Kill a Bird with Two Stones: Closing the Convergence Gaps in Non-Strongly Convex Optimization by Directly Accelerated SVRG with Double Compensation and Snapshots

Yuanyuan Liu 1  Fanhua Shang 2  Weixin An 1  Hongying Liu 1 3  Zhouchen Lin 4 5 3

Abstract

Recently, some accelerated stochastic variance reduction algorithms such as Katyusha and ASVRG-ADMM achieve faster convergence than non-accelerated methods such as SVRG and SVRG-ADMM. However, there are still some gaps between the oracle complexities and their lower bounds. To fill in these gaps, this paper proposes a novel Directly Accelerated stochastic Variance reduction algorithm with two Snapshots (DAVIS) for non-strongly convex (non-SC) unconstrained problems. Our theoretical results show that DAVIS achieves the optimal convergence rate $O(1/(nS^2))$ and optimal gradient complexity $O(n + \sqrt{nL}/\epsilon)$, which is identical to its lower bound. To the best of our knowledge, this is the first directly accelerated algorithm that attains the lower bound and improves the convergence rate from $O(1/S^2)$ to $O(1/(nS^2))$. Moreover, we extend DAVIS and theoretical results to non-SC problems with an equality constraint, and prove that the proposed DAVIS-ADMM algorithm with double snapshots for each variable also attains the optimal convergence rate $O(1/(nS))$ and optimal oracle complexity $O(n + L/\epsilon)$ for such problems, and it is at least by a factor $n/S$ faster than existing accelerated stochastic algorithms, where $n \gg S$ in general.

1. Introduction

Consider the following finite-sum composite convex minimization problem

$$\min_{x \in \mathbb{R}^d} \left\{ F(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(x) + h(x) \right\},$$

(1)

where each $f_i(\cdot)$ is convex and $h(\cdot)$ is convex but possibly non-smooth. We define $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$. This problem arises frequently in machine learning, signal processing, statistics, and operations research (Bubeck, 2015), such as regularized empirical risk minimization. In many real-world applications, the number of component functions (i.e., $n$) is usually very large so that even first-order methods become computationally burdensome due to high per-iteration complexity $O(nd)$. Stochastic gradient descent (SGD) (Robbins & Monro, 1951) uses the gradient of only one (or a small batch of) randomly chosen $f_i$ to estimate full gradient in each iteration, and enjoys a significantly lower cost $O(d)$.

In recent years, stochastic (or incremental) variance reduction methods have received extensive attention due to their low per-iteration cost and ability to handle large-scale problems. In particular, research on variance reduction methods (e.g., SAG (Roux et al., 2012), SDCA (Shalev-Shwartz & Zhang, 2013), SVRG (Johnson & Zhang, 2013), SAGA (Defazio et al., 2014), and their proximal variants, e.g., Prox-SVRG (Xiao & Zhang, 2014)), and stochastic variance reduced algorithms of the alternating direction method of multipliers (ADMM) (e.g., SAG-ADMM (Zhong & Kwok, 2014), SDCA-ADMM (Suzuki, 2014) and SVRG-ADMM (Zheng & Kwok, 2016)) have made exciting progress, e.g., linear convergence for strongly convex (SC) problems.

For solving the SC problem (1), the oracle complexity (i.e., the number of Incremental First-order Oracle calls and Proximal Oracle calls needed to find an $\epsilon$-suboptimal solution) of the stochastic variance reduction methods mentioned above is $O((n + \kappa) \log(1/\epsilon))$, while the complexity of accelerated deterministic methods including AGD (Nesterov, 1983) and APG (Beck & Teboulle, 2009) is $O(n\sqrt{n} \log(1/\epsilon))$, where $\kappa$ is the condition number. Obviously, the complexities show that the variance reduction methods always converge
faster than accelerated batch methods as long as \( \kappa \leq \mathcal{O}(n^2) \). For non-strongly convex (non-SC) problems, they seem to yield slower convergence rates, e.g., \( \mathcal{O}(1/S) \) for SVRG vs. \( \mathcal{O}(1/S^2) \) for AGD and APG, where \( S \) is the length of the outer-loop or the number of iterations.

The momentum acceleration techniques for deterministic optimization have been widely researched, e.g., the heavy-ball method (Polyak, 1964), Nesterov’s accelerated gradient methods (Nesterov, 1983; 2013) and the optimized gradient method (Kim & Fessler, 2016). Recently, there has been a surge in interest in accelerating stochastic variance reduced methods such as (Frostig et al., 2015; Lin et al., 2015a; Mahdavi et al., 2013; Nitanda, 2014; Allen-Zhu, 2018; Murata & Suzuki, 2017; Hien et al., 2019). Lin et al. (2015a) presented an indirect acceleration (Catalyst) framework, which achieves the complexity of \( \mathcal{O}(n + \sqrt{nL/\epsilon}) \log^2(1/\epsilon) \) for non-SC problems, where \( L \) is a Lipschitz constant. Here, the methods via dummy regularization or reductions are regarded as indirect ones. As the direct acceleration of SVRG, Katyusha (Allen-Zhu, 2018) introduced the idea of negative momentum (i.e., Katyusha momentum). By combining Katyusha momentum with Nesterov’s momentum, Katyusha achieves the complexity \( \mathcal{O}(n + \sqrt{nL/\epsilon}) \log(1/\epsilon) \) for SC problems, which matches the complexity lower bound for minimizing convex finite-sum functions, proved by Lan & Zhou (2018b). Besides, several accelerated methods were proposed, e.g., APCG (Lin et al., 2015b), SDPC (Zhang & Xiao, 2015), Point-SAGA (Defazio, 2016) and RPDG (Lan & Zhou, 2018a). In particular, Allen-Zhu (2018) also proved that Katyusha directly (i.e., Katyusha\(^\text{NA} \)) attains the complexity \( \mathcal{O}(n/\sqrt{\epsilon} + \sqrt{nL/\epsilon}) \) for non-SC problems. Although by using reduction techniques, Katyusha obtains an improved complexity \( \mathcal{O}(n \log(1/\epsilon) + \sqrt{nL/\epsilon}) \), which is still worse than the optimal oracle bound in (Woodworth & Srebro, 2016), i.e., \( \mathcal{O}(n + \sqrt{nL/\epsilon}) \). More recently, Lan et al. (2019) proposed a directly accelerated (Varag) method, which obtains the complexity of \( \mathcal{O}(n \log_2(n) + \sqrt{nL/\epsilon}) \) for non-SC problems. However, similar to Varag, VRADA (Song et al., 2020) has an extra \( \log_2(n) \) factor compared with the complexity lower bound, \( \Omega(n + \sqrt{nL/\epsilon}) \). It is then natural to ask whether there exists a directly accelerated stochastic method that can attain the optimal oracle complexity.

This paper also considers the minimization problem (1) with a structured regularizer \( h(Ax) \), such as graph-guided fused Lasso (Kim et al., 2009), where \( A \in \mathbb{R}^{d_1 \times d} \) is a given matrix. As the generalization of Problem (1), such problems can be formulated as the equality-constrained finite-sum problem,

\[
\min_{x \in \mathbb{R}^{d_1}, w \in \mathbb{R}^{d}} \left\{ f(x) + h(w), \text{ s.t., } Ax = w \right\},
\]

where \( A \in \mathbb{R}^{d_1 \times d} \). In fact, the algorithm proposed in this paper and its convergence result can be extended to the more general problem (2) with the constraint \( Ax + Bw = c \), where \( A \in \mathbb{R}^{d_2 \times d}, B \in \mathbb{R}^{d_2 \times d_1}, c \in \mathbb{R}^{d_2} \). For the SC and equality-constrained problem (2), Suzuki (2014) and Zheng & Kwok (2016) proved that their variance reduction stochastic ADMM methods attain linear convergence for the special (i.e., the constraint in (2) is \( Ax = w \)) and general ADMM forms (i.e., the constraint in (2) becomes \( Ax + Bw = c \)), respectively. In SAG-ADMM and SVRG-ADMM, the convergence rate \( \mathcal{O}(1/S^2) \) can be guaranteed for non-SC problems, which implies that there remains a gap in the convergence rates of between the stochastic ADMM and accelerated batch algorithms, i.e., \( \mathcal{O}(1/S) \) vs. \( \mathcal{O}(1/S^2) \).

