implemented a version of `ProxSVRG` as described in Poon et al. (2018, Equation (8) Option II), `SAGA` as described in Poon et al. (2018, Equation (6)), `RDA` as described in Lee and Wright (2012, Algorithm 1), and `PStorm` as described in Xu and Xu (2020, Algorithm 1)[6]. (i) **Step size strategy**: For `ProxSVRG`, `SAGA`, and `S-PStorm`, we used a constant step size strategy by setting $\alpha_k \equiv 0.1/L_g$, which follows the choice made in Xiao and Zhang (2014). We remark that L_g can be estimated by $1/4$ since the data set is normalized instance-wise (see Xiao and Zhang (2014, Section 4.1) for the reason). For `RDA`, the step size was set as $\alpha_k = \sqrt{k}/\gamma$.[7] We tuned γ by choosing its value from the set $\{10^j\}_{j\in\{-4,-3,\ldots,2\}}$ using the 32 test instances obtained from the datasets a9a, covtype, phishing, and w8a, and found that $\gamma = 10^{-2}$ worked the best. For `PStorm`, we used $\alpha_k = \frac{4^{1/3}/(8L_g)}{(k+4)^{1/3}}$ as suggested in Xu and Xu (2020, Theorem 2). (ii) **Algorithm specific parameters**: `ProxSVRG` is a double loop algorithm and we set the inner loop length to 1, i.e., the parameter P in Poon et al. (2018, Equation (8) Option II) was set to 1. For `RDA` the prox-function h was chosen as the square of the ℓ_2 norm. For `PStorm` we used $\beta_k = \frac{1+24\alpha_k^2 L_g^2 - \frac{\alpha_{k+1}}{\alpha_k}}{1+4\alpha_k^2 L_g^2}$, and for `S-PStorm` we used $\beta_k = \frac{1}{k+1}$. The ζ parameter is chosen in an adaptive way to improve the practical performance. In particular, ζ is initialized to 1 and increased by 1 after an iteration is completed. Although this choice is not covered by the convergence theory, one could cap the number of adjustments made to ζ, in which case it is covered by the theory. For all algorithms, the batch size was set to 256 and the starting point was the zero vector. (iii) **Termination conditions**: A test instance was terminated when either 1000 epochs was reached, or a 12 hour time limit was reached. We note that `SAGA` terminated immediately on all test instances associated with the datasets avazu-app.tr, kdd2010, news20, real-sim, and url because the storage of the gradient look-up table exceeded the memory limit.

4.2 Numerical Results

Experiments were run on a cluster with 16 AMD Opteron Processor 6128 2.0 GHz CPUs and 32 GB memory.

Support Identification Performance. We considered four metrics for measuring an algorithm's performance on support identification. Specifically, we computed the supports of the iterates $\{x_{kb} \mid k = 1\cdots, 1000\}$ with $b = \lceil N/m\rceil$, where m was the mini-batch size. The sequence $\{x_{kb}\}$ can be thought of as the "major iterates" resulting after each full data-pass. The first metric was the *total number of identifications*, which measured the number of iterates in $\{x_{kb}\}$ that correctly identified the support $\mathcal{S}(x^*)$ (the larger the better); the second metric was the *first identification*, which was the smallest $k_0 \in [1000]$ such that $x_{k_0 b}$ identified the support $\mathcal{S}(x^*)$ (the smaller the better); the third metric was the *first consistent identification*, which was the smallest $K \in [1000]$ such that all $\{x_{kb}\}_{k\geq K}$ identified the support $\mathcal{S}(x^*)$ (the smaller the better); the last metric was the *last iterate support recovery*, which was defined as $1 - \frac{|\mathcal{S}(x_{1000b})\Delta\mathcal{S}(x^*)|}{|\mathcal{S}(x^*)|}$ (the closer to 1 the better) with Δ being the set symmetric difference. The *last iterate support recovery* metric was introduced because we observed that all five algorithms failed to identify the support $\mathcal{S}(x^*)$ on some test instances generated by the larger datasets (e.g., url) as a result of not getting an accurate enough approximate solution. Nonetheless, when the algorithms termi-

[6] The code is publicly available at https://github.com/Yutong-Dai/S-PStorm.

[7] The original paper used β_k to denote the step size. See part (1) of Lemma A.5 for how to map β_k to α_k.

nated, the last iterates still had sparse structure, and the *last iterate support recovery* metric measured how close the algorithm was to identifying the true support.

For every test instance solved by a given algorithm, we repeated the experiments for 3 independent runs and for each run compute the four metrics, which are then averaged to obtain the final values of the metrics for the algorithms. For a given test instance and metric, we assigned scores from $\{1, 2, 3, 4, 5\}$ to the 5 algorithms based on their ranked performances. The better an algorithm performed, the higher the score it received. The best performer received a score of 5, the second best performer received a score of 4, and so forth.[8] For the first three metrics, if an algorithm failed to identify the support before it terminated, we assigned the algorithm a score of 0. For each metric, we summed over all test instances to get the final scores for each algorithm and then normalized the scores so that the scores for all algorithms under a given metric summed to one.

We present the normalized scores for the 5 algorithms over the 4 metrics in Figure 1, and provide the raw data for these metrics in Appendix B.2. One can see that `S-PStorm` consistently outperformed the other algorithms on all 4 metrics by a significant margin.

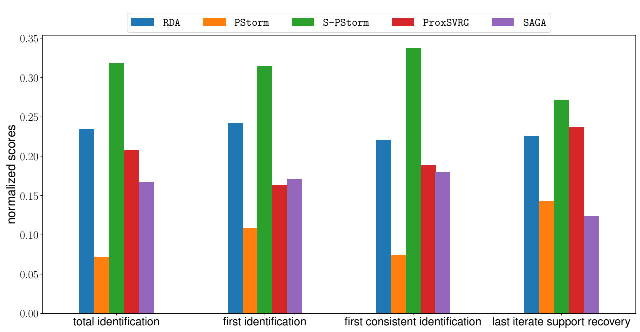


Figure 1: Normalized scores for four metrics that evaluate the performance of the support identification.

Solution Quality. We measure the solution quality of an algorithm by computing the optimal objective function value gap. Specifically, for a given test instance, denote $F^* = \min_j\{F_j^{\text{best}}\}$, where $F_j^{\text{best}} = \min_{b\in[1000]}\{F(x_{kb}^j)\}$ with $j \in$ {`ProxSVRG`, `SAGA`, `RDA`, `PStorm`, `S-PStorm`} and $\{x_{kb}^j\}$ generated by the jth algorithm. If algorithm j failed on a given problem instance (due to insufficient memory), we set $F_j^{\text{best}} = \infty$. Then, we compute the optimal objective function value gap as $(F_j^{\text{best}} - F^*)/\max\{1, F^*\}$ for all j. The results are visualized in Figure 2. The deeper the blue color of a rectangle for an algorithm, the better it performed in terms of achieving a lower objective value. On the flip side, the deeper the red color of a rectangle for an algorithm, the worse it performed in terms of achieving a lower objective value. In Appendix B.1, we provide a discussion on the performance gap for the different methods.

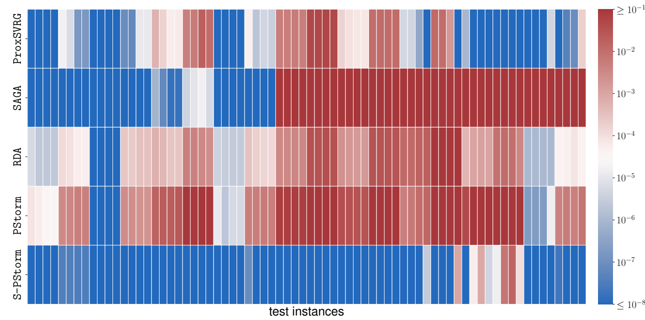


Figure 2: Visualization of objective value gaps for different methods. Each rectangular represents a test instance.

Together Figure 1 and Figure 2 illustrate that `S-PStorm` performed significantly better in both support identification and achieving better objective function values.

Lastly, in Appendix B.2, we illustrate how the distance to the optimal solution $\|x_k - x^*\|$ (x^* is obtained using the `FaRSA-Group` algorithm) and error ϵ_k in the gradient estimator converge to 0. It can be observed empirically that the rates at which $\{\epsilon_k\}$ converges to 0 and $\{x_k\}$ converges to x^* agree with our $\mathcal{O}(\sqrt{\log k/k})$ convergence result (see Remark 3.2 and Remark 3.3).

5 CONCLUSION

This paper proposes a new variance-reduced and stabilized stochastic proximal gradient method `S-PStorm` for stochastic optimization with structured sparsity. Compared with existing methods, `S-PStorm` has two new advantages. In terms of theoretical results, `S-PStorm` has the consistent support identification property, which has not been proved for `RDA`. Regarding the efficiency and deployability, `S-PStorm` neither requires any exact gradient evaluations nor needs to store a history of stochastic gradients. Numerical experiments on regularized logistic loss problems show that `S-PStorm` outperforms popular methods in terms of both support identification and final objective function values obtained.

Future directions. First, it would be interesting to investigate whether our consistent support identification results extend to the non-convex setting. Second, our convergence and support identification results rely on exact evaluations of proximal operator, but some proximal operators, for example, overlapping group ℓ_1 regularizers (Obozinski et al., 2011; Yuan et al., 2013), do not admit closed-form solutions. We believe our results can be extended to this setting provided a subproblem solver is carefully designed to produce inexact proximal operator solutions geared towards support identification (Dai and Robinson, 2022).

[8] When two or more algorithms obtained the same value for a metric, we assign them all the same score.

Acknowledgements

We thank the reviewers for their constructive comments that helped improve the paper. The authors Yutong Dai, Frank E. Curtis, and Daniel P. Robinson were supported by the US National Science Foundation grant DMS-2012243. The author Guanyi Wang was supported by the Singapore MOE under AcRF Tier-1 grant 22-5539-A0001.

References

Beck, A. (2017). *First-order methods in optimization*, volume 25. SIAM.

Chen, T., Curtis, F. E., and Robinson, D. P. (2017). A reduced-space algorithm for minimizing ℓ_1-regularized convex functions. *SIAM Journal on Optimization*, 27(3):1583–1610.

Curtis, F. E., Dai, Y., and Robinson, D. P. (2022). A subspace acceleration method for minimization involving a group sparsity-inducing regularizer. *SIAM Journal on Optimization*, 32(2):545–572.

Cutkosky, A. and Orabona, F. (2019). Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32.

Dai, Y. and Robinson, D. P. (2022). Inexact proximal-gradient methods with support identification. *arXiv preprint arXiv:2211.02214*.

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

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hazan, E. et al. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325.

Huang, Z.-S. and Lee, C.-p. (2022). Training structured neural networks through manifold identification and variance reduction. In *International Conference on Learning Representations*.

Lee, S. and Wright, S. J. (2012). Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13(6).

Liu, Z., Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. L. (2022). Meta-storm: Generalized fully-adaptive variance reduced sgd for unbounded functions. *arXiv preprint arXiv:2209.14853*.

Na, S., Dereziński, M., and Mahoney, M. W. (2022). Hessian averaging in stochastic newton methods achieves superlinear convergence. *Mathematical Programming*, pages 1–48.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Obozinski, G., Jacob, L., and Vert, J.-P. (2011). Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*.

Pham, N. H., Nguyen, L. M., Phan, D. T., and Tran-Dinh, Q. (2020). Proxsarah: An efficient algorithmic framework for stochastic composite nonconvex optimization. *J. Mach. Learn. Res.*, 21(110):1–48.

Pinelis, I. (1994). Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706.

Poon, C., Liang, J., and Schoenlieb, C. (2018). Local convergence properties of saga/prox-svrg and acceleration. In *International Conference on Machine Learning*, pages 4124–4132. PMLR.

Rockafellar, R. T. (1970). *Convex analysis*, volume 18. Princeton university press.

Rosasco, L., Villa, S., and Vũ, B. C. (2020). Convergence of stochastic proximal gradient algorithm. *Applied Mathematics & Optimization*, 82(3):891–917.

Sun, Y., Jeong, H., Nutini, J., and Schmidt, M. (2019). Are we there yet? manifold identification of gradient-related proximal methods. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1110–1119. PMLR.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. (2022). A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, 191(2):1005–1071.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. (2019). Spiderboost and momentum: Faster variance reduction algorithms. *Advances in Neural Information Processing Systems*, 32.

Wright, S. J. (2012). Accelerated block-coordinate relaxation for regularized optimization. *SIAM Journal on Optimization*, 22(1):159–186.

Xiao, L. (2009). Dual averaging method for regularized stochastic learning and online optimization. *Advances in Neural Information Processing Systems*, 22.

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

Xu, Y. and Xu, Y. (2020). Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization.

Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141.

Yuan, L., Liu, J., and Ye, J. (2013). Efficient methods for overlapping group lasso. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2104–2116.

A Proofs of Results in Section 3

A.1 Proof of Theorem 3.1

We first establish some useful lemmas. The first lemma establishes an upper bound on $\left(\prod_{j=i}^{k}(1-\beta_j)\right)$, which will be used later to prove the variance reduction property.

Lemma A.1. *Under Assumption 3.3 and with $\underline{k} = \lceil (2c) - 1 \rceil$, it holds for all $k \geq \underline{k}$ and $i \in \{2, 3, \cdots, k\}$ that*

$$\left(\prod_{j=i}^{k}(1-\beta_j)\right) \leq \exp\left(-\frac{\underline{k} - \min\{\underline{k}, i\}}{2}\right)\left(\frac{\max\{\underline{k}, i\}+1}{k+2}\right)^c.$$

Proof. One can see from Assumption 3.3 that

$$\beta_j = \begin{cases} \frac{1}{2} & \text{if } j < \underline{k} \\ \frac{c}{j+1} & \text{if } j \geq \underline{k}. \end{cases}$$

It follows from the above inequality and the fact that $1 - x \leq \exp(-x)$ for all $x \in \mathbb{R}$ that

$$\begin{aligned}
\left(\prod_{j=i}^{k}(1-\beta_j)\right) &\leq \exp\left(-\sum_{j=i}^{k}\beta_j\right) = \begin{cases} \exp\left(-\sum_{j=i}^{k}\frac{c}{j+1}\right) & \text{if } i \geq \underline{k}, \\ \exp\left(-\sum_{j=i}^{\underline{k}-1}\frac{1}{2} - \sum_{j=\underline{k}}^{k}\frac{c}{j+1}\right) & \text{if } i < \underline{k}, \end{cases} \\
&= \exp\left(-\frac{\underline{k} - \min\{\underline{k}, i\}}{2} - \sum_{j=\max\{\underline{k}, i\}}^{k}\frac{c}{j+1}\right) \\
&\leq \exp\left(-\frac{\underline{k} - \min\{\underline{k}, i\}}{2} - \int_{x=\max\{\underline{k}, i\}}^{k+1}\frac{c}{x+1}\mathrm{d}x\right) \\
&= \exp\left(-\frac{\underline{k} - \min\{\underline{k}, i\}}{2}\right)\left(\frac{\max\{\underline{k}, i\}+1}{k+2}\right)^c,
\end{aligned}$$

where the second inequality follows from $\int_a^{b+1}\frac{1}{x}dx < \sum_{j=a}^{b}\frac{1}{i}$ for any $0 < a \leq b$. This completes the proof. □

The next lemma establishes, for all k, a relationship between the stochastic gradient error $\epsilon_k = d_k - \nabla f(x_k)$ and a martingale. This is useful for an Azuma-Hoeffding-type inequality that will be used to prove a variance reduction property.

Lemma A.2. *For all $k \geq 2$, with the convention that $\prod_{i=l}^{u} a_i = 1$ if $l > u$, consider $\{e_{ki}\}_{i=0}^{k}$ with*

$$e_{ki} := \begin{cases} 0 & i = 0, \\ \left(\prod_{j=2}^{k}(1-\beta_j)\right) A_1 & i = 1, \\ \left(\prod_{j=i+1}^{k}(1-\beta_j)\right) A_i + \left(\prod_{j=i}^{k}(1-\beta_j)\right) B_i & 2 \leq i \leq k, \end{cases}$$

where $A_i := v_i - \nabla f(x_i)$ and $B_i := \nabla f(x_{i-1}) - u_i$ for all $i \geq 1$ with v_i and u_i defined as in Algorithm 1.

1. *Consider $\{S_{kt}\}_{t=0}^{\infty}$ with $S_{kt} := \sum_{i=0}^{t} e_{ki}$ for all $0 \leq t \leq k$ and $S_{kt} = S_{kk}$ for all $t > k$. Under Assumption 3.2(1), $\{S_{kt}\}_{t=0}^{\infty}$ forms a martingale with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Specifically, with $\mathcal{F}_0 = \mathcal{F}_1 = \sigma(x_1)$ and $\mathcal{F}_t$ is the σ-algebra generated by $\{\{\Xi_{1,i}\}_{i=1}^{m}, \ldots, \{\Xi_{(t-1),i}\}_{i=1}^{m}\}$ (of which $\{\{\xi_{1,i}\}_{i=1}^{m}, \cdots, \{\xi_{(t-1),i}\}_{i=1}^{m}\}$ is a realization) for all $t \in \{2, \cdots, k\}$, and $\mathcal{F}_t = \mathcal{F}_k$ for all $t > k$.*
2. *With $\{S_{kt}\}_{t=0}^{\infty}$ defined as in part 1, one has that $S_{kk} = \epsilon_k$.*
3. *Under Assumption 3.2 and Assumption 3.3 and with $\underline{k} = \lceil (2c) - 1 \rceil$, it holds almost surely that*

$$\|e_{ki}\| \leq \begin{cases} \sigma \exp\left(-\frac{\underline{k}-2}{2}\right)\left(\frac{\underline{k}+1}{k+2}\right)^c & \text{if } i = 1, \\ (2\sigma + 2L_g(G_r + G_d)\zeta\underline{\alpha})\frac{1}{2}\exp\left(-\frac{\underline{k}-i}{2}\right)\left(\frac{\underline{k}+1}{k+2}\right)^c & \text{if } 2 \leq i \leq \underline{k}, \\ (2\sigma + 2L_g(G_r + G_d)\zeta\underline{\alpha})\frac{c}{i}\left(\frac{i+1}{k+2}\right)^c & \text{if } \underline{k}+1 \leq i \leq k. \end{cases}$$

Proof. Consider part 1. We have $S_{k0} = e_{k0} = 0$, and for all $1 \le t \le k$, one finds $S_{kt} - S_{k(t-1)} = e_{kt}$, so that

$$\mathbb{E}_{\xi\sim\mathcal{P}}\left[S_{kt}|\mathcal{F}_t\right] = \mathbb{E}_{\xi\sim\mathcal{P}}\left[S_{k(t-1)} + e_{kt}|\mathcal{F}_t\right] = S_{k(t-1)} + \mathbb{E}_{\xi\sim\mathcal{P}}\left[e_{kt}|\mathcal{F}_t\right]. \tag{12}$$

Assumption 3.2(1) implies that $\mathbb{E}_{\xi\sim\mathcal{P}}\left[e_{kt}|\mathcal{F}_t\right] = 0$, which may then be combined with (12) to conclude that $\mathbb{E}_{\xi\sim\mathcal{P}}\left[S_{kt}|\mathcal{F}_t\right] = S_{k(t-1)}$ for all $1 \le t \le k$. On the other hand, for all $t > k$, we trivially have $\mathbb{E}_{\xi\sim\mathcal{P}}\left[S_{kt}|\mathcal{F}_t\right] = \mathbb{E}_{\xi\sim\mathcal{P}}\left[S_{k(t-1)}|\mathcal{F}_t\right] = S_{k(t-1)}$. Therefore, $\{S_{kt}\}_{t=0}^{\infty}$ forms a martingale.

Consider part 2. For all $k \ge 2$, one finds that

$$\begin{aligned}
\epsilon_k &= d_k - \nabla f(x_k)\\
&= (1-\beta_k)\epsilon_{k-1} + A_k + (1-\beta_k)B_k\\
&= (1-\beta_k)(1-\beta_{k-1})\epsilon_{k-2} + (1-\beta_k)A_{k-1} + A_k + (1-\beta_k)(1-\beta_{k-1})B_{k-1} + (1-\beta_k)B_k\\
&= \left(\prod_{j=2}^{k}(1-\beta_j)\right)\epsilon_1 + \sum_{i=2}^{k}\left(\prod_{j=i+1}^{k}(1-\beta_j)\right)A_i + \sum_{i=2}^{k}\left(\prod_{j=i}^{k}(1-\beta_j)\right)B_i.
\end{aligned}$$

Since $\epsilon_1 = A_1$, the desired conclusion follows that $\epsilon_k = \sum_{i=0}^{k} e_{ki} = S_{kk}$.

We now prove part 3. Consider the following two cases:

Case I: For $i = 1$, it follows from the triangular inequality and Assumption 3.2(3) that, almost surely, one finds

$$\begin{aligned}
\|e_{k1}\| = \left\|\left(\prod_{j=2}^{k}(1-\beta_j)\right)\epsilon_1\right\| &\le \left(\prod_{j=2}^{k}(1-\beta_j)\right)\|\epsilon_1\|\\
&= \left(\prod_{j=2}^{k}(1-\beta_j)\right)\left\|\frac{1}{m}\sum_{i'=1}^{m}\nabla\ell(x_1;\xi_{1i'}) - \nabla f(x_1)\right\|\\
&\le \sigma\left(\prod_{j=2}^{k}(1-\beta_j)\right).
\end{aligned}$$

It follows from Lemma A.1 that, almost surely, one finds

$$\|e_{k1}\| \le \sigma\exp\left(-\frac{k-2}{2}\right)\left(\frac{k+1}{k+2}\right)^c.$$

Case II: For any i with $2 \le i \le k$, it follows almost surely that

$$\begin{aligned}
&\|e_{ki}\| \qquad (13)\\
&= \left\|\left(\prod_{j=i+1}^{k}(1-\beta_j)\right)A_i + \left(\prod_{j=i}^{k}(1-\beta_j)\right)B_i\right\|\\
&= \left\|\left(\prod_{j=i+1}^{k}(1-\beta_j)\right)(1-\beta_i+\beta_i)A_i + \left(\prod_{j=i}^{k}(1-\beta_j)\right)B_i\right\|\\
&= \left\|\beta_i\left(\prod_{j=i+1}^{k}(1-\beta_j)\right)A_i + \left(\prod_{j=i}^{k}(1-\beta_j)\right)(A_i+B_i)\right\|\\
&\le \sigma\beta_i\left(\prod_{j=i+1}^{k}(1-\beta_j)\right) + \left(\prod_{j=i}^{k}(1-\beta_j)\right)\left\|\frac{1}{m}\sum_{i'=1}^{m}\nabla\ell(x_i;\xi_{ii'}) - \frac{1}{m}\sum_{i'=1}^{m}\nabla\ell(x_{i-1};\xi_{ii'}) - (\nabla f(x_i) - \nabla f(x_{i-1}))\right\|\\
&\le \sigma\beta_i\left(\prod_{j=i+1}^{k}(1-\beta_j)\right) + 2L_g\left(\prod_{j=i}^{k}(1-\beta_j)\right)\|x_i - x_{i-1}\|, \qquad (14)
\end{aligned}$$

where the first inequality holds by Assumption 3.2(3) and the second inequality holds by Assumption 3.1(2). Since $y_{i-1} = \text{prox}_{\alpha_{i-1}r}(x_{i-1} - \alpha_{i-1}d_{i-1})$, it follows from Beck (2017, Theorem 6.39) that $\frac{x_{i-1}-y_{i-1}}{\alpha_{i-1}} - d_{i-1} \in \partial r(y_{i-1})$. Hence, it follows from Assumption 3.2(2) that $\left\|\frac{x_{i-1}-y_{i-1}}{\alpha_{i-1}} - d_{i-1}\right\| \leq G_r$. It follows from line 12 of Algorithm 1, Assumption 3.2(4), the triangular inequality, and the previous inequality that

$$\begin{aligned}\|x_i - x_{i-1}\| &= \zeta\beta_{i-1}\|x_{i-1} - y_{i-1}\| \\ &\leq \zeta\beta_{i-1}(\|x_{i-1} - y_{i-1} - \alpha_{i-1}d_{i-1}\| + \alpha_{i-1}\|d_{i-1}\|) \\ &\leq \zeta\beta_{i-1}\alpha_{i-1}(G_r + G_d).\end{aligned}$$

Combining (14) and the above inequality, one finds almost surely that

$$\|e_{ki}\| \leq \sigma\beta_i\left(\prod_{j=i+1}^{k}(1-\beta_j)\right) + 2L_g(G_r+G_d)\zeta\beta_{i-1}\alpha_{i-1}\left(\prod_{j=i}^{k}(1-\beta_j)\right). \tag{15}$$

It follows from Assumption 3.3 that $\beta_k = \min\{\frac{1}{2}, \frac{c}{(k+1)}\}$ for all $k \geq 2$. Therefore, since $2(1-\beta_i) \geq 1$, one finds

$$\beta_i\left(\prod_{j=i+1}^{k}(1-\beta_j)\right) \leq 2\beta_i\left(\prod_{j=i}^{k}(1-\beta_j)\right) \leq 2\beta_{i-1}\left(\prod_{j=i}^{k}(1-\beta_j)\right), \tag{16}$$

It follows from (15), (16), and $\alpha_i \equiv \underline{\alpha}$ that almost surely one finds

$$\|e_{ki}\| \leq \big(2\sigma + 2L_g(G_r+G_d)\zeta\underline{\alpha}\big)\beta_{i-1}\left(\prod_{j=i}^{k}(1-\beta_j)\right).$$

Applying Lemma A.1 to the above inequality, one finds almost surely that

$$\|e_{ki}\| \leq \begin{cases} \big(2\sigma + 2L_g(G_r+G_d)\zeta\underline{\alpha}\big)\frac{1}{2}\exp\left(-\frac{\underline{k}-i}{2}\right)\left(\frac{\underline{k}+1}{k+2}\right)^c & \text{if } 2 \leq i \leq \underline{k}, \\ \big(2\sigma + 2L_g(G_r+G_d)\zeta\underline{\alpha}\big)\frac{c}{i}\left(\frac{i+1}{k+2}\right)^c & \text{if } \underline{k}+1 \leq i \leq k. \end{cases}$$

Combining the two cases above give the results claimed in part 3. ☐

The last lemma bounds $\sum_{i=1}^{k}\|e_{ki}\|^2$, which will appear in the Azuma-Hoeffding type inequality.