For the equality-constrained composite convex problem (2), Xu et al. (2017) proposed a faster variant of SVRG-ADMM with an adaptive penalty parameter scheme. Liu et al. (2021) presented a momentum accelerated variant of SVRG-ADMM (called ASVRG-ADMM), and Li & Lin (2017) proposed an accelerated stochastic ADMM for solv-
Table 2. Comparison of convergence rates and oracle complexities of the stochastic ADMM methods for solving Problem (2), where those of ASVRG-ADMM are obtained with a boundedness assumption on the constraint sets of primal and dual variables (see Section 4.3 for details). Note that we can easily achieve the lower bounds for Problem (2) by using (Xie et al., 2020).

<table>
<thead>
<tr>
<th>Algorithms (Zhong &amp; Kwok, 2014) (Zheng &amp; Kwok, 2016) (Liu et al., 2021)</th>
<th>This paper</th>
<th>Lower Bound (Xie et al., 2020)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convergence rates</td>
<td>$\mathcal{O}(\frac{n}{\epsilon})$</td>
<td>$\mathcal{O}(\frac{n}{\epsilon})$</td>
</tr>
<tr>
<td>Oracle complexities</td>
<td>$\mathcal{O}(\frac{n}{\epsilon} + \frac{L}{\epsilon})$</td>
<td>$\mathcal{O}(\frac{n}{\epsilon} + \frac{L}{\epsilon})$</td>
</tr>
<tr>
<td>Boundedness assumption</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Motivations: (I) For solving the non-SC problem (1) (e.g., $\ell_1$-norm regularized problems), Katyusha (Allen-Zhu, 2018) and Varag (Lan et al., 2019) attain the oracle complexities of $\mathcal{O}(n/\sqrt{\epsilon} + \sqrt{nL}/\epsilon)$ and $\mathcal{O}(n\log_2(n)/\sqrt{\epsilon} + \sqrt{nL}/\epsilon)$, respectively. However, the lower bound of the oracle complexity is $\mathcal{O}(n + \sqrt{nL}/\epsilon)$ (Woodworth & Srebro, 2016). That is, there are some gaps between the convergence results in (Allen-Zhu, 2018; Lan et al., 2019) and the lower bound of $\mathcal{O}(n + \sqrt{nL}/\epsilon)$.

(II) Although by adding a SC proximal term into non-SC problems as in (Frostig et al., 2015; Lin et al., 2015a; Allen-Zhu, 2018), one can achieve faster convergence, this may hurt the performance of the algorithms in both theory and practice (Allen-Zhu & Yuan, 2016). The difficulty for the indirect methods is that it is really hard to choose the proximal parameter properly. **Can we design a simple algorithm for Problem (1) to close the gap in theory?**

(III) Moreover, for solving the non-SC structure-regularized problem (2), Table 2 shows that there is a big gap of convergence rates between prior works and the lower bound in (Xie et al., 2020). **Can we obtain the optimal convergence rate in both theory and practice?**

Our Main Contributions: To fill in the gaps, we propose a novel directly accelerated stochastic variance reduced gradient (DAVIS) method, which has two snapshots and new momentum accelerated rules with a new compensated stochastic gradient operator. We prove that DAVIS obtains an optimal convergence rate, $\mathcal{O}(1/(nS^2))$. Moreover, we prove that the oracle complexity of DAVIS is $\mathcal{O}(n + \sqrt{nL}/\epsilon)$, which is identical to the lower bound in (Woodworth & Srebro, 2016). That is, our oracle complexity is by a factor $1/\sqrt{\epsilon}$ lower than Katyusha, and by a factor $\log_2(n)$ better than Varag, as shown in Table 1. To the best of our knowledge, this is the first directly accelerated method that attains the optimal complexity bound for the non-SC problem (1).

To answer the above-mentioned question, we also extend DAVIS to solve Problem (2) with a structured regularizer. For important emerging equality-constrained non-SC problems (e.g., graph-guided fused Lasso (Kim et al., 2009)), the best-known convergence rate of existing accelerated stochastic ADMM methods such as (Liu et al., 2021) is $\mathcal{O}(1/S^2)$ in the case with an assumption of boundedness on the constraint sets of primal and dual variables (see details in Section 4.3) or $\mathcal{O}(1/S)$ in the case without the assumption of boundedness. Our second main result is that we propose a directly accelerated stochastic ADMM (DAVIS-ADMM) algorithm, and prove that it attains the optimal rate of $\mathcal{O}(1/(nS))$ without the assumption of boundedness, as shown in Table 2. Moreover, DAVIS-ADMM attains the optimal gradient complexity $\mathcal{O}(n + L/\epsilon)$, which matches the lower bound in (Xie et al., 2020). To the best of our knowledge, this is the first method that obtains the optimal theoretical result for stochastic ADMMs.

2. Related Work

Throughout this paper, $\|\cdot\|$ denotes the Euclidean norm. We mostly focus on two types of problems (1) and (2), where each component function is $L$-smooth.

Assumption 2.1 (Smoothness). Each component function $f_i(\cdot)$ is $L$-smooth, i.e., for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|.$$

2.1. Stochastic Variance Reduction Methods for Unconstrained Optimization

In recent years, stochastic variance reduction methods such as (Roux et al., 2012; Shalev-Shwartz & Zhang, 2013; Johnson & Zhang, 2013; Defazio et al., 2014; Xiao & Zhang, 2014) have received extensive attention. Variance reduced gradient estimators such as SVRG (Johnson & Zhang, 2013) and SAGA (Defazio et al., 2014) use gradient information from previous iterates to construct a better estimate of the gradient at the current iterate $x_k$. In particular, the popular SVRG estimator (i.e., $\nabla f_i(x_k)$) in (Johnson & Zhang,
Another possible solution is primal-dual hybrid gradient (PDHG) method. Together with the dual variable \( \lambda \), the update steps of stochastic ADMMs are 

\[
\begin{align*}
    w_k &= \arg \min_w \left\{ b(w) + \phi_k(x_{k-1}, w) \right\}, \\
    x_k &= \arg \min_x \left\{ x^T \nabla f_{ik}(x_{k-1}) + \frac{\| x - x_{k-1} \|^2}{2\eta_k} + \phi_k(x, w_k) \right\}, \\
    \lambda_k &= \lambda_{k-1} + A x_k - w_k,
\end{align*}
\]

where \( \phi_k(x, w) = \frac{\beta}{2} \| A x - w + \lambda_{k-1} \|^2 \), \( \beta > 0 \) is a penalty parameter, \( \eta_k > 0 \) is a parameter, and \( \| x \|_Q^2 = x^T Q x \) with a given positive semi-definite matrix \( Q \) (Ouyang et al., 2013).

Another possible solution is primal-dual hybrid gradient methods such as (Zhu & Chan, 2008; Goldstein et al., 2015). Note that we mainly focus on accelerating stochastic ADMM and refrain from discussing dual and primal-dual stochastic algorithms. Some researchers have adopted the variance reduced techniques mentioned above for ADMM, e.g., (Zhong & Kwok, 2014; Suzuki, 2014; Zheng & Kwok, 2016). More recently, together with variance reduction techniques, some momentum accelerated stochastic ADMM algorithms (Li & Lin, 2017; Xu et al., 2017; Liu et al., 2021) have been proposed for solving Problem (2).
We define a new compensated stochastic variance reduction gradient estimator for our DA VIS algorithm as follows:

$$\tilde{\nabla}_{ik}(x) = \nabla f_{i}(x) - \nabla f_{i}(\tilde{x}^{s-1}) + \nabla f(\tilde{x}^{s-1})$$

SVRG estimator

$$+ m\theta_k(\tilde{x}^{s-1} - \tilde{x}^{s-1})/\eta.$$ Compensated estimator

It is clear that our stochastic variance reduction gradient estimator consists of two terms, i.e., the SVRG estimator independently proposed in (Johnson & Zhang, 2013; Zhang et al., 2013) and a new compensated estimator. Note that the new compensated term is introduced into the proposed gradient estimator $\tilde{\nabla}_{ik}(x)$, and plays a key role to offset the residual term in the upper bound of Lemma 3.3 (see the discussion in Section 3.4 for details).

3.3.2. Momentum Acceleration

We first define a new proximal stochastic gradient decent scheme for our algorithm.

**Definition 3.2 (Prox-SGrad).** The proximal stochastic gradient decent (Prox-SGrad) is defined as:

$$\text{Prox-SGrad}(x) = \arg \min_z \left\{ h(z) + \langle \tilde{\nabla}_{ih}(x), z \rangle + \frac{m\theta_k}{2\eta} \| z - \tilde{x}^s \|^2 \right\}$$

$$= \text{prox}_{\frac{m\theta_k}{2\eta}h} \left( \delta_k^* - \frac{\eta}{m\theta_k} \tilde{\nabla}_{ih}(x) \right),$$

where $\delta_k^* = p_k^* + 2(\tilde{x}^{s-1} - \tilde{x}^{s-1})$, and $p_k^*$ is defined below, and prox is the standard proximal operator as in (Xiao & Zhang, 2014; Allen-Zhu, 2018).

Next we design a new momentum acceleration scheme in each inner loop, and first give a new update rule for $y_k^s$ at the $k$-th iteration of the $s$-th epoch:

$$y_k^s = \frac{\theta}{m} p_k^s + (1 - \frac{\theta}{m}) x_k^s. \quad (4)$$

Here $p_k^s = z_{k-1}^s - \tilde{x}^{s-1} + \tilde{x}^{s-1}$ is designed to get a structured recursive form (i.e., $\| x^s - z_{k-1}^s \|^2 - \| x^s - z_k^s \|^2$) in our upper bound of one-iteration in Lemma 3.4 below. Moreover, we define the following momentum acceleration update rule for $x_k^s$:

$$x_k^s = \frac{\theta}{m} (z_k^s - p_k^s) + y_k^s, \quad (5)$$

where $z_k^s = \text{Prox-SGrad}(y_k^s)$. We give the following upper bound for our stochastic updates in one iteration (e.g., the $k$-th iteration) of Algorithm 1.

### 3.4. Optimal Convergence Guarantees

In this subsection, we analyze the convergence property of our DAVIS algorithm. Theorem 3.6 below shows that...
DAVIS can improve the best-known convergence rate of some accelerated methods (e.g., Katyusha) from $O(1/S^2)$ to $O(1/(nS^2))$ for the non-SC problem (1). We also suppose that the distance between an initial point $\bar{x}^0$ and an optimal solution $x^*$ can be bounded by a constant $c$, i.e., $\|x^* - \bar{x}^0\| \leq c$, which is a basic condition (see Appendix B for our proof sketch and detailed proofs).

3.4.1. Core Lemmas

The proof of our main result relies on the one-iteration inequality in Lemma 3.5 below, which is a key lemma to obtain our theoretical result in Theorem 3.6. Lemma 3.5 consists of two key upper bounds in Lemmas 3.3 and 3.4. By using our double snapshot scheme in each outer loop of Algorithm 1, we can obtain the following result.

**Lemma 3.3 (Upper bound of double snapshot update).** Suppose that Assumption 2.1 holds. Let $\{\pi^s\}$ be the sequence generated by our double snapshot scheme in Algorithm 1, we have

$$F(\pi^{s-1}) - F(x^*) \leq (1 - \theta_s)(F(\pi^{s-1}) - F(x^*)) + R^s + \frac{\theta^2_s}{2m\eta} \|x^* - \pi^{s-1}\|^2 - \|x^* - \pi^{s-1}\|^2,$$

where $R^s = \left(\frac{\theta^2_s}{2m\eta} - \frac{\theta^2_s}{2m\eta}\right)\|\pi^{s-1} - \pi^{s-1}\|^2$.

Note that the upper bound in Lemma 3.3 has an additional term, which may be positive but we shall cancel it by using the upper bound of our stochastic update schemes in Lemma 3.4 below. Using our stochastic momentum accelerated scheme in each inner loop, we give the following result.

**Lemma 3.4 (Upper bound of one-iteration).** Suppose that Assumption 2.1 holds. Let $\{x^s_k, z^s_k\}$ be the sequence generated by Algorithm 1. Then

$$\mathbb{E}[F(x^s_k) - F(x^*)] \leq \frac{\theta^2_s}{2m\eta}(\|x^* - z^{s-1}_k\|^2 - \|x^* - z^{s-1}_k\|^2) + C^s + \mathbb{E}\left[1 - \frac{\theta_s}{m}(F(\pi^{s-1}) - F(x^*))\right],$$

where $C^s = \left(\frac{\theta^2_s}{2m\eta} - \frac{\theta^2_s}{2m\eta}\right)\|\pi^{s-1} - \pi^{s-1}\|^2$.

**Remark 1.** In Lemma 3.4, the term $C^s$ is produced by our gradient estimator, and is used to compensate the additional term $R^s$ in Lemma 3.3 (i.e., $R^s + C^s = 0$). As we expected, the designed stochastic descent step can be used to counter-act the additional term, as shown in the detailed proof for Lemma 3.5 below.

Using Lemmas 3.3 and 3.4, we give the following upper bound of one iteration in Algorithm 1.

**Lemma 3.5 (Upper bound of one-epoch).** Suppose that Assumption 2.1 holds. Let $\{x^s_k\}$ be the sequence generated by Algorithm 1. Then we have

$$\mathbb{E}[F(\bar{x}^s) - F(x^*)] \leq \left(1 - \theta_s\right)\left(F(\bar{x}^{s-1}) - F(x^*)\right) + \frac{\theta^2_s}{2m\eta}(\|x^* - \bar{x}^{s-1}\|^2 - \|x^* - \bar{x}^{s-1}\|^2).$$

3.4.2. Optimal Convergence Results

We can obtain the inequality of one-epoch by using Lemma 3.5, and telescope the inequality over all epochs to obtain the following result.

**Theorem 3.6.** Suppose that each component function $f_i(\cdot)$ is $L$-smooth. Let $\bar{x}^s = \frac{1}{m}\sum_{k=1}^m x^s_k$ (i.e., the average point of the previous epoch), then the following result holds

$$\mathbb{E}[F(\bar{x}^s) - F(x^*)] \leq O\left(\frac{L\|x^* - \bar{x}^0\|^2}{mS^2}\right).$$

Choosing $m = \Theta(n)$, Algorithm 1 achieves an $\epsilon$-suboptimal solution using at most $O(n + \sqrt{nL/\epsilon})$ iterations.