Lemma A.3. *Under Assumption 3.2 and Assumption 3.3, there exits a constant $C_1 > 0$ that is independent of k such that, for all $k \geq \underline{k} = \lceil(2c) - 1\rceil$, one finds*

$$\sum_{i=1}^{k}\|e_{ki}\|^2 \leq C_1\big(\sigma + L_g(G_r+G_d)\zeta\underline{\alpha}\big)^2\max\left\{\left(\frac{\underline{k}+1}{k+2}\right)^{2c}, \frac{c^2}{k+2}\right\} \textit{ almost surely.}$$

Proof. It follows from Lemma A.2(3) that, almost surely,

$$\begin{aligned}\sum_{i=1}^{k}\|e_{ki}\|^2 &= \|e_{k1}\|^2 + \sum_{i=2}^{\underline{k}}\|e_{ki}\|^2 + \sum_{i=\underline{k}+1}^{k}\|e_{ki}\|^2 \\ &\leq \sigma^2\exp\left(-(\underline{k}-2)\right)\left(\frac{\underline{k}+1}{k+2}\right)^{2c} + \sum_{i=2}^{\underline{k}}\big(2\sigma + 2L_g(G_r+G_d)\zeta\underline{\alpha}\big)^2\frac{1}{4}\exp\left(-(\underline{k}-i)\right)\left(\frac{\underline{k}+1}{k+2}\right)^{2c} \\ &\quad + \sum_{i=\underline{k}+1}^{k}\big(2\sigma + 2L_g(G_r+G_d)\zeta\underline{\alpha}\big)^2\frac{c^2}{i^2}\left(\frac{i+1}{k+2}\right)^{2c}.\end{aligned} \tag{17}$$

With respect to each of three terms above, for some C_{11} that is independent of k, one finds

$$\sigma^2 \exp\left(-(\underline{k}-2)\right) \cdot \left(\frac{\underline{k}+1}{k+2}\right)^{2c} = \sigma^2 e^2 \exp(-\underline{k}) \cdot \left(\frac{\underline{k}+1}{k+2}\right)^{2c} \tag{18}$$

$$\sum_{i=2}^{\underline{k}} \left(2\sigma + 2L_g(G_r+G_d)\zeta\underline{\alpha}\right)^2 \frac{1}{4} \exp\left(-(\underline{k}-i)\right) \cdot \left(\frac{\underline{k}+1}{k+2}\right)^{2c} \leq \left(\sigma + L_g(G_r+G_d)\zeta\underline{\alpha}\right)^2 \frac{e}{e-1} \cdot \left(\frac{\underline{k}+1}{k+2}\right)^{2c} \tag{19}$$

$$\sum_{i=\underline{k}+1}^{k} \left(2\sigma + 2L_g(G_r+G_d)\zeta\underline{\alpha}\right)^2 \frac{c^2}{i^2} \cdot \left(\frac{i+1}{k+2}\right)^{2c} \leq \frac{\left(2\sigma + 2L_g(G_r+G_d)\zeta\underline{\alpha}\right)^2 c^2 C_{11}}{k+2}, \tag{20}$$

where (19) holds since the geometric series $\sum_{i=2}^{\underline{k}} \exp\left(-(\underline{k}-i)\right) = \sum_{i=2}^{\underline{k}} \frac{\exp(i)}{\exp(\underline{k})} = \frac{e-e^{2-\underline{k}}}{e-1} \leq \frac{e}{e-1}$ and (20) hold since

$$\begin{aligned}
\sum_{i=1}^{k} \frac{(i+1)^{2c}}{i^2} &= \sum_{i=1}^{1} \frac{(i+1)^{2c}}{i^2} + \sum_{i=2}^{k} \frac{(i+1)^{2c}}{i^2} \\
&\leq 4^c + \sum_{i=2}^{k} \frac{(1.5i)^{2c}}{i^2} \\
&\leq 4^c + (1.5)^{2c} \int_{i=2}^{k+1} i^{2c-2} \mathrm{d}i \\
&= 4^c + (1.5)^{2c} \left(\frac{(k+1)^{2c-1}}{2c-1} - \frac{2^{2c-1}}{2c-1}\right) \leq C_{11}(k+1)^{2c-1} \leq C_{11}(k+2)^{2c-1}.
\end{aligned} \tag{21}$$

Combining (17)-(20), one finds almost surely that

$$\begin{aligned}
\sum_{i=1}^{k} \|e_{ki}\|^2 &\leq \left(\sigma + L_g(G_r+G_d)\zeta\underline{\alpha}\right)^2 \left(C_{12}\left(\frac{\underline{k}+1}{k+2}\right)^{2c} + \left(4C_{11} + \frac{e}{e-1}\right)\frac{c^2}{k+2}\right) \\
&\leq \left(\sigma + L_g(G_r+G_d)\zeta\underline{\alpha}\right)^2 \left(C_{12} + 4C_{11} + \frac{e}{e-1}\right) \max\left\{\left(\frac{\underline{k}+1}{k+2}\right)^{2c}, \frac{c^2}{k+2}\right\},
\end{aligned}$$

where we use the fact that $\sigma^2 e^2 \exp(-\underline{k}) \leq C_{12}\left(\sigma + L_g(G_r+G_d)\zeta\underline{\alpha}\right)^2$ for some $C_{12} > 0$ that is independent of k. We complete the proof by setting $C_1 = \left(C_{12} + 4C_{11} + \frac{e}{e-1}\right)$. $\square$

Now, we are ready to formally prove Theorem 3.1.

Theorem 3.1. Let Assumption 3.1–Assumption 3.3 hold, let $\epsilon_k = d_k - \nabla f(x_k)$ for all $k \in \mathbb{N}_+$, and define $\underline{k} = \lceil (2c)-1 \rceil$. Then, for any $k \geq \underline{k}$ and any $\eta_k \in (0,1)$, the event $\mathcal{E}_k := \{\|\epsilon_k\| \leq U(k)\}$ holds with probability at least $1-\eta_k$, where for some constant $C \in \mathbb{R}_+$ independent of k, one defines

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

(Specifically, the constant is $C = \sqrt{2C_1}$, where C_1 is defined in Lemma A.3.)

Proof. It follows from Lemma A.3 that almost surely one finds

$$\sum_{i=1}^{k} \|e_{ki}\|^2 \leq C_1\left(\sigma + L_g(G_r+G_d)\zeta\underline{\alpha}\right)^2 \max\left\{\left(\frac{\underline{k}+1}{k+2}\right)^{2c}, \frac{c^2}{k+2}\right\} =: h(k).$$

Based on Lemma A.2(1), we have for $k \geq \underline{k}$ that $\{S_{kt}\}_{t=0}^{k}$ forms a martingale with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{k}$. For any $\rho_k > 0$, using the Azuma-Hoeffding type inequality (Pinelis, 1994, Theorem 3.5)[9] on the martingale $\{S_{kt}\}_{t=0}^{k}$,

[9] See Remark A.1 for details on applying this theorem.

together with $\|e_{ki}\|_\infty \le \|e_{ki}\|$ (e_{ki} is defined in Lemma A.2) and the fact that $S_{kk} = \epsilon_k$ (Lemma A.2(2)), we have

$$\mathbb{P}\left[\|\epsilon_k\| \ge \rho_k\right] = \mathbb{P}\left[\|S_{kk}\| \ge \rho_k\right] \le \mathbb{P}\left[\sup_{t\in[k]} \|S_{kt}\| \ge \rho_k\right] \le 2\exp\left(-\frac{\rho_k^2}{2h(k)}\right). \tag{22}$$

For any $\eta_k \in (0,1)$, by setting $\rho_k = U(k) = \sqrt{2h(k)\log(2/\eta_k)}$ in (22), we have $\mathbb{P}\big[\,\|\epsilon_k\| \ge U(k)\big] \le \eta_k$, which implies that the event $\mathcal{E}_k = \{\|\epsilon_k\| \le U(k)\}$ holds with probability at least $1-\eta_k$. This completes the proof. □

Remark A.1. We define the f used in (Pinelis, 1994, Theorem 3.5) when cited in the proof of Theorem 3.1 above as $f = \{S_{k0}, S_{k1}, \cdots, S_{kk}, S_{kk}, \dots\}$ with $f_j = S_{kj}$ for all $1 \le j \le k$ and $f_j = S_{kk}$ for all $j > k$. As proved in Lemma A.2(1), f is a martingale. Consequently, the d_j and f^* appearing in (Pinelis, 1994, Theorem 3.5) are defined as $d_j = S_{kj} - S_{k(j-1)} = e_{kj}$ and $f^* = \sup_{j\in[k]}\{\|f_j\|\} = \sup_{j\in[k]}\{\|S_{kj}\|\}$, respectively. As proved in Lemma A.3, we have $\sum_{j=1}^k \|d_j\|_\infty^2 \le \sum_{j=1}^k \|d_j\|_2^2 = \sum_{j=1}^k \|e_{kj}\|_2^2 \le h(k)$ almost surely.

A.2 Proof of Corollary 3.1

Corollary 3.1 Let $\eta_k = \frac{\eta_0}{k^2}$ for all $k \ge 1$ with $\eta_0 \in (0, 6/\pi^2)$. Define the event $\mathcal{E}_k := \{\|\epsilon_k\| \le U(k)\}$ and recall that $\underline{k} = \lceil (2c) - 1\rceil$. Under Assumption 3.1–Assumption 3.3, the event $\mathcal{E} := \bigcap_{k\ge\underline{k}}^\infty \mathcal{E}_k$ holds with probability at least $1 - \frac{\eta_0\pi^2}{6}$.

Proof. It follows from the stated conditions, the union bound from probability, and Theorem 3.1 that

$$\begin{aligned}
\mathbb{P}\left[\bigcap_{k=\underline{k}}^\infty \{\|\epsilon_k\| \le U(k)\}\right] &= \mathbb{P}\left[\bigcap_{k=\underline{k}}^\infty \mathcal{E}_k\right] = 1 - \mathbb{P}\left[\left(\bigcap_{k=\underline{k}}^\infty \mathcal{E}_k\right)^c\right] \qquad \text{(here } c \text{ is the set complement operator)}\\
&= 1 - \mathbb{P}\left[\bigcup_{k=\underline{k}}^\infty \mathcal{E}_k^c\right] \ge 1 - \sum_{k\ge\underline{k}}^\infty \mathbb{P}\left[\mathcal{E}_k^c\right] = 1 - \sum_{k\ge\underline{k}}^\infty \mathbb{P}\left[\|\epsilon_k\| > U(k)\right]\\
&\ge 1 - \sum_{k\ge\underline{k}}^\infty \eta_k \ge 1 - \sum_{k=1}^\infty \frac{\eta_0}{k^2} = 1 - \frac{\eta_0\pi^2}{6},
\end{aligned}$$

where the last equality holds by the Basel equality $\sum_{k=1}^\infty \frac{1}{k^2} = \frac{\pi^2}{6}$. □

A.3 Proof of Theorem 3.2

Theorem 3.2. Let $\underline{\alpha} = \mu_f/L_g^2$, $\zeta \in (0,2)$, $\theta \ge 2$, $c = (2\theta L_g^2)/(\zeta\mu_f^2) > 2$, and $\underline{k} = \lceil 2c - 1\rceil$. Set $\eta_k = \eta_0/k^2$ for all $k \ge 1$ with $\eta_0 \in (0, 6/\pi^2)$. Then, under Assumption 3.1–Assumption 3.3, there exists a constant $C_3 \in \mathbb{R}_+$ independent of k, such that the event $\mathcal{E}_k^x := \left\{\|x_k - x^*\|^2 \le \bar{c}_1 \frac{\|x_{\underline{k}} - x^*\|^2}{k^\theta} + \bar{c}_2 \cdot \frac{\log\frac{2k}{\eta_0}}{k}\right\}$ with $\bar{c}_1 := (\underline{k}+2)^\theta$ and $\bar{c}_2 := C_3\zeta\left(\frac{\mu_f^2}{L_g^4} + \frac{2}{L_g^2}\left(1 + \frac{\mu_f}{L_g}\right)^2\right)(\sigma + L_g(G_r + G_d)\zeta\underline{\alpha})^2$ satisfies

$$\mathbb{P}\left[\bigcap_{k\ge\underline{k}}^\infty \mathcal{E}_k^x\right] \ge 1 - \eta_0\pi^2/6 > 0\,.$$