**Remark 2.** Theorem 3.6 shows that DAVIS achieves the optimal convergence rate $O(1/(nS^2))$, while most existing directly accelerated methods including (Allen-Zhu, 2018; Zhou et al., 2018) attain the rate $O(1/S^2)$. In particular, Algorithm 1 also attains the optimal oracle complexity $O(n + \sqrt{nL/\epsilon})$, which matches the lower bound in (Woodworth & Srebro, 2016). In contrast, the complexities of Katyusha (Allen-Zhu, 2018), Varag (Lan et al., 2019) and VRADA (Song et al., 2020) are $O(n/\sqrt{\epsilon} + \sqrt{nL/\epsilon})$, $O(n \log_2(n) + \sqrt{nL/\epsilon})$ and $O(n \log_2(n) + \sqrt{nL/\epsilon})$, respectively. In other words, DAVIS has both the optimal oracle complexity and optimal convergence rate for solving Problem (1). To the best of our knowledge, this is the first time that the optimal complexity bound is obtained through a directly accelerated algorithm for general convex finite-sum optimization in the literature.

3.5. Comparison with Existing Algorithms

There are some main differences between our DAVIS algorithm and existing accelerated stochastic algorithms such as Katyusha (Allen-Zhu, 2018), Varag (Lan et al., 2019), and VRADA (Song et al., 2020).

- One vs two snapshots: Both snapshots (i.e., $\bar{x}^s$ and $\pi^s$) are introduced into our update rules in (4) and (5), while only one snapshot $\bar{x}^s$ is applied in most algorithms such as Katyusha, Varag and VRADA. Similarly, we use double snapshots for each variable in equality-constrained problems.

- SVRG estimator vs our compensated estimator: Most accelerated algorithms such as Katyusha, Varag and
Algorithm 2 DAVIS-ADMM for Problem (2)

Input: \( S \) and \( m \).

Initialize: \( \bar{x}^0, \bar{w}^0, \bar{\lambda}^0, z_1^0, \theta_1 = 1, \) and \( \eta \).

1: for \( s = 1, 2, \ldots, S \) do
2: Update the snapshots \( \bar{x}^s-1, \bar{w}^s-1 \) and \( \bar{\lambda}^s-1 \) via (6);
3: Compute the full gradient at the snapshot \( \bar{x}^s-1 \),
4: \( \nabla f(\bar{x}^s-1) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\bar{x}^s-1) \);
5: for \( k = 1, 2, \ldots, m \) do
6: \( w_k^s = \arg \min_w \left\{ h(w) + \frac{\beta}{2m}\| A z_k^s - w + \lambda_k^s - 1 \|^2 \right\} \);
7: \( y_k^s = \frac{\theta}{m} p_k^s + \left( 1 - \frac{\theta}{m} \right) \bar{x}^s-1 \);
8: \( \hat{\nabla} I_k(y_k^s) = g_{I_k}(y_k^s) + \frac{m\theta}{m} Q_\eta(\bar{x}^s-1 - \bar{x}^{s-1}) \);
9: \( z_k^s = \arg \min_z \left\{ \langle \hat{\nabla} I_k(y_k^s), z \rangle + \phi_\eta(z, w_k^s) + \frac{m\theta}{m} \| z - p_k^s - 2(\bar{x}^s-1 - \bar{x}^{s-1}) \|^2 \right\} \);
10: \( x_k^s = \frac{\theta}{m} z_k^s + \left( 1 - \frac{\theta}{m} \right) \bar{x}^{s-1} \);
11: \( \lambda_k^s = A z_k^s - w_k^s + \lambda_k^{s-1} \);
12: end for
13: \( z^{s+1} = z_m^s \), \( \lambda^{s+1} = \lambda_m^s \), \( \theta_s = \sqrt{\theta^4_s + 4\theta^2_s - \theta^2_s} \);
14: end for
15: Output: \( \bar{x}^S, \bar{w}^S \).

4. An Optimal Stochastic ADMM Algorithm for Equality Constrained Optimization

In this section, we propose a novel directly accelerated stochastic variance reduction ADMM (DAVIS-ADMM) algorithm with double-snapshots to solve Problem (2) (see Algorithm 2 for the details of DAVIS-ADMM). The proposed DAVIS-ADMM algorithm improves the best-known convergence rate from O(1/S^2) to O(1/(ns)) . We first let \( Q_s = \gamma I - \frac{y_s}{\sqrt{T}} A^T A \) with \( \gamma \geq \eta \| A^T A \|_2 + 1 \) to ensure that \( Q_s \geq I_n \), where \( \| \cdot \|_2 \) is the spectral norm, and \( \| C \|_2 \) is the spectral norm of a matrix \( C \).

4.1. Double-Snapshot Scheme in DAVIS-ADMM

Existing stochastic variance reduction ADMMs such as SVRG-ADMM (Zheng & Kwok, 2016) and ASVRG-ADMM (Liu et al., 2021) use the three snapshots \( \bar{x}^s, \bar{w}^s \) and \( \bar{\lambda}^s \), while three additional snapshots \( \bar{x}^s, \bar{w}^s \) and \( \bar{\lambda}^s \) are also designed for our DAVIS-ADMM algorithm. Note that all the snapshots in our DAVIS-ADMM algorithm are updated only in outer loop. More specifically, the update rules of the first three snapshots are defined as: \( \bar{x}^s = \frac{1}{m} \sum_{k=1}^{m} z_k^s \), \( \bar{w}^s = \frac{\theta}{m} p^s + (1 - \frac{\theta}{m}) \bar{w}^{s-1} \). And the update rules of the three new snapshots in our DAVIS-ADMM are given as follows:

\[
\begin{align*}
\bar{x}^{s-1} &= \theta_s z^{s-1} + (1 - \theta_s) \bar{x}^{s-1}, \\
\bar{w}^{s-1} &= \theta_s p^{s-1} + (1 - \theta_s) \bar{w}^{s-1}, \\
\bar{\lambda}^{s-1} &= A \bar{z}^{s-1} - p^{s-1} + \bar{\lambda}^{s-2}.
\end{align*}
\]

Let \( \phi(z, w) = \frac{1}{\alpha_m} \| A z - w + \lambda^{s-2} \|^2 \), the auxiliary variables are defined as follows: \( p^{s-1} = \arg \min_p \left\{ h(p) + \frac{\beta}{2m} \right\} \).

Closing the Convergence Gaps in Non-Strongly Convex Optimization

VRADA use the SVRG estimator proposed in (Johnson & Zhang, 2013; Zhang et al., 2013). In contrast, our gradient estimator is introduced to compensate the residual term in Lemma 3.3, and thus is one of main contributions of this paper.

- Negative momentum vs compensated momentum: Both Varag and Katyusha use the negative momentum proposed in (Allen-Zhu, 2018). By combining with our defined gradient estimator, we also design a new momentum accelerated rule in (5) to achieve an optimal convergence rate. Note that \( \hat{\theta} \) can be defined as: \( \hat{\theta} = \frac{2}{s+1} \) satisfying \( \frac{1}{\hat{\theta}^2} \geq \frac{1}{\theta^2} \). Different from Katyusha, the coefficient of our momentum is \( \theta_s / m \), which can reduce the impact of the variance bound on the upper bound of one iteration in Lemma 3.4 from \( O(1) \) to \( O(1/m) \). Moreover, both Varag and DAVIS only require the solution of one subproblem in each iteration instead of two subproblems in Katyusha.
4.3. Optimal Convergence Rate for Problem (2)

We analyze the convergence property of DAVIS-ADMM. Let \( x^* \) be an optimal solution of Problem (2), \( w^* \) and \( \lambda^* \) be the corresponding solutions, we give the convergence criterion (Zheng & Kwok, 2016) for our analysis.

**Definition 4.2** (Convergence criterion). Given a constant \( \delta \geq 0 \), a nonnegative convergence criterion is defined as: 
\[
\phi(\tilde{x}^S, \tilde{w}^S) = \mathcal{P}(\tilde{x}^S, \tilde{w}^S) + \delta \|Ax^S - \tilde{w}^S\| \geq 0,
\]
where \( \mathcal{P}(x, w) = f(x) - f(x^*) - \nabla f(x^*)^T(x - x^*) + h(w) - h(w^*) - \nabla h(w^*)^T(w - w^*) \), and \( \nabla h(w) \) is the (sub)gradient of \( h(\cdot) \) at \( w \).

Previous work such as (He & Yuan, 2012; Azadi & Sra, 2014) requires the assumption of boundedness on the constraint sets of primal and dual variables (i.e., suppose \( x \in \mathcal{X} \) and \( \lambda \in \Lambda \), where \( \mathcal{X} \) and \( \Lambda \) are compact convex sets with diameters \( D_X = \sup_{x_1, x_2 \in \mathcal{X}} \|x_1 - x_2\| \) and \( D_\Lambda = \sup_{\lambda_1, \lambda_2 \in \Lambda} |\lambda_1 - \lambda_2| \), respectively) when proving the convergence of ADMMs, and ASVRG-ADMM (Liu et al., 2021) obtains the convergence rate \( O(1/S^2) \) with the boundedness assumption. However, we can remove such assumption, which is in fact a strong assumption, and provide the following convergence result (see Appendix C for our proof sketch and detailed proofs).

**Theorem 4.3.** Suppose Assumption 2.1 holds. Let the constant \( c_1 = 2\|AA^T\|_2 \|x^* - \tilde{x}^0\|^2 + 2\|\lambda^* - \tilde{\lambda}^0\|^2 + 8\delta^2 + 10\|\lambda^*\|^2 \) and choose \( m = \Theta(n) \), then
\[
\mathbb{E}[\phi(\tilde{x}^S, \tilde{w}^S)] \leq O\left(\frac{2\phi(x^0, \tilde{w}^0) + \|ar{x}^* - \tilde{x}^0\|^2}{n(S+1)} + \frac{c_1 \beta}{n(S+1)}\right).
\]

**Remark 3.** Theorem 4.3 shows that without the boundedness assumption, the convergence rate of DAVIS-ADMM is \( O(1/(ns)) \), while the best-known convergence result as in (Liu et al., 2021) with a strong boundedness assumption is \( O(1/S^2) \), as shown in Table 2. That is, DAVIS-ADMM improves the best-known convergence rate from \( O(1/S^2) \) to \( O(1/(ns)) \) without the boundedness assumption, which matches the lower bound in (Xie et al., 2020). The upper bound in Theorem 4.3 only relies on the constant \( c_1 \), while the theoretical result of ASVRG-ADMM (Liu et al., 2021) requires that \( \mathcal{X} \) and \( \Lambda \) are bounded with the diameters \( D_X \) and \( D_\Lambda \). Moreover, our DAVIS-ADMM algorithm and convergence results can be extended to the deterministic setting. When the mini-batch size is \( b = n \), and \( m = 1 \), DAVIS-ADMM degenerates to its deterministic version, and the convergence rate of our deterministic DAVIS-ADMM algorithm becomes \( O(1/S) \), which is consistent with the optimal convergence rate of accelerated deterministic methods such as (Ouyang & Xu, 2020).

5. Experimental Results

In this section, we evaluate the performance of our algorithms for solving the non-SC problems (1) and (2), and the detailed experimental setup is given in Appendix D.

5.1. \( \ell_1 \)-Norm Regularized Logistic Regression

We first apply Algorithm 1 to solve the binary non-strongly convex \( \ell_1 \)-norm regularized logistic regression problem,
\[
\min_x \left\{ \frac{1}{n} \sum_{i=1}^{n} \log \left( 1 + \exp \left( -b_i a_i^T x \right) \right) + \lambda \|x\|_1 \right\},
\]
where \( a_i \in \mathbb{R}^d, b_i \in \{-1, +1\}, i = 1, 2, \ldots, n \). We also compare our DAVIS algorithm with the state-of-the-art accelerated methods such as SVRG++ (Allen-Zhu & Yuan, 2016), Katyusha (Allen-Zhu, 2018) and Varag (Lan et al., 2019). The experimental results of all the accelerated methods on Adult and Covtype are shown in Figure 1, where the regularization parameter is \( 10^{-5} \). We observe that our DAVIS consistently outperforms other accelerated methods, which empirically verifies our theoretical result.

5.2. Graph-Guided Fused Lasso

We also evaluate the performance of DAVIS-ADMM for solving the graph-guided fused lasso problem:
\[
\min_x \left\{ \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \lambda \|Ax\|_1 \right\},
\]
where \( f_i(\cdot) \) is the logistic loss function, \( \lambda \geq 0 \) is the regularization parameter, \( A = [G; I] \), and \( G \) is the sparsity pattern of the graph obtained by sparse inverse covariance selection as in (Banerjee et al., 2008).

Figure 2 shows the experimental results of SVRG-ADMM (Zheng & Kwok, 2016), ASVRG-ADMM (Liu et al., 2021) and DAVIS-ADMM. All the results show that the accelerated methods (i.e., ASVRG-ADMM and DAVIS-ADMM) outperform the non-accelerated stochastic ADMM method, i.e., SVRG-ADMM. In particular, DAVIS-ADMM converges much faster than the other methods, which is consistent with our convergence guarantee.
Figure 2. Comparison of all the methods for solving graph-guided fused Lasso problems on Adult and Covtype, where the regularization parameter is $\lambda = 10^{-5}$.

6. Conclusions

In this paper, we proposed two efficient directly optimal stochastic variance reduction algorithms for unconstrained and equality-constrained finite-sum problems, respectively. The proposed algorithms have simple update rules, and thus their per-iterations have similar computational costs as existing accelerated methods (e.g., Katyusha and ASVRG-ADMM) for the two classes of optimization problems, respectively. In particular, we theoretically analyzed the convergence properties of our algorithms, and our theoretical results show that our algorithms obtain the optimal convergence rates and optimal oracle complexities for both non-SC unconstrained and equality-constrained problems, respectively. They are also identical to the lower bounds provided in (Woodworth & Srebro, 2016; Xie et al., 2020). That is, our algorithms are a factor $n$ faster than both existing accelerated algorithms (e.g., Katyusha) for Problem (1), i.e., $O(1/nS^2)$ vs. $O(1/S^2)$, and a factor $\frac{n}{S}$ faster than accelerated stochastic ADMM algorithms for Problem (2), i.e., $O(1/nS)$ vs. $O(1/S^2)$.

Our directly accelerated algorithms (including DAVIS and DAVIS-ADMM) can be easily extended to solve more complex problems, e.g., the stochastic nested composition optimization problem (Wang et al., 2017). As our future work, our method can allow the applications of non-uniform sampling and non-Euclidean Bregman distance for solving more different types of optimization problems.

Acknowledgements

We thank all the reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China (Nos. 61876221, 61876220, 61976164, and 61836009), and Natural Science Basic Research Program of Shaanxi (Program No. 2022GY-061). Z. Lin was supported by the major key project of PCL (grant no. PCL2021A12), the NSF China (No. 61731018), and Project 2020BD006 supported by PKU-Baidu Fund.