Proof. Since the proximal operator is non-expansive (Beck, 2017, Theorem 6.42) and $x^* = \operatorname{prox}_{\alpha_k r}(x^* - \alpha_k\nabla f(x^*))$, it follows that

$$\begin{aligned}
\|y_k - x^*\|^2 &= \left\|\operatorname{prox}_{\alpha_k r}(x_k - \alpha_k d_k) - \operatorname{prox}_{\alpha_k r}(x^* - \alpha_k\nabla f(x^*))\right\|^2\\
&\le \|x_k - x^* - \alpha_k(d_k - \nabla f(x^*))\|^2\\
&= \|x_k - x^*\|^2 - 2\alpha_k(x_k - x^*)^T(d_k - \nabla f(x^*)) + \alpha_k^2\|d_k - \nabla f(x^*)\|^2\\
&= \|x_k - x^*\|^2 - 2\alpha_k(x_k - x^*)^T(\epsilon_k + \nabla f(x_k) - \nabla f(x^*)) + \alpha_k^2\|d_k - \nabla f(x^*)\|^2\,.
\end{aligned} \tag{23}$$

It follows from Assumption 3.1 that f is μ_f-strongly convex, and therefore

$$(x_k - x^*)^T(\nabla f(x_k) - \nabla f(x^*)) \geq \mu_f \|x_k - x^*\|^2 . \tag{24}$$

It follows from (23) that

$$\begin{aligned}
&\|y_k - x^*\|^2 \qquad (25)\\
&\leq \|x_k - x^*\|^2 - 2\alpha_k(x_k - x^*)^T(\nabla f(x_k) - \nabla f(x^*)) - 2\alpha_k(x_k - x^*)^T\epsilon_k + \alpha_k^2 \|d_k - \nabla f(x^*)\|^2\\
&\overset{(i)}{\leq} (1 - 2\mu_f\alpha_k)\|x_k - x^*\|^2 - 2\alpha_k(x_k - x^*)^T\epsilon_k + \alpha_k^2\|d_k - \nabla f(x^*)\|^2\\
&= (1 - 2\mu_f\alpha_k)\|x_k - x^*\|^2 + 2\alpha_k(x^* - x_k)^T\epsilon_k + \alpha_k^2\|d_k - \nabla f(x_k) + \nabla f(x_k) - \nabla f(x^*)\|^2\\
&\overset{(ii)}{=} (1 - 2\mu_f\alpha_k)\|x_k - x^*\|^2 + 2\alpha_k(x^* - x_k)^T\epsilon_k + \alpha_k^2\left(\|\epsilon_k\|^2 + 2\epsilon_k^T(\nabla f(x_k) - \nabla f(x^*)) + \|\nabla f(x_k) - \nabla f(x^*)\|^2\right)\\
&\overset{(iii)}{\leq} (1 - 2\mu_f\alpha_k)\|x_k - x^*\|^2 + 2\alpha_k\|x_k - x^*\|\|\epsilon_k\| + \alpha_k^2\left(\|\epsilon_k\|^2 + 2L_g\|\epsilon_k\|\|x_k - x^*\| + L_g^2\|x_k - x^*\|^2\right)\\
&= (1 - 2\mu_f\alpha_k + \alpha_k^2L_g^2)\|x_k - x^*\|^2 + (2\alpha_k + 2L_g\alpha_k^2)\|x_k - x^*\|\|\epsilon_k\| + \alpha_k^2\|\epsilon_k\|^2 , \qquad (26)
\end{aligned}$$

where (i) follows from (24), (ii) follows from the definition of ϵ_k, and (iii) follows from Assumption 3.1 and the Cauchy-Schwarz inequality. When the event $\mathcal{E}_k = \{\|\epsilon_k\| \leq U(k)\}$ happens ($U(k)$ defined in Theorem 3.1), it follows from line 12 in Algorithm 1, $\zeta\beta_k < 1$, (26), and Theorem 3.1 that

$$\begin{aligned}
&\|x_{k+1} - x^*\|^2\\
&= \|\zeta\beta_k(y_k - x^*) + (1 - \zeta\beta_k)(x_k - x^*)\|^2\\
&\leq \zeta\beta_k\|y_k - x^*\|^2 + (1 - \zeta\beta_k)\|x_k - x^*\|^2\\
&\leq \zeta\beta_k\left((1 - 2\mu_f\alpha_k + \alpha_k^2L_g^2)\|x_k - x^*\|^2 + (2\alpha_k + 2L_g\alpha_k^2)\|x_k - x^*\|\|\epsilon_k\| + \alpha_k^2\|\epsilon_k\|^2\right) + (1 - \zeta\beta_k)\|x_k - x^*\|^2\\
&\leq (1 - \zeta\beta_k(2\mu_f\alpha_k - \alpha_k^2L_g^2))\|x_k - x^*\|^2 + (2\alpha_k + 2L_g\alpha_k^2)\|x_k - x^*\|\zeta\beta_kU(k) + \alpha_k^2\zeta\beta_kU(k)^2; \qquad (27)
\end{aligned}$$

we emphasize that the first inequality above follows from the convexity of the 2-norm-squared. Therefore, (27) holds with probability at least $1 - \eta_k$ since the event $\mathcal{E}_k = \{\|\epsilon_k\| \leq U(k)\}$ happens with probability at least $1 - \eta_k$.

Define $s_k^2 = \|x_k - x^*\|^2$ and since $\alpha_k \equiv \underline{\alpha} = \mu_f/L_g^2$, then (27) becomes

$$\begin{aligned}
s_{k+1}^2 &\leq \left(1 - \zeta\beta_k\frac{\mu_f^2}{L_g^2}\right)s_k^2 + \frac{2\mu_f}{L_g^2}\left(1 + \frac{\mu_f}{L_g}\right)\zeta\beta_kU(k)s_k + \frac{\mu_f^2}{L_g^4}\zeta\beta_kU(k)^2.\\
&= (1 - c_0\zeta\beta_k)s_k^2 + c_1\zeta\beta_kU(k)s_k + c_2\zeta\beta_kU(k)^2, \qquad (28)
\end{aligned}$$

with $c_0 = \frac{\mu_f^2}{L_g^2}, c_1 = \frac{2\mu_f}{L_g^2}\left(1 + \frac{\mu_f}{L_g}\right)$, and $c_2 = \frac{\mu_f^2}{L_g^4}$. The second term in the above inequality can be upper bounded as

$$c_1\zeta\beta_kU(k)s_k = 2\left(\frac{c_1\sqrt{\rho\zeta\beta_k}}{2}s_k\right)\left(\frac{\sqrt{\zeta\beta_k}}{\sqrt{\rho}}U(k)\right) \leq \frac{\rho\zeta\beta_kc_1^2}{4}s_k^2 + \frac{\zeta\beta_k}{\rho}U(k)^2 \quad \text{for all } \rho > 0,$$

by using Young's inequality. Combining this result with (28), one obtains

$$s_{k+1}^2 \leq \left[1 - \left(c_0\zeta - \rho\zeta\frac{c_1^2}{4}\right)\beta_k\right]s_k^2 + \left[c_2\zeta + \frac{\zeta}{\rho}\right]\beta_kU(k)^2.$$

Now setting $\rho = \frac{2\mu_f^2}{L_g^2c_1^2}$, it follows from this inequality that

$$s_{k+1}^2 \leq \left[1 - \frac{\zeta\mu_f^2}{2L_g^2}\beta_k\right]s_k^2 + \left[\frac{\zeta\mu_f^2}{L_g^4} + \frac{\zeta c_1^2L_g^2}{2\mu_f^2}\right]\beta_kU(k)^2 = (1 - \gamma_k)s_k^2 + c_3\beta_kU(k)^2, \tag{29}$$

where $\gamma_k = \frac{\zeta\mu_f^2}{2L_g^2}\beta_k$ and $c_3 = \frac{\zeta\mu_f^2}{L_g^4} + \frac{\zeta c_1^2L_g^2}{2\mu_f^2}$.

Conditioning on the event $\mathcal{E} = \bigcap_{i\geq \underline{k}}^{\infty} \mathcal{E}_i$ happens, it follows from (29), for all $k \geq \underline{k}$, that

$$\begin{aligned} s_{k+1}^2 &\leq (1-\gamma_k)s_k^2 + c_3\beta_k U(k)^2 \\ &\leq (1-\gamma_k)(1-\gamma_{k-1})s_{k-1}^2 + c_3 \sum_{i=k-1}^{k} \left(\prod_{j=i+1}^{k} (1-\gamma_j)\beta_i U(i)^2 \right) \\ &\leq \text{(expanding recursively on } s_{k-1}^2) \\ &\leq \left[\prod_{i=\underline{k}}^{k} (1-\gamma_i) \right] \cdot s_{\underline{k}}^2 + c_3 \cdot \sum_{i=\underline{k}}^{k} \left[\prod_{j=i+1}^{k} (1-\gamma_j) \right] \beta_i U(i)^2, \end{aligned} \tag{30}$$

where we use the convention that $\prod_{i=l}^{u} a_i = 1$ if $l > u$ for any $a_i \in \mathbb{R}$ and $(l,u) \in \mathbb{Z}_+^2$. Then using a similar argument as from Lemma A.1, one can establish, for any $i \geq 2$, that

$$\begin{aligned} \prod_{j=i}^{k}(1-\gamma_j) \leq \exp\left(-\sum_{j=i}^{k}\gamma_j\right) &= \exp\left(-\frac{\zeta\mu_f^2}{2L_g^2}\cdot\sum_{j=i}^{k}\min\left\{\frac{1}{2}, \frac{c}{j+1}\right\}\right) \\ &= \exp\left(-\frac{\zeta\mu_f^2}{2L_g^2}\cdot\frac{\underline{k}-\min\{\underline{k},i\}}{2} - \frac{\zeta\mu_f^2}{2L_g^2}\cdot\sum_{j=\max\{\underline{k},i\}}^{k}\frac{c}{j+1}\right) \\ &\leq \exp\left(-\frac{\zeta\mu_f^2}{2L_g^2}\cdot\frac{\underline{k}-\min\{\underline{k},i\}}{2}\right)\cdot\left(\frac{\max\{\underline{k},i\}+2}{k+1}\right)^{\zeta\mu_f^2 c/(2L_g^2)} \\ &= \exp\left(-\frac{\zeta\mu_f^2}{2L_g^2}\cdot\frac{\underline{k}-\min\{\underline{k},i\}}{2}\right)\cdot\left(\frac{\max\{\underline{k},i\}+2}{k+1}\right)^{\theta}. \end{aligned}$$

Combing the above inequality with (30) we obtain, for any $k \geq \underline{k}$, that

$$s_{k+1}^2 \leq \left(\frac{\underline{k}+2}{k+1}\right)\cdot s_{\underline{k}}^2 + c_3\cdot\sum_{i=\underline{k}}^{k}\left[\left(\frac{(i+1)+2}{k+1}\right)^{\theta}\right]\frac{c}{i+1}U(i)^2. \tag{31}$$

It follows from Theorem 3.1 that

$$U(i)^2 = \begin{cases} G^2\left(\frac{\underline{k}+1}{i+2}\right)^{2c}\log\frac{2}{\eta_i} & \text{if } i < \bar{k}, \\ G^2\frac{c^2}{i+2}\log\frac{2}{\eta_i} & \text{if } i \geq \bar{k}, \end{cases} \tag{32}$$

where $G = C(\sigma + L_g(G_r + G_d)\zeta\underline{\alpha})$ and $\bar{k} = \max\left\{\underline{k}, \left\lceil\frac{(\underline{k}+1)^{2c/(2c-1)}}{c^{2/(2c-1)}} - 2\right\rceil\right\}$. Then it follows from (32) that

$$\begin{aligned} & c_3\cdot\sum_{i=\underline{k}}^{k}\left[\left(\frac{(i+1)+2}{k+1}\right)^{\theta}\right]\frac{c}{i+1}U(i)^2 \\ \leq\ & \frac{c_3\cdot c\cdot G^2}{(k+1)^{\theta}}\left[\sum_{i=\underline{k}}^{\min\{\bar{k}-1,k\}}\frac{(i+3)^{\theta}}{i+1}\frac{(\underline{k}+1)^{2c}}{(i+2)^{2c}}\log\frac{2}{\eta_i} + \sum_{i=\min\{\bar{k}-1,k\}+1}^{k}\frac{(i+3)^{\theta}}{i+1}\frac{c^2}{i+2}\log\frac{2}{\eta_i}\right], \end{aligned} \tag{33}$$

where we use the convention that $\sum_{i=l}^{u} a_i = 0$ if $l > u$ for any $a_i \in \mathbb{R}$ and $(l,u) \in \mathbb{Z}_+^2$.