References


Appendix for “Kill a Bird with Two Stones: Closing the Convergence Gaps in Non-Strongly Convex Optimization by Directly Accelerated SVRG with Double Compensation and Snapshots”

Appendix A: Preliminaries

Before giving the convergence analysis of our algorithms, we first present the following property and lemma.

**Lemma 4.** ((Allen-Zhu, 2018)) The variance reduction stochastic gradient estimator proposed in (Johnson & Zhang, 2013; Zhang et al., 2013) is defined as:

\[
\tilde{\nabla} f_i(x_k) = \nabla f_i(x_k) - \nabla f_i(x_{k-1}) + \nabla f(x_{k-1}).
\]

Suppose that each \( f_i(x) \) is convex and \( L \)-smooth, then the following inequality holds

\[
\mathbb{E}\left[ \|\tilde{\nabla} f_i(x_k) - \nabla f(x_k)\|^2 \right] \leq 2L\left[ f(x_{k-1}) - f(x_k) + \langle \nabla f(x_k), x_k - x_{k-1} \rangle \right].
\]

The convergence analysis for the proposed algorithms requires the above upper bound on the term \( \mathbb{E}[\|\tilde{\nabla} f_i(x_k) - \nabla f(x_k)\|^2] \) as in (Allen-Zhu, 2018). Moreover, we need to extend the expected variance upper bound in Lemma 4 to the mini-batch setting (see Lemma 7 below).

**Property 1.** Given any \( x_1, x_2, x_3, x_4 \in \mathbb{R}^d \), then we have

\[
\langle x_1 - x_2, x_1 - x_3 \rangle = \frac{1}{2}(\|x_1 - x_2\|^2 + \|x_1 - x_3\|^2 - \|x_2 - x_3\|^2)
\]

\[
\langle x_1 - x_2, x_3 - x_4 \rangle = \frac{1}{2}(\|x_1 - x_4\|^2 - \|x_1 - x_3\|^2 + \|x_2 - x_3\|^2 - \|x_2 - x_4\|^2).
\]

Appendix B: Proofs for Section 3

In this section, we give detailed proofs for the convergence analysis of DAVIS (i.e., Algorithm 1), which mainly include the proofs for Lemmas 3.3, 3.4 and 3.5, and Theorem 3.6 in Section 3.4 in the main paper.

Now we sketch the proof of Theorem 3.6 as follows: The proof of Theorem 3.6 relies on telescoping the upper bound of one-epoch in Lemma 3.5. Moreover, Lemmas 3.3 and 3.4 in the main paper play a key role for obtaining the upper bound of one-epoch in Lemma 3.5. That is, we first give the upper bound in Lemma 3.3 by using the proposed double snapshot scheme in Algorithm 1, and the residual term \( R \) is also produced. For each inner loop of Algorithm 1, we obtain the upper bound of one-iteration in Lemma 3.5 by using both the proposed momentum acceleration scheme and the compensated stochastic gradient estimator. As a result, the compensated term \( C \) is introduced in the upper bound in Lemma 3.4, which can be used to offset by the residual term \( R \) in Lemma 3.3. Therefore, we obtain a tight upper bound of one-epoch in Lemma 3.5 by using Lemmas 3.3 and 3.4.
Proof of Lemma 3.3 (Upper bound of double snapshot update)

Proof. We first recall the following iteration scheme of our deterministic gradient descent step,

$$\tilde{x}^{s-1} = \arg\min_z \left\{ h(z) + \langle \nabla f(\tilde{x}^{s-1}), z \rangle + \frac{\theta_s}{2m\eta} \| z - \tilde{x}^{s-1} \|^2 \right\},$$

and $\tilde{x}^{s-1}$ is required to satisfy the following optimal condition,

$$\nabla f(\tilde{x}^{s-1}) + \xi + \frac{\theta_s}{m\eta} (\tilde{x}^{s-1} - \tilde{x}^{s-1}) = 0,$$

(8)

where $\xi \in \partial h(\tilde{x}^{s-1})$ is a sub-gradient of $h(\cdot)$ at $\tilde{x}^{s-1}$.

Since $f(\cdot)$ is $L$-smooth, and $\tilde{x}^{s-1} = \theta_s \tilde{x}^{s-1} + (1 - \theta_s)\tilde{x}^{s-1}$, the following results hold

$$F(\tilde{x}^{s-1}) = h(\tilde{x}^{s-1}) + f(\tilde{x}^{s-1})$$

$$\leq h(\tilde{x}^{s-1}) + f(\tilde{x}^{s-1}) + \langle \nabla f(\tilde{x}^{s-1}), \tilde{x}^{s-1} - \tilde{x}^{s-1} \rangle + \frac{L}{2}\| \tilde{x}^{s-1} - \tilde{x}^{s-1} \|^2$$

$$\leq h(\tilde{x}^{s-1}) + f(\tilde{x}^{s-1}) + \theta_s \langle \nabla f(\tilde{x}^{s-1}), \tilde{x}^{s-1} - \tilde{x}^{s-1} \rangle + \frac{\theta_s^2}{2\eta} \| \tilde{x}^{s-1} - \tilde{x}^{s-1} \|^2$$

$$\quad - \theta_s \langle \nabla f(\tilde{x}^{s-1}), x^* - \tilde{x}^{s-1} \rangle$$

$$= h(\tilde{x}^{s-1}) + f(\tilde{x}^{s-1}) + \theta_s \langle \nabla f(\tilde{x}^{s-1}), \tilde{x}^{s-1} - \tilde{x}^{s-1} \rangle + \frac{\theta_s^2}{2\eta} \| \tilde{x}^{s-1} - \tilde{x}^{s-1} \|^2$$

(9)

$$\quad + \theta_s \left( \xi + \frac{\theta_s}{m\eta} (\tilde{x}^{s-1} - \tilde{x}^{s-1}), x^* - \tilde{x}^{s-1} \right)$$

$$\leq \theta_s F(x^*) + (1 - \theta_s) F(\tilde{x}^{s-1}) + \frac{\theta_s^2}{2m\eta} (\| x^* - \tilde{x}^{s-1} \|^2 - \| x^* - \tilde{x}^{s-1} \|^2)$$

$$\quad + \frac{\theta_s^2}{2\eta} \| \tilde{x}^{s-1} - \tilde{x}^{s-1} \|^2,$$

where the first inequality holds due to the smoothness of $f(\cdot)$, the third equality holds due to the optimal condition in Eq. (8) and Property 1, and the last inequality holds due to the convexities of $h(\cdot)$ and $f(\cdot)$, and the following fact that

$$\frac{\theta_s^2}{m\eta} \langle \tilde{x}^{s-1} - \tilde{x}^{s-1}, x^* - \tilde{x}^{s-1} \rangle = \frac{\theta_s^2}{2m\eta} (\| x^* - \tilde{x}^{s-1} \|^2 - \| x^* - \tilde{x}^{s-1} \|^2 - \| \tilde{x}^{s-1} - \tilde{x}^{s-1} \|^2).$$

Note that the above equality holds due to Property 1.

Therefore, we have

$$F(\tilde{x}^{s-1}) - F(x^*)$$

$$\leq (1 - \theta_s)(F(\tilde{x}^{s-1}) - F(x^*))+ \frac{\theta_s^2}{2m\eta} (\| x^* - \tilde{x}^{s-1} \|^2 - \| x^* - \tilde{x}^{s-1} \|^2) + R^s.$$

(10)

This completes the proof.
Proof of Lemma 3.4 (Upper Bound of One-iteration)

In this subsection, we will prove the upper bound for our stochastic gradient descent step in each inner loop of Algorithm 1. We first recall the main update rules and the optimal condition in our stochastic gradient descent step (i.e., for a fixed $k$).

Let $g^s_k = \nabla f_i(y^s_k) - \nabla f_i(x^{s-1}) + \nabla f(x^{s-1})$, our compensated stochastic variance reduction gradient estimator is rewritten as follows:

$$\tilde{\nabla}_i g^s_k = g^s_k + \frac{m\theta_s}{\eta} (x^{s-1} - \bar{x}^{s-1}) .$$

And the update rule of $z^s_k$ is

$$z^s_k \triangleq \arg \min_z \left\{ h(z) + \left\langle \tilde{\nabla}_i g^s_k, z \right\rangle + \frac{m\theta_s}{2\eta} \| z - p^s_k - 2(z^{s-1} - \bar{x}^{s-1}) \|^2 \right\} ,$$

which implies that $z^s_k$ is required to satisfy the following optimal condition:

$$\tilde{\nabla}_i g^s_k + \zeta^s_k + \frac{m\theta_s}{2\eta} [z^s_k - p^s_k - 2(z^{s-1} - \bar{x}^{s-1})] = 0 , \tag{11}$$

where $\zeta^s_k \in \partial h(z^s_k)$.

Moreover, the main update rules of our stochastic gradient descent step are defined as follows:

$$y^s_k = \frac{\theta_s}{m} p^s_k + \left( 1 - \frac{\theta_s}{m} \right) x^{s-1} ,$$

$$x^s_k = \frac{\theta_s}{m} (z^s_k - p^s_k) + y^s_k \tag{12}$$

Below we give the detailed proof of Lemma 3.4 in the main paper.

Proof of Lemma 3.4:
Proof. Using the smoothness of $f(\cdot)$, we get

\[
\mathbb{E}[F(x_k^s)] \\
\leq \mathbb{E}\left[h(x_k^s) + f(y_k^s) + \langle \nabla f(y_k^s), x_k^s - y_k^s \rangle + \frac{L}{2}\|x_k^s - y_k^s\|^2\right] \\
\equiv \mathbb{E}\left[h(x_k^s) + f(y_k^s) + \left\langle \nabla f(y_k^s), \frac{\theta_s}{m} (z_k^s - p_k^s) \right\rangle + \frac{L \theta_s^2}{2m^2}\|z_k^s - p_k^s\|^2\right] \\
= \mathbb{E}\left[h(x_k^s) + f(y_k^s) + \frac{\theta_s}{m} \left\langle \nabla i_k(y_k^s), z_k^s - p_k^s \right\rangle\right] \\
+ \mathbb{E}\left[\left\langle \nabla f(y_k^s) - \nabla i_k(y_k^s), \frac{\theta_s}{m} (z_k^s - p_k^s) \right\rangle + \frac{L \theta_s^2}{2m^2}\|z_k^s - p_k^s\|^2\right] \\
= \mathbb{E}\left[h(x_k^s) + f(y_k^s) + \frac{\theta_s}{m} \left\langle \nabla i_k(y_k^s), x^* - p_k^s \right\rangle - \frac{\theta_s}{m} \left\langle \nabla i_k(y_k^s), x^* - z_k^s \right\rangle\right] \\
+ \mathbb{E}\left[\left\langle \nabla f(y_k^s) - \nabla i_k(y_k^s), \frac{\theta_s}{m} (z_k^s - p_k^s) \right\rangle + \frac{L \theta_s^2}{2m^2}\|z_k^s - p_k^s\|^2\right] \\
\equiv \mathbb{E}\left[h(x_k^s) + f(y_k^s) + \frac{\theta_s}{m} \left\langle \nabla i_k(y_k^s), x^* - p_k^s \right\rangle\right] \\
+ \frac{\theta_s^2}{2m} \left(\|x^* - p_k^s - 2(\bar{x}^{s-1} - \bar{x}^{s-1})\|^2 - \|x^* - z_k^s\|^2 - \|z_k^s - p_k^s - 2(\bar{x}^{s-1} - \bar{x}^{s-1})\|^2\right) \\
+ \mathbb{E}\left[\left\langle \nabla f(y_k^s) - \nabla i_k(y_k^s), \frac{\theta_s}{m} (z_k^s - p_k^s) \right\rangle + \frac{L \theta_s^2}{2m^2}\|z_k^s - p_k^s\|^2\right] \\
= \mathbb{E}\left[h(x_k^s) + \frac{\theta_s}{m} \left\langle \zeta_k^s - z_k^s + \frac{\theta_s}{m} \left\langle \zeta_k^s, x^* - z_k^s \right\rangle + \frac{\theta_s}{m} \langle \nabla i_k(y_k^s), x^* - z_k^s \rangle \right\rangle + \frac{L \theta_s^2}{2m^2}\|z_k^s - p_k^s\|^2\right] \\
+ \mathbb{E}\left[\left\langle \nabla f(y_k^s) - \nabla i_k(y_k^s), \frac{\theta_s}{m} (z_k^s - p_k^s) \right\rangle - \frac{L \theta_s^2}{2m^2}\|z_k^s - p_k^s\|^2, \right]
\]  

(13)

where the first inequality follows from the smoothness of $f(\cdot)$ (i.e., $f(x_k^s) \leq f(y_k^s) + \langle \nabla f(y_k^s), x_k^s - y_k^s \rangle + \frac{L}{2}\|x_k^s - y_k^s\|^2$); the equality $\equiv$ holds due to the fact that $x_k^s = \frac{\theta_s}{m} (z_k^s - p_k^s) + y_k^s$; and the equality $\equiv$ holds due to the optimal condition in Eq. (11) and Property 1, that is,

\[
\frac{\theta_s}{m} \langle \nabla i_k(y_k^s), x^* - z_k^s \rangle \\
= \frac{\theta_s^2}{\eta} \langle z_k^s - p_k^s - 2(\bar{x}^{s-1} - \bar{x}^{s-1}), x^* - z_k^s \rangle + \frac{\theta_s}{m} \langle \zeta_k^s, x^* - z_k^s \rangle \\
= \frac{\theta_s^2}{2\eta} \left(\|x^* - p_k^s - 2(\bar{x}^{s-1} - \bar{x}^{s-1})\|^2 - \|x^* - z_k^s\|^2 - \|z_k^s - p_k^s - 2(\bar{x}^{s-1} - \bar{x}^{s-1})\|^2\right) \\
+ \frac{\theta_s}{m} \langle \zeta_k^s, x^* - z_k^s \rangle.
\]

Next we need to bound the terms $A_1$, $A_2$, and $A_3$ in the inequality (13). And we first bound the term $A_1$. Using the update
where the first inequality holds due to the convexity of $h(\cdot)$, and the second inequality follows from the facts that $\zeta_k \in \partial h(z_k)$ and $\langle \zeta_k, x^* - z_k \rangle \leq h(x^*) - h(z_k)$.

Using the definitions of $\nabla_i(x_k^s)$, $p_k^s$ and Property 1, the term $A_2$ in the inequality (13) is rewritten as follows:

$$A_2 = E \left[ \frac{\theta_s}{m} \langle \nabla_i(x_k^s), x^* - p_k^s \rangle \right]$$

$$= \frac{\theta_s}{m} \langle \nabla f(y_k^s) + \frac{m \theta_s}{\eta} (\bar{x}^s - \bar{x}^s), x^* - p_k^s \rangle$$

$$= \frac{\theta_s}{m} \langle \nabla f(y_k^s), x^* - p_k^s \rangle + \frac{\theta_s^2}{2\eta} E \left[ \| x^* - z_{k-1}^s \|^2 - \| x^* - p_k^* - 2(\bar{x}^s - \bar{x}^s) \|^2 + \| \bar{x}^s - \bar{x}^s \|^2 \right].$$

Furthermore, we give the upper bound of the term $A_3$ in the inequality (13) as follows:

$$A_3 = E \left[ \frac{\theta_s}{m} \langle \nabla f(y_k^s) - \nabla_i(y_k^s), z_k^s - p_k^s \rangle \right]$$

$$= \frac{\theta_s}{m} E \left[ \langle \nabla f(y_k^s) - \nabla_i(y_k^s), z_k^s - p_k^s - 2(\bar{x}^s - \bar{x}^s) \rangle \right]$$

$$+ \frac{\theta_s}{m} E \left[ \langle \nabla f(y_k^s) - \nabla_i(y_k^s), 2(\bar{x}^s - \bar{x}^s) \rangle \right]$$

$$\leq \frac{\theta_s}{m} E \left[ \langle \nabla f(y_k^s) - \nabla_i(y_k^s), z_k^s - p_k^s - 2(\bar{x}^s - \bar{x}^s) \rangle \right] - \frac{2\theta_s^2}{\eta} \| \bar{x}^s - \bar{x}^s \|^2$$

$$\leq \frac{\eta}{2m^2} \left\| \nabla f(y_k^s) - \nabla_i(y_k^s) \right\|^2 + \frac{\theta_s^2}{2\eta} \| z_k^s - p_k^s - 2(\bar{x}^s - \bar{x}^s) \|^2 - \frac{2\theta_s^2}{\eta} \| \bar{x}^s - \bar{x}^s \|^2$$

$$\leq \frac{1}{m} \left[ f(\bar{x}^s) - f(y_k^s) + (\nabla f(y_k^s), y_k^s - \bar{x}^s) \right]$$

$$+ \frac{\theta_s^2}{2\eta} \| z_k^s - p_k^s - 2(\bar{x}^s - \bar{x}^s) \|^2 - \frac{3\theta_s^2}{2\eta} \| \bar{x}^s - \bar{x}^s \|^2,$$

where the equality $\triangleq$ holds due to the definition of the gradient estimator in Definition 1 and the facts that $E[\nabla f(y_k^s) - g_k^s] = 0$.
and
\[
\frac{\theta_s}{m} E \left[ \left( \nabla f(y_k^*) - \nabla \tilde{u}_k(y_k^*), 2(\pi^{s-1} - \bar{x}^{s-1}) \right) \right] \\
= \frac{\theta_s}{m} E \left[ \left( \nabla f(y_k^*) - g_k^* - \frac{m \theta_s}{\eta} (\pi^{s-1} - \bar{x}^{s-1}), 2(\pi^{s-1} - \bar{x}^{s-1}) \right) \right] \\
= - \frac{2\theta_s^2}{\eta} \| \pi^{s-1} - \bar{x}^{s-1} \|^2.
\]

Note that the equality \( b \leq \) in (16) follows from the Young’s inequality, and the inequality \( c \leq \) in (16) holds due to the following fact
\[
\frac{\eta}{2m^2} E \left[ \| \nabla f(y_k^*) - \nabla \bar{u}_k(y_k^*) \|^2 \right] \\
= \frac{\eta}{2m^2} E \left[ \| \nabla f(y_k^*) - g_k^* \|^2 \right] + \frac{\theta_s^2}{2\eta} E \left[ \| \pi^{s-1} - \bar{x}^{s-1} \|^2 \right] \\
+ \frac{\theta_s^2}{2m} E \left[ \langle \nabla f(y_k^*), \bar{x}^{s-1} - \pi^{s-1} \rangle \right] \\
= \frac{\eta}{2m^2} E \left[ \| \nabla f(y_k^*) - g_k^* \|^2 \right] + \frac{\theta_s^2}{2\eta} \| \pi^{s-1} - \bar{x}^{s-1} \|^2 \\
\leq \frac{1}{m} \left( f(\pi^{s-1}) - f(y_k^*) + \langle \nabla f(y_k^*), y_k^* - \pi^{s-1} \rangle \right) + \frac{\theta_s^2}{2\eta} \| \pi^{s-1} - \bar{x}^{s-1} \|^2,
\]
where the first equality holds due to the definition of our gradient operator, the second equality holds due to the fact \( E[\nabla f(y_k^*) - g_k^*] = 0 \), and the inequality follows from Lemma 4 with the setting \( \eta \leq 1/L \) (i.e., \( L\eta \leq 1 \)).

Combing the equality (15) and the inequality (16), we have
\[
A_2 + A_3 \\
\leq \left( \nabla f(y_k^*), \frac{\theta_s}{m} (x^* - p_k^*) + \frac{1}{m} (y_k^* - \bar{x}^{s-1}) \right) \\
+ \frac{\theta_s^2}{2\eta} \| z_k^* - p_k^* - 2(\pi^{s-1} - \bar{x}^{s-1}) \|^2 - \frac{\theta_s^2}{\eta} \| \pi^{s-1} - \bar{x}^{s-1} \|^2 \\
+ \frac{\theta_s^2}{2\eta} \left( \| x^* - p_k^* - \pi^{s-1} + \bar{x}^{s-1} \|^2 - \| x^* - p_k^* - 2(\pi^{s-1} - \bar{x}^{s-1}) \|^2 \right) \leq \frac{\theta_s}{m} f(x^*) + \left( 1 - \frac{\theta_s}{m} \right) \left( f(\pi^{s-1}) - f(y_k^*) \right) \\
+ \frac{\theta_s^2}{2\eta} \| z_k^* - p_k^* - 2(\pi^{s-1} - \bar{x}^{s-1}) \|^2 - \frac{\theta_s^2}{\eta} \| \pi^{s-1} - \bar{x}^{s-1} \|^2 \\
+ \frac{\theta_s^2}{2\eta} \left( \| x^* - p_k^* - \pi^{s-1} + \bar{x}^{s-1} \|^2 - \| x^* - p_k^* - 2(\pi^{s-1} - \bar{x}^{s-1}) \|^2 \right),
\]
where the first inequality follows from the update rule of \( y_k^* = \frac{\theta_s}{m} p_k^* + \left( 1 - \frac{\theta_s}{m} \right) \pi^{s-1} \) and the following fact that
\[
\left( \nabla f(y_k^*), \frac{\theta_s}{m} (x^* - p_k^*) + \frac{1}{m} (y_k^* - \pi^{s-1}) \right) \\
= \left( \nabla f(y_k^*), \frac{\theta_s}{m} x^* + \left( 1 - \frac{\theta_s}{m} \right) \pi^{s-1} + \frac{1}{m} y_k^* - y_k^* \right) + \frac{1}{m} [f(\pi^{s-1}) - f(y_k^*)] \leq f \left( \frac{\theta_s}{m} x^* + \left( 1 - \frac{\theta_s}{m} \right) \pi^{s-1} + \frac{1}{m} y_k^* \right) - f(y_k^*) + \frac{1}{m} [f(\pi^{s-1}) - f(y_k^*)] \leq \frac{\theta_s}{m} f(x^*) + \left( 1 - \frac{\theta_s}{m} \right) \left( f(\pi^{s-1}) - f(y_k^*) \right).
\]
Note that the first inequality in (18) holds due to the property of $f$ (i.e., $\langle \nabla f(x), y - x \rangle \leq f(y) - f(x)$), and the last inequality in (18) follows from the convexity of $f(\cdot)$.

Using the above analysis and combining the inequalities (13), (14) and (18), we have

\[
\mathbb{E}[F(x^*_k) - F(x^*)] \\
\leq \left( 1 - \frac{\theta_s}{m} \right) [F(\bar{x}^{s-1}) - F(x^*)] + \left( 1 - \frac{\theta_s}{m} \