It follows from (33) and $\theta \geq 2$, there exists constants $\{C_{30}, C_{31}, C_{32}, C_{34}\} \subset \mathbb{R}_+$, which are independent of k, such that, for all $k \geq \underline{k}$, one obtains,

$$\begin{aligned}
& c_3 \cdot \sum_{i=\underline{k}}^{k} \left[\left(\frac{(i+1)+2}{k+1} \right)^{\theta} \right] \frac{c}{i+1} U(i)^2 \\
&= \frac{c_3 \cdot c \cdot G^2}{(k+1)^{\theta}} \left[\sum_{i=\underline{k}}^{\min\{\bar{k}-1,k\}} \frac{(i+3)^{\theta}}{i+1} \frac{(\underline{k}+1)^{2c}}{(i+2)^{2c}} \log \frac{2}{\eta_i} + \sum_{i=\min\{\bar{k}-1,k\}+1}^{k} \frac{(i+3)^{\theta}}{i+1} \frac{c^2}{i+2} \log \frac{2}{\eta_i} \right] \\
&= \begin{cases} \frac{c_3 \cdot c \cdot G^2}{(k+1)^{\theta}} \left[\sum_{i=\underline{k}}^{k} \frac{(i+3)^{\theta}}{i+1} \frac{(\underline{k}+1)^{2c}}{(i+2)^{2c}} \log \frac{2}{\eta_i} \right] & \text{if } \underline{k} \leq k < \bar{k}, \\ \frac{c_3 \cdot c \cdot G^2}{(k+1)^{\theta}} \left[\sum_{i=\underline{k}}^{\bar{k}-1} \frac{(i+3)^{\theta}}{i+1} \frac{(\underline{k}+1)^{2c}}{(i+2)^{2c}} \log \frac{2}{\eta_i} + \sum_{i=\bar{k}}^{k} \frac{(i+3)^{\theta}}{i+1} \frac{c^2}{i+2} \log \frac{2}{\eta_i} \right] & \text{if } \underline{k} \leq \bar{k} \leq k, \end{cases} \\
&\leq \begin{cases} \frac{c_3 \cdot c \cdot G^2 (\underline{k}+1)^{2c} \log \frac{2}{\eta_k}}{(k+1)^{\theta}} \left[\sum_{i=\underline{k}}^{k} \frac{(i+3)^{\theta}}{i+1} \frac{1}{(i+2)^{2c}} \right] & \text{if } \underline{k} \leq k < \bar{k}, \quad \text{(due to } \eta_i \geq \eta_k) \\ \frac{c_3 \cdot c^3 \cdot G^2 \log \frac{2}{\eta_k}}{(k+1)^{\theta}} \left[C_{30} + \sum_{i=\bar{k}}^{k} \frac{(i+3)^{\theta}}{i+1} \frac{1}{i+2} \right] & \text{if } \underline{k} \leq \bar{k} \leq k, \quad \text{(due to } \eta_i \geq \eta_k, \bar{k}, \text{ and } \underline{k} \text{ are both constants)} \end{cases} \\
&\leq \begin{cases} C_{31} \frac{c_3 \cdot c \cdot G^2 (\underline{k}+1)^{2c} \log \frac{2}{\eta_k}}{(k+1)^{\theta}} \left(\int_1^k t^{\theta-1-2c} dt \right) & \text{if } \underline{k} \leq k < \bar{k}, \\ C_{32} \frac{c_3 \cdot c^3 \cdot G^2 \log \frac{2}{\eta_k}}{(k+1)^{\theta}} \left(\int_1^k t^{\theta-2} dt \right) & \text{if } \underline{k} \leq \bar{k} \leq k, \end{cases} \\
&\leq \begin{cases} C_{31} \frac{c_3 \cdot c \cdot G^2 (\underline{k}+1)^{2c} \log \frac{2}{\eta_k}}{(k+1)^{\theta}} \cdot \frac{1}{2c-\theta} & \text{if } \underline{k} \leq k < \bar{k}, \quad \text{(due to } c > \theta) \\ C_{32} \frac{c_3 \cdot c^3 \cdot G^2 \log \frac{2}{\eta_k}}{(k+1)^{\theta}} k^{\theta-1} \cdot \frac{1}{\theta-1} & \text{if } \underline{k} \leq \bar{k} \leq k, \end{cases} \\
&\leq \begin{cases} C_{31} \frac{c_3 \cdot c \cdot G^2 (\underline{k}+1)^{2c} \log \frac{2}{\eta_k}}{k+1} \cdot \frac{1}{2c-\theta} & \text{if } \underline{k} \leq k < \bar{k}, \quad \text{(due to } \theta \geq 2 > 1) \\ C_{32} \frac{c_3 \cdot c^3 \cdot G^2 \log \frac{2}{\eta_k}}{k+1} \cdot \frac{1}{\theta-1} & \text{if } \underline{k} \leq \bar{k} \leq k, \end{cases} \\
&\leq c_3 G^2 C_{34} \frac{\log \frac{2}{\eta_k}}{k+1}. \qquad (34)
\end{aligned}$$

Combining (31) with (34), for all $k \geq \underline{k}$, gives

$$s_{k+1}^2 \leq \left(\frac{\underline{k}+2}{k+1} \right)^{\theta} \cdot s_{\underline{k}}^2 + c_3 G^2 C_{34} \frac{\log \frac{2}{\eta_k}}{k+1}, \qquad (35)$$

which implies, for all $k \geq \underline{k}$, that

$$\begin{aligned}
\|x_k - x^*\|^2 &\leq \left(\frac{\underline{k}+2}{k} \right)^{\theta} \left\|x_{\underline{k}} - x^*\right\|^2 + c_3 G^2 C_{34} \frac{2 \log \frac{2k}{\eta_0}}{k} \\
&= \left(\frac{\underline{k}+2}{k} \right)^{\theta} \left\|x_{\underline{k}} - x^*\right\|^2 + \zeta \left(\frac{\mu_f^2}{L_g^4} + \frac{2}{L_g^2} \left(1 + \frac{\mu_f}{L_g} \right)^2 \right) (\sigma + L_g(G_r + G_d)\zeta\underline{\alpha})^2 C_{34} \frac{\log \frac{2k}{\eta_0}}{k} \\
&= \bar{c}_1 \frac{\left\|x_{\underline{k}} - x^*\right\|^2}{k^{\theta}} + \bar{c}_2 \frac{\log \frac{2k}{\eta_0}}{k},
\end{aligned}$$

where we set $C_3 = C_{34}$. It follows from the definition of $\mathcal{E}_k^x$ and the above result that $\mathbb{P}\left[\bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k^x \mid \bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k\right] = 1$. In conclusion, for any given $\eta_0 \in (0, 6/\pi^2)$, it follows from Corollary 3.1 that

$$\begin{aligned}
\mathbb{P}\left[\bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k^x\right] &= \frac{\mathbb{P}\left[\bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k^x, \bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k\right]}{\mathbb{P}\left[\bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k \mid \bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k^x\right]} \geq \mathbb{P}\left[\bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k^x, \bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k\right] \\
&= \mathbb{P}\left[\bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k^x \mid \bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k\right] \mathbb{P}\left[\bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k\right] = \mathbb{P}\left[\bigcap_{k \geq \underline{k}}^{\infty} \mathcal{E}_k\right] \geq 1 - \eta_0 \pi^2/6 > 0,
\end{aligned}$$

which completes the proof. □

A.4 Proof of Theorem 3.3

Theorem 3.3. Given $\alpha > 0$, $d \in \mathbb{R}^n$, and the optimal solution x^* to problem (1), denote $z = x - \alpha d$ and $y = \mathrm{prox}_{\alpha r}(z)$. Let Assumption 3.4 hold. If

$$\left\| \frac{[z - x^*]_{g_i}}{\alpha} + \nabla_{g_i} f(x^*) \right\| < \delta^* \text{ for all } i \notin \mathcal{S}(x^*),$$

then $\mathcal{S}(y) \subseteq \mathcal{S}(x^*)$. Furthermore, if $\|y - x^*\| < \Delta^*$, then $\mathcal{S}(x^*) \subseteq \mathcal{S}(y)$ so that, in fact, $\mathcal{S}(y) = \mathcal{S}(x^*)$.

Proof. We start with the first claim $\mathcal{S}(y) \subseteq \mathcal{S}(x^*)$. It follows from Assumption 3.4 and the triangular inequality that, for all $i \notin \mathcal{S}(x^*)$, one has

$$\begin{aligned}
\left\| \frac{[z - x^*]_{g_i}}{\alpha} \right\| &= \left\| \frac{[z - x^*]_{g_i}}{\alpha} + \nabla_{g_i} f(x^*) - \nabla_{g_i} f(x^*) \right\| \\
&\le \left\| \frac{[z - x^*]_{g_i}}{\alpha} + \nabla_{g_i} f(x^*) \right\| + \|\nabla_{g_i} f(x^*)\| \\
&< \delta^* + \|\nabla_{g_i} f(x^*)\| \le \delta_{\min} + \|\nabla_{g_i} f(x^*)\| < \lambda_i.
\end{aligned}$$

Since $[x^*]_{g_i} = 0$ for all $i \notin \mathcal{S}(x^*)$, it follows that $\frac{[z-x^*]_{g_i}}{\alpha} \in \partial r_i([x^*]_{g_i})$[10]. It follows from the optimality condition for the proximal problem (Beck, 2017, Theorem 6.39) that this is true if and only if $[x^*]_{g_i} = \mathrm{prox}_{\alpha r_i}([z]_{g_i})$ for all $i \notin \mathcal{S}(x^*)$, which further implies $[y]_{g_i} = [x^*]_{g_i} = 0$ for all $i \notin \mathcal{S}(x^*)$. Consequently, $(\mathcal{S}(x^*))^c \subseteq (\mathcal{S}(y))^c$, which implies $\mathcal{S}(y) \subseteq \mathcal{S}(x^*)$.

Now we prove the second claim $\mathcal{S}(x^*) \subseteq \mathcal{S}(y)$. Note that $\|[y - x^*]_{g_i}\| \le \|y_k - x^*\|$ for any $i \in [n_{\mathcal{G}}]$. Therefore, when $\|y - x^*\| < \Delta^*$, for $i \in \mathcal{S}(x^*)$, $[y]_{g_i}$ cannot be 0 for all $i \in \mathcal{S}(x^*)$. Otherwise, $\Delta^* \le \|[x^*]_{g_i}\| < \Delta^*$ for $i \in \mathcal{S}(x^*)$. This proves that $\mathcal{S}(x^*) \subseteq \mathcal{S}(y_k)$. □

A.5 Proof of Theorem 3.4

Theorem 3.4. Let Assumption 3.1–Assumption 3.4 hold, $\zeta \in (0, 2)$, $\theta \ge 2$, $c = (2\theta L_g^2)/(\zeta \mu_f^2) > 2$, and $\underline{k} = \lceil 2c - 1 \rceil$. Consider the sequence $\{y_k\}$ of Algorithm 1 and define the event $\mathcal{E}_k^{\mathrm{id}} = \{\mathcal{S}(y_k) = \mathcal{S}(x^*)\}$ for all $k \ge 1$. Then, there exists constants $\{C_{41}, C_{42}\} \subseteq \mathbb{R}_+^n$ that are independent of k, $k_{\delta^*} = (C_{41}/\delta^*)^4$ and $k_{\Delta^*} = (C_{42}/\Delta^*)^4$ such that, with $K := \max\{k_{\delta^*}, k_{\Delta^*}, \underline{k}\}$, it follows that

$$\mathbb{P}\left[\bigcap_{k \ge K}^{\infty} \mathcal{E}_k^{\mathrm{id}} \right] \ge 1 - \frac{\eta_0 \pi^2}{6} > 0.$$

Proof. Denote $z_k = x_k - \alpha_k d_k$ for all $k \ge 1$, then it follows from Assumption 3.1(2) and the triangular inequality that

$$\begin{aligned}
\left\| \frac{z_k - x^*}{\alpha_k} + \nabla f(x^*) \right\| &\le \frac{1}{\alpha_k} \|x_k - x^*\| + \|d_k - \nabla f(x^*)\| \\
&\le \frac{1}{\alpha_k} \|x_k - x^*\| + \|d_k - \nabla f(x_k)\| + \|\nabla f(x_k) - \nabla f(x^*)\| \\
&\le \left(\frac{1}{\alpha_k} + L_g \right) \|x_k - x^*\| + \|d_k - \nabla f(x_k)\|. \qquad (36)
\end{aligned}$$

Conditioning on the events $\bigcap_{k \ge \underline{k}}^{\infty} \mathcal{E}_i$ and $\bigcap_{k \ge \underline{k}}^{\infty} \mathcal{E}_i^x$ happening (with $\mathcal{E}_k$ defined in Theorem 3.1 and $\mathcal{E}_k^x$ defined in Theorem 3.2), it follows from $\alpha_k \equiv \underline{\alpha}$ for all k (Assumption 3.3), Corollary 3.1, and Theorem 3.2 that, there exists a constant $C_{41} > 0$ that is independent of k, for all $k \ge \underline{k}$,

$$\left(\frac{1}{\alpha_k} + L_g \right) \|x_k - x^*\| + \|d_k - \nabla f(x_k)\| \le C_{41} \sqrt{\frac{\log k}{k}}. \qquad (37)$$

[10] The subdifferential is given by $\partial \|x\| = \{v \in \mathbb{R}^n \mid \|v\| \le 1\}$.

Combining (36) and (37), we know for all $k \geq k_{\delta^*}$ that $C_{41}\sqrt{\frac{\log k}{k}} \leq C_{41}/k^4 < C_{41}/k_{\delta^*}^4 = \delta^*$ [11]. Together with Theorem 3.3 and the definition of y_k (line 11 of Algorithm 1), we have $\mathcal{S}(y_k) \subseteq \mathcal{S}(x^*)$ for all $k \geq \max\{k_{\delta^*}, \underline{k}\}$.

It follows from the non-expansiveness (Beck, 2017, Theorem 6.42) of the proximal operator, $x^* = \text{prox}_{\alpha_k r}(x^* - \alpha_k \nabla f(x^*))$, the definition of y_k (line 11 of Algorithm 1), and the triangular inequality that

$$\begin{aligned}\|y_k - x^*\| &= \left\|\text{prox}_{\alpha_k r}(x_k - \alpha_k \nabla f(x_k)) - \text{prox}_{\alpha_k r}(x^* - \alpha_k \nabla f(x^*))\right\| \\ &\leq \|(x_k - x^*) - \alpha_k(d_k - \nabla f(x^*))\| \\ &\leq \|x_k - x^*\| + \alpha_k \|d_k - \nabla f(x^*)\|. \end{aligned} \tag{38}$$

Again, **conditioning on the events $\bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i$ and $\bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i^x$ happening**, it follows from $\alpha_k \equiv \underline{\alpha}$ for all k (Assumption 3.3), Corollary 3.1, Theorem 3.2, and (38) that, there exist a constant $C_{42} > 0$ that is independent of k, such that for all $k \geq \underline{k}$, $\|y_k - x^*\| \leq C_{42}\sqrt{\frac{\log k}{k}}$. Therefore, when $k \geq k_{\Delta^*}$, it follows that $C_{42}\sqrt{\frac{\log k}{k}} \leq C_{42}/k^4 < C_{42}/k_{\Delta^*}^4 = \Delta^*$. Together with Theorem 3.3 and the definition of y_k (line 11 of Algorithm 1), we have $\mathcal{S}(x^*) \subseteq \mathcal{S}(y_k)$ for all $k \geq \max\{k_{\Delta^*}, \underline{k}\}$. Therefore, when $k \geq K = \max\{k_{\delta^*}, k_{\Delta^*}, \underline{k}\}$, together with the fact that $\mathbb{P}\left[\bigcap_{k\geq K}^{\infty} \mathcal{E}_k^{\text{id}} \mid \bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i, \bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i^x\right] = 1$, it follows that

$$\begin{aligned}\mathbb{P}\left[\bigcap_{k\geq K}^{\infty} \mathcal{E}_k^{\text{id}}\right] &= \frac{\mathbb{P}\left[\bigcap_{k\geq K}^{\infty} \mathcal{E}_k^{\text{id}}, \bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i, \bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i^x\right]}{\mathbb{P}\left[\bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i, \bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i^x \mid \bigcap_{k\geq K}^{\infty} \mathcal{E}_k^{\text{id}}\right]} \geq \mathbb{P}\left[\bigcap_{k\geq K}^{\infty} \mathcal{E}_k^{\text{id}}, \bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i, \bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i^x\right] \\ &= \mathbb{P}\left[\bigcap_{k\geq K}^{\infty} \mathcal{E}_k^{\text{id}} \mid \bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i, \bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i^x\right] \mathbb{P}\left[\bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i, \bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i^x\right] \\ &= \mathbb{P}\left[\bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i, \bigcap_{k\geq\underline{k}}^{\infty} \mathcal{E}_i^x\right] \geq 1 - \frac{\eta_0 \pi^2}{6}, \end{aligned}$$

which completes the proof. □

A.6 Proofs for additional lemmas

Lemma A.4. *Denote $\mathcal{F}_1 = \sigma(x_1)$ and, for all $k \geq 2$, denote $\mathcal{F}_k$ as the σ-algebra generated by the random variables $\{\{\Xi_{1,i}\}_{i=1}^m, \ldots, \{\Xi_{(k-1),i}\}_{i=1}^m\}$ (of which $\{\{\xi_{1,i}\}_{i=1}^m, \ldots, \{\xi_{(k-1),i}\}_{i=1}^m\}$ is a realization) so that $\{\mathcal{F}_k\}$ is a filtration. If (i) there exists a constant $c_e > 0$ such that for all $k \geq 1$, $\mathbb{P}\{\|d_k - \nabla f(x_k)\| \leq c_e \mid \mathcal{F}_k\} = 1$ and (ii) there exists a constant c_α such that for a given $\alpha > 0$ and all $k \geq 1$, $\mathbb{P}\{\chi(x_k; \alpha) \leq c_\alpha \mid \mathcal{F}_k\} = 1$, then there exists a constant $G_d > 0$ such that for all $k \geq 1$, it holds that $\mathbb{P}\{\|d_k\| \leq G_d \mid \mathcal{F}_k\} = 1$.*

Proof. To see why the implication holds, we define $\tilde{y}_k = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k))$ and we make an algorithmic choice $\alpha_k \equiv \alpha$ for all k. Since $y_k = \text{prox}_{\alpha r}(x_k - \alpha d_k)$, then $\frac{x_k - y_k}{\alpha} - d_k \in \partial r(y_k)$. It follows from Assumption 3.2.2 that $\left\|\frac{x_k - y_k}{\alpha} - d_k\right\| \leq G_r$. By the triangle inequality, we have

$$\begin{aligned}\|d_k\| &\leq G_r + \frac{\|x_k - y_k\|}{\alpha} \leq G_r + \frac{\|x_k - \tilde{y}_k\|}{\alpha} + \frac{\|y_k - \tilde{y}_k\|}{\alpha} \\ &= G_r + \chi_k(\alpha) + \frac{\|\text{prox}_{\alpha r}(x_k - \alpha d_k) - \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k))\|}{\alpha} \\ &\leq G_r + \chi_k(\alpha) + \|d_k - \nabla f(x_k)\| \leq G_r + c_\alpha + c_e, \end{aligned}$$

where the penultimate inequality holds by the non-expansiveness of the proximal operator(Beck, 2017, Theorem 6.42). □

Lemma A.5. *Consider the* `RDA` *algorithm with its update defined as*

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \left\{d_k^T x + r(x) + \frac{\rho_k}{k}\|x\|^2\right\} \text{ with } d_k = \frac{k-1}{k} d_{k-1} + \frac{1}{k}\nabla \ell(x_k; \xi_k),$$

where ξ_k is a i.i.d sample from $\mathcal{P}$.

[11] We use the inequality $\sqrt{\frac{\log x}{x}} < \frac{1}{x^{1/4}}$ for all $x > 1$.

1. *If $\rho_k = \underline{\alpha}\sqrt{k}$ for a given $k \geq 1$ and $\underline{\alpha}$ is defined as in Assumption 3.3, then the update can be equivalently written as*

$$x_{k+1} = \text{prox}_{\alpha_k r}\left(-\alpha_k d_k\right) \text{ with } \alpha_k = \frac{\sqrt{k}}{\underline{\alpha}}.$$

2. *Assume r is $\mu_r > 0$ strongly convex. Further, assume that there are constants $\{G, D\} \subset (0, \infty)$ such that, for all $k \geq 1$, it holds that $\|\nabla \ell(x_k; \xi_k)\| \leq G$ and $\|x^*\| \leq D$. If $\rho_k = \underline{\alpha}\sqrt{k}$, then*

$$\mathbb{E}\left[\|x_k - x^*\|^2\right] \leq \frac{2\left(\underline{\alpha}D^2 + G^2/\underline{\alpha}\right)}{\mu_r}\frac{1}{\sqrt{k}}.$$

Moreover, for any $\epsilon > 0$ and $k \geq 1$, it holds that

$$\mathbb{P}\left[\|x_k - x^*\| \geq \epsilon\right] \leq \sqrt{\frac{2\left(\underline{\alpha}D^2 + G^2/\underline{\alpha}\right)}{\mu_r \epsilon^2}}\frac{1}{k^{1/4}}.$$

3. *Assume f and r are $\mu_f > 0$ and $\mu_r > 0$ strongly convex, respectively. Further, assume that there are constants $\{G, D\} \subset (0, \infty)$ such that, for all $k \geq 1$, it holds that $\|\nabla \ell(x_k; \xi_k)\| \leq G$ and $\|x^*\| \leq D$. If $\rho_k = \underline{\alpha}\sqrt{k}$, then for all $k \geq 1$, it holds that*

$$\mathbb{P}\left[\mathcal{S}\left(x_{k+1}\right) = \mathcal{S}\left(x^*\right)\right] \geq 1 - \eta_k^{\texttt{RDA}},$$

where $\eta_k^{\texttt{RDA}} = \max\left\{\mathcal{O}\left(\frac{1}{\delta^* \cdot k^{1/4}}\right), \mathcal{O}\left(\frac{1}{\Delta^* \cdot k^{1/4}}\right)\right\}$.

Proof. For part 1, let $\beta_k = \underline{\alpha}\sqrt{k}$. Then

$$\begin{aligned}
\arg\min_{x \in \mathbb{R}^n}\left\{d_k^T x + r(x) + \frac{\rho_k}{k}\|x\|^2\right\} &= \arg\min_{x \in \mathbb{R}^n}\left\{d_k^T x + r(x) + \frac{1}{\frac{1}{\underline{\alpha}}\sqrt{k}}\|x\|^2\right\} \\
&= \arg\min_{x \in \mathbb{R}^n}\left\{d_k^T x + r(x) + \frac{1}{\alpha_k}\|x\|^2\right\} \\
&= \arg\min_{x \in \mathbb{R}^n}\left\{\frac{1}{\alpha_k}\|x + \alpha_k d_k\|^2 + r(x)\right\} = \text{prox}_{\alpha_k r}\left(-\alpha_k d_k\right).
\end{aligned}$$

For part 2, it follows from Xiao (2009, Equation 22) and Xiao (2009, Corollary 2) that for all $k \geq 1$,

$$\mathbb{E}\left[\|x_k - x^*\|^2\right] \leq \frac{2}{\mu_r k}\left(\underline{\alpha}D^2 + G^2/\underline{\alpha}\right)\sqrt{k} = \frac{2\left(\underline{\alpha}D^2 + G^2/\underline{\alpha}\right)}{\mu_r}\frac{1}{\sqrt{k}}.$$

It follows from Jensen' inequality that

$$\mathbb{E}\left[\|x_k - x^*\|\right] \leq \sqrt{\mathbb{E}\left[\|x_k - x^*\|^2\right]} \leq \sqrt{\frac{2\left(\underline{\alpha}D^2 + G^2/\underline{\alpha}\right)}{\mu_r}}\frac{1}{k^{1/4}},$$

which together with the Markov inequality implies that

$$\mathbb{P}\left[\|x_k - x^*\| \geq \epsilon\right] \leq \sqrt{\frac{2\left(\underline{\alpha}D^2 + G^2/\underline{\alpha}\right)}{\mu_r \epsilon^2}}\frac{1}{k^{1/4}}.$$

For part 3, consider three events $\mathcal{E}_{k,1}^{\texttt{RDA}} := \{\|d_k - \nabla f(x^*)\| < \delta^*/2\}$, $\mathcal{E}_{k,2}^{\texttt{RDA}} := \{(1/\alpha_k + L_g)\|x_k - x^*\| < \delta^*/2\}$, and $\mathcal{E}_{k,3}^{\texttt{RDA}} := \{\|x_{k+1} - x^*\| < \Delta^*\}$. It follows from Lee and Wright (2012, Theorem 11, equation (31)) and part 2 of this lemma that

$$\mathbb{P}\left[\left(\mathcal{E}_{k,1}^{\texttt{RDA}}\right)^c\right] \leq \mathcal{O}\left(\frac{1}{\delta^* \cdot k^{1/4}}\right), \tag{39}$$

$$\mathbb{P}\left[\left(\mathcal{E}_{k,2}^{\texttt{RDA}}\right)^c\right] \leq \mathcal{O}\left(\frac{1}{\delta^* \cdot k^{3/4}}\right), \text{ and} \tag{40}$$

$$\mathbb{P}\left[\left(\mathcal{E}_{k,2}^{\texttt{RDA}}\right)^c\right] \leq \mathcal{O}\left(\frac{1}{\Delta^* \cdot k^{1/4}}\right). \tag{41}$$

It follows from the union bound and (39)–(41) that

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}_{k,1}^{\texttt{RDA}}\bigcap\mathcal{E}_{k,2}^{\texttt{RDA}}\bigcap\mathcal{E}_{k,3}^{\texttt{RDA}}\right] &= 1-\mathbb{P}\left[\left(\mathcal{E}_{k,1}^{\texttt{RDA}}\right)^c\bigcup\left(\mathcal{E}_{k,2}^{\texttt{RDA}}\right)^c\bigcup\left(\mathcal{E}_{k,3}^{\texttt{RDA}}\right)^c\right]\\ &\geq 1-\left(\mathbb{P}\left[\left(\mathcal{E}_{k,1}^{\texttt{RDA}}\right)^c\right]+\mathbb{P}\left[\left(\mathcal{E}_{k,2}^{\texttt{RDA}}\right)^c\right]+\mathbb{P}\left[\left(\mathcal{E}_{k,3}^{\texttt{RDA}}\right)^c\right]\right)\\ &= 1-\max\left\{\mathcal{O}\left(\frac{1}{\delta^*\cdot k^{1/4}}\right),\mathcal{O}\left(\frac{1}{\Delta^*\cdot k^{1/4}}\right)\right\},\end{aligned}$$

which together with Theorem 3.3 implies that, for any chosen $k\geq 1$,

$$\mathbb{P}\left[\mathcal{S}\left(x_{k+1}\right)=\mathcal{S}\left(x^*\right)\right]\geq 1-\eta_k^{\texttt{RDA}},$$

which completes the proof. □

B Experiments

B.1 Discussions on the performance gaps in different methods

First, `ProxSVRG` performs poorly on test instances induced by the datasets phishing, rcv1, real-sim, and news20. It can be checked that these datasets cover different sample sizes and decision variable dimensions. We attribute the cause of poor performance to the inner loop length parameter, which is difficult to choose to work on all test instances. In the experiments, we set it to 1 for all cases to follow the original paper's experimental setting (Xiao and Zhang, 2014).

Second, `SAGA` performed quite well on the first 32 test instances, where the memory limit is not violated, and failed on the remaining 48 test instances (marked as the darkest red) because the program terminates immediately due to memory limits being exceeded.

Third, `RDA` appears to perform poorly compared with `PStorm` (`S-PStorm`) because the prox step of `RDA` only applies to its initial point $x_0=0$ with updated search direction $-\alpha_k d_k$ (see Lemma A.5(1)), whereas `PStorm` (`S-PStorm`) applies the prox step at the up-to-date iterate x_k.

Finally, one can see that `S-PStorm` significantly outperforms `PStorm`. We attribute this to a combination of the stabilization that we introduced and that the step size for `PStorm` was designed for nonconvex problems (for our tests, nonetheless, we fine-tuned the step size for `PStorm` to be fair).

B.2 Additional results

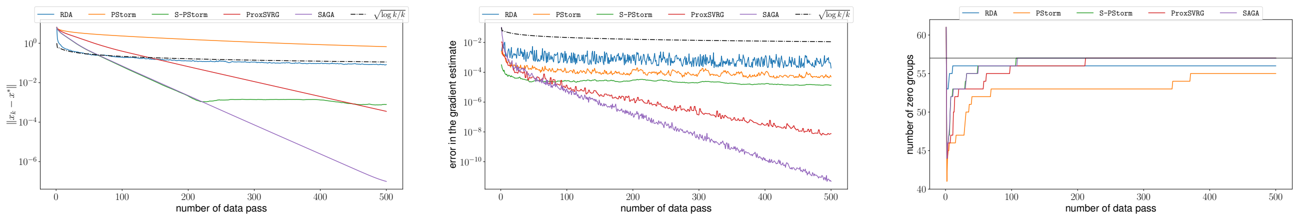
We visualize three metrics: the distance to the optimal solution $\|x_k-x^*\|$ (x^* is obtained by the `FaRSA-Group` algorithm), the error in the gradient evaluation ϵ_k (defined in Theorem 3.1), and the sparse structure of major iterates, which can be found in the first, second, and third column of Figure 3, respectively. The first metric measures the convergence speed of $\{x_k\}$, the second metric shows how fast the error in the stochastic gradient estimator d_k (defined in Algorithm 1 line 9) diminishing to zero, and the third metric visualizes the progress made with respect to support identification.

For demonstration, we only show results on six moderate-size datasets with randomly picked problem parameters $\Lambda=0.1$ and number of groups $\lfloor 0.5n\rfloor$. We remark that in some plots, lines that represent different algorithms could visually overlap. For example, the green line (`S-PStorm`) and purple line (`SAGA`) overlap in the first column for dataset phishing and rcv1.[12] We also emphasize that `SAGA` does not appear in the Figure 3(f) due to memory limitation.

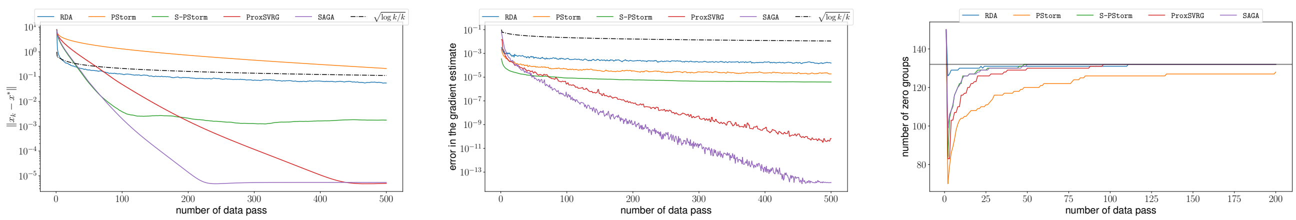
From the first and the second column of Figure 3, it can be observed that the rates at which the x_k converges to x^* and ϵ_k converges to 0 seem to be bounded by $\mathcal{O}(\sqrt{\log k/k})$, which matches our theoretical results in Theorem 3.1 and Theorem 3.2. We can also observe that for the relatively large datasets rcv1 and real-sim, 1000 data passes is not enough to obtain accurate estimates of x^*, but a decent ratio of zeros groups is identified nonetheless.

[12] The reason is that the numerical difference between $\|x_k-x^*\|$ for `S-PStorm` and `SAGA` is of order 10^{-2}.

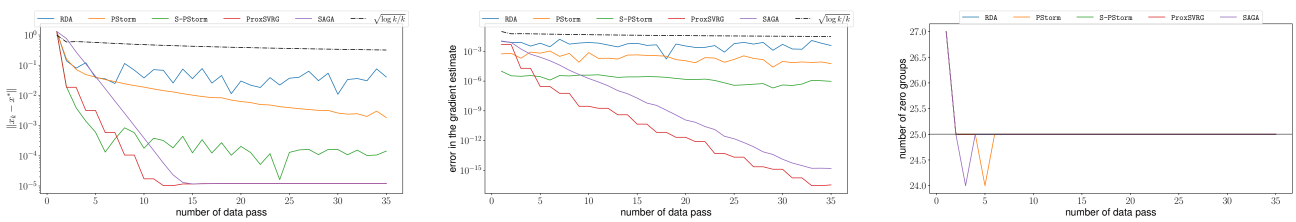
Figure 3: Visualization of three metrics: the distance to optimal solution $\|x_k - x^*\|$ (the first column), error in the gradient evaluation ϵ_k (the second column), and the progress of support identification on different datasets (the third column). We added a dotted reference line corresponding to $\sqrt{\log k/k}$ (for $k \geq 2$) for the plots in the first and second columns. In addition, we added a horizontal black reference line for the plots in the third column to indicate the number of zero groups at the optimal solution x^*.



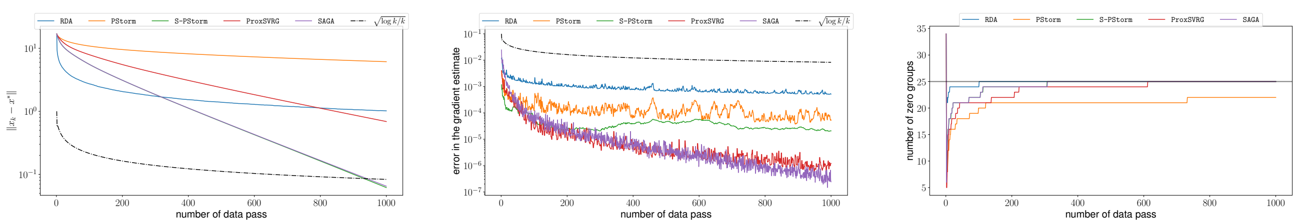
(a) a9a



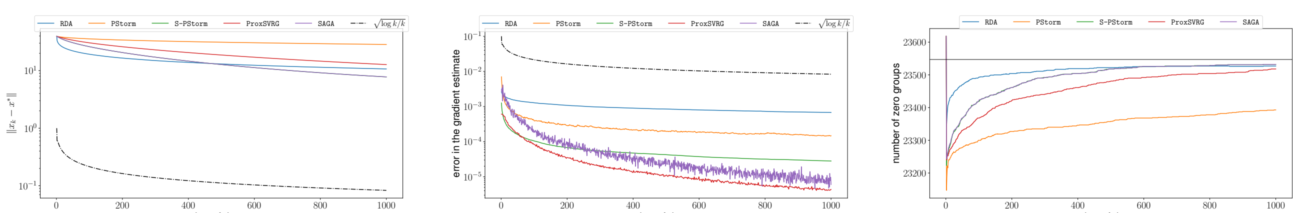
(b) w8a



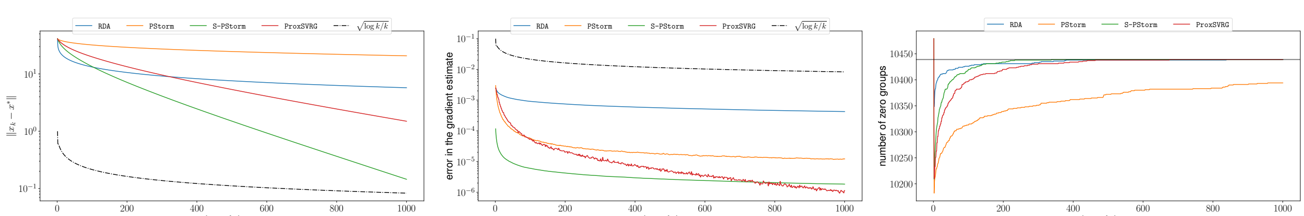
(c) covtype



(d) phishing



(e) rcv1



(f) real-sim

Finally, we provide the raw data for the metrics of total identification (Table 3), first identification (Table 4), first consistent identification (Table 5), and the last iterate support recovery (Table 6); for an explanation of their precise meaning, revisit Section 4.2. All results (excluding `FaRSA-Group` which is a deterministic algorithm) are reported as the average of three independent runs. In all tables, the problem instance is formatted as (dataset name)-(value of Λ)-(ratio of # of groups).

We remark that NaN represents that a particular method failed to identify the support within 1000 data pass. We also removed the instances that all five methods failed to identify the support.

Table 3: Total number of support identifications.

instance	`ProxSVRG`	`SAGA`	`RDA`	`PStorm`	`S-PStorm`
a9a-0.1-0.25	976.0	987.0	1000.0	899.0	988.0
a9a-0.1-0.5	790.0	893.0	75.0	NaN	895.0
a9a-0.1-0.75	932.0	965.0	998.0	547.0	966.0
a9a-0.1-1.0	932.0	965.0	998.0	547.0	966.0
a9a-0.01-0.25	NaN	NaN	NaN	NaN	93.0
a9a-0.01-0.5	578.0	788.0	706.0	NaN	786.0
a9a-0.01-0.75	652.0	825.0	84.0	NaN	825.0
a9a-0.01-1.0	652.0	825.0	84.0	NaN	825.0
covtype-0.1-0.25	998.0	999.0	1000.0	998.0	1000.0
covtype-0.1-0.5	1000.0	999.0	1000.0	999.0	1000.0
covtype-0.1-0.75	1000.0	999.0	1000.0	999.0	1000.0
covtype-0.1-1.0	1000.0	999.0	1000.0	999.0	1000.0
covtype-0.01-0.25	1000.0	1000.0	1000.0	1000.0	1000.0
covtype-0.01-0.5	1000.0	1000.0	1000.0	1000.0	1000.0
covtype-0.01-0.75	994.0	991.0	999.0	837.0	1000.0
covtype-0.01-1.0	994.0	991.0	999.0	837.0	1000.0
phishing-0.1-0.25	792.0	896.0	991.0	NaN	895.0
phishing-0.1-0.5	390.0	694.0	901.0	NaN	695.0
phishing-0.1-0.75	420.0	710.0	502.0	NaN	710.0
phishing-0.1-1.0	326.0	662.0	28.0	NaN	667.0
w8a-0.1-0.25	960.0	979.0	997.0	749.0	980.0
w8a-0.1-0.5	906.0	952.0	891.0	120.0	954.0
w8a-0.1-0.75	886.0	942.0	951.0	NaN	942.0
w8a-0.1-1.0	886.0	942.0	951.0	NaN	942.0
w8a-0.01-0.5	164.0	581.0	NaN	NaN	580.0
w8a-0.01-0.75	NaN	185.0	NaN	NaN	195.0
w8a-0.01-1.0	NaN	185.0	NaN	NaN	195.0
real-sim-0.1-0.25	NaN	NaN	NaN	NaN	212.0
real-sim-0.1-0.5	326.0	NaN	163.0	NaN	664.0
news20-0.1-0.25	38.0	NaN	NaN	NaN	26.0
news20-0.1-0.5	6.0	NaN	NaN	NaN	312.0
url-combined-0.1-0.25	4.0	NaN	NaN	NaN	2.0
avazu-app.tr-0.1-0.25	6.0	NaN	4.0	3.0	3.0
avazu-app.tr-0.1-0.5	2.0	NaN	3.0	2.0	2.0
avazu-app.tr-0.1-0.75	2.0	NaN	2.0	2.0	2.0
avazu-app.tr-0.1-1.0	NaN	NaN	2.0	1.0	1.0
avazu-app.tr-0.01-0.25	6.0	NaN	3.0	NaN	3.0
avazu-app.tr-0.01-0.5	2.0	NaN	2.0	NaN	2.0
avazu-app.tr-0.01-0.75	2.0	NaN	1.0	NaN	2.0
avazu-app.tr-0.01-1.0	NaN	NaN	2.0	NaN	1.0

Table 4: First support identification.

instance	ProxSVRG	SAGA	RDA	PStorm	S-PStorm
a9a-0.1-0.25	25.0	14.0	1.0	102.0	13.0
a9a-0.1-0.5	211.0	108.0	660.0	NaN	106.0
a9a-0.1-0.75	69.0	36.0	3.0	454.0	35.0
a9a-0.1-1.0	69.0	36.0	3.0	454.0	35.0
a9a-0.01-0.25	NaN	NaN	NaN	NaN	908.0
a9a-0.01-0.5	423.0	213.0	264.0	NaN	215.0
a9a-0.01-0.75	349.0	176.0	77.0	NaN	176.0
a9a-0.01-1.0	349.0	176.0	77.0	NaN	176.0
covtype-0.1-0.25	3.0	1.0	1.0	1.0	1.0
covtype-0.1-0.5	1.0	1.0	1.0	1.0	1.0
covtype-0.1-0.75	1.0	1.0	1.0	1.0	1.0
covtype-0.1-1.0	1.0	1.0	1.0	1.0	1.0
covtype-0.01-0.25	1.0	1.0	1.0	1.0	1.0
covtype-0.01-0.5	1.0	1.0	1.0	1.0	1.0
covtype-0.01-0.75	3.0	5.0	1.0	1.0	1.0
covtype-0.01-1.0	3.0	5.0	1.0	1.0	1.0
phishing-0.1-0.25	209.0	105.0	10.0	NaN	106.0
phishing-0.1-0.5	611.0	307.0	100.0	NaN	306.0
phishing-0.1-0.75	581.0	291.0	492.0	NaN	291.0
phishing-0.1-1.0	675.0	339.0	734.0	NaN	334.0
w8a-0.1-0.25	41.0	22.0	4.0	252.0	21.0
w8a-0.1-0.5	95.0	49.0	110.0	881.0	47.0
w8a-0.1-0.75	115.0	59.0	38.0	NaN	59.0
w8a-0.1-1.0	115.0	59.0	38.0	NaN	59.0
w8a-0.01-0.5	837.0	420.0	NaN	NaN	421.0
w8a-0.01-0.75	NaN	816.0	NaN	NaN	806.0
w8a-0.01-1.0	NaN	816.0	NaN	NaN	806.0
real-sim-0.1-0.25	NaN	NaN	NaN	NaN	789.0
real-sim-0.1-0.5	675.0	NaN	838.0	NaN	337.0
news20-0.1-0.25	963.0	NaN	NaN	NaN	504.0
news20-0.1-0.5	995.0	NaN	NaN	NaN	523.0
url-combined-0.1-0.25	9.0	NaN	NaN	NaN	5.0
avazu-app.tr-0.1-0.25	3.0	NaN	1.0	1.0	1.0
avazu-app.tr-0.1-0.5	3.0	NaN	1.0	1.0	1.0
avazu-app.tr-0.1-0.75	3.0	NaN	1.0	1.0	1.0
avazu-app.tr-0.1-1.0	NaN	NaN	1.0	1.0	1.0
avazu-app.tr-0.01-0.25	3.0	NaN	1.0	NaN	1.0
avazu-app.tr-0.01-0.5	3.0	NaN	1.0	NaN	1.0
avazu-app.tr-0.01-0.75	3.0	NaN	2.0	NaN	1.0
avazu-app.tr-0.01-1.0	NaN	NaN	1.0	NaN	1.0

Table 5: First consistent support identification.

instance	ProxSVRG	SAGA	RDA	PStorm	S-PStorm
a9a-0.1-0.25	25.0	14.0	1.0	102.0	13.0
a9a-0.1-0.5	211.0	108.0	NaN	NaN	106.0
a9a-0.1-0.75	69.0	36.0	3.0	454.0	35.0
a9a-0.1-1.0	69.0	36.0	3.0	454.0	35.0
a9a-0.01-0.25	NaN	NaN	NaN	NaN	908.0
a9a-0.01-0.5	423.0	213.0	299.0	NaN	215.0
a9a-0.01-0.75	349.0	176.0	NaN	NaN	176.0
a9a-0.01-1.0	349.0	176.0	NaN	NaN	176.0
covtype-0.1-0.25	3.0	3.0	1.0	11.0	1.0
covtype-0.1-0.5	1.0	3.0	1.0	5.0	1.0
covtype-0.1-0.75	1.0	3.0	1.0	5.0	1.0
covtype-0.1-1.0	1.0	3.0	1.0	5.0	1.0
covtype-0.01-0.25	1.0	1.0	1.0	1.0	1.0
covtype-0.01-0.5	1.0	1.0	1.0	1.0	1.0
covtype-0.01-0.75	29.0	24.0	8.0	879.0	1.0
covtype-0.01-1.0	29.0	24.0	8.0	879.0	1.0
phishing-0.1-0.25	209.0	105.0	10.0	NaN	106.0
phishing-0.1-0.5	611.0	307.0	100.0	NaN	306.0
phishing-0.1-0.75	581.0	291.0	520.0	NaN	291.0
phishing-0.1-1.0	675.0	339.0	997.0	NaN	334.0
w8a-0.1-0.25	41.0	22.0	4.0	252.0	21.0
w8a-0.1-0.5	95.0	49.0	110.0	881.0	47.0
w8a-0.1-0.75	115.0	59.0	65.0	NaN	59.0
w8a-0.1-1.0	115.0	59.0	65.0	NaN	59.0
w8a-0.01-0.5	837.0	420.0	NaN	NaN	421.0
w8a-0.01-0.75	NaN	816.0	NaN	NaN	806.0
w8a-0.01-1.0	NaN	816.0	NaN	NaN	806.0
real-sim-0.1-0.25	NaN	NaN	NaN	NaN	789.0
real-sim-0.1-0.5	675.0	NaN	838.0	NaN	337.0
news20-0.1-0.25	963.0	NaN	NaN	NaN	NaN
news20-0.1-0.5	995.0	NaN	NaN	NaN	523.0
url-combined-0.1-0.25	9.0	NaN	NaN	NaN	5.0
avazu-app.tr-0.1-0.25	3.0	NaN	1.0	1.0	1.0
avazu-app.tr-0.1-0.5	3.0	NaN	1.0	1.0	1.0
avazu-app.tr-0.1-0.75	3.0	NaN	1.0	1.0	1.0
avazu-app.tr-0.1-1.0	NaN	NaN	1.0	1.0	1.0
avazu-app.tr-0.01-0.25	3.0	NaN	4.0	NaN	1.0
avazu-app.tr-0.01-0.5	3.0	NaN	4.0	NaN	1.0
avazu-app.tr-0.01-0.75	3.0	NaN	2.0	NaN	1.0
avazu-app.tr-0.01-1.0	NaN	NaN	1.0	NaN	1.0

Table 6: Last iterate sparsity.

instance	FaRSAGroup	ProxSVRG	SAGA	RDA	PStorm	S-PStorm
a9a-0.1-0.25	26	26.0	26.0	26.0	26.0	26.0
a9a-0.1-0.5	57	57.0	57.0	56.0	56.0	57.0
a9a-0.1-0.75	86	86.0	86.0	86.0	86.0	86.0
a9a-0.1-1.0	117	117.0	117.0	117.0	117.0	117.0
a9a-0.01-0.25	20	19.0	19.0	18.0	16.0	20.0
a9a-0.01-0.5	44	44.0	44.0	44.0	38.0	44.0
a9a-0.01-0.75	65	65.0	65.0	66.0	58.0	65.0
a9a-0.01-1.0	96	96.0	96.0	97.0	89.0	96.0
covtype-0.1-0.25	11	11.0	11.0	11.0	11.0	11.0
covtype-0.1-0.5	25	25.0	25.0	25.0	25.0	25.0
covtype-0.1-0.75	38	38.0	38.0	38.0	38.0	38.0
covtype-0.1-1.0	52	52.0	52.0	52.0	52.0	52.0
covtype-0.01-0.25	10	10.0	10.0	10.0	10.0	10.0
covtype-0.01-0.5	22	22.0	22.0	22.0	22.0	22.0
covtype-0.01-0.75	33	33.0	33.0	33.0	33.0	33.0
covtype-0.01-1.0	47	47.0	47.0	47.0	47.0	47.0
phishing-0.1-0.25	12	12.0	12.0	12.0	9.0	12.0
phishing-0.1-0.5	25	25.0	25.0	25.0	22.0	25.0
phishing-0.1-0.75	43	43.0	43.0	43.0	41.0	43.0
phishing-0.1-1.0	59	59.0	59.0	59.0	55.0	59.0
w8a-0.1-0.25	57	57.0	57.0	57.0	57.0	57.0
w8a-0.1-0.5	132	132.0	132.0	132.0	132.0	132.0
w8a-0.1-0.75	208	208.0	208.0	208.0	206.0	208.0
w8a-0.1-1.0	281	281.0	281.0	281.0	279.0	281.0
w8a-0.01-0.5	79	79.0	79.0	77.0	67.0	79.0
w8a-0.01-0.75	150	149.0	150.0	149.0	129.0	150.0
w8a-0.01-1.0	214	213.0	214.0	213.0	189.0	214.0
real-sim-0.1-0.25	5211	5210.0	NaN	5210.0	5187.0	5211.0
real-sim-0.1-0.5	10439	10439.0	NaN	10439.0	10394.0	10439.0
news20-0.1-0.25	338697	338697.0	NaN	338661.0	338674.0	338698.0
news20-0.1-0.5	677496	677496.0	NaN	677459.0	677464.0	677496.0
url-combined-0.1-0.25	807983	807983.0	NaN	807982.0	807974.0	807983.0
avazu-app.tr-0.1-0.25	249995	249995.0	NaN	249995.0	249995.0	249995.0
avazu-app.tr-0.1-0.5	499993	499993.0	NaN	499993.0	499993.0	499993.0
avazu-app.tr-0.1-0.75	749990	749990.0	NaN	749990.0	749990.0	749990.0
avazu-app.tr-0.1-1.0	999988	999987.0	NaN	999988.0	999988.0	999988.0
avazu-app.tr-0.01-0.25	249980	249980.0	NaN	249980.0	249973.0	249980.0
avazu-app.tr-0.01-0.5	499978	499978.0	NaN	499979.0	499970.0	499978.0
avazu-app.tr-0.01-0.75	749976	749976.0	NaN	749976.0	749972.0	749976.0
avazu-app.tr-0.01-1.0	999973	999814.0	NaN	999973.0	999965.0	999973.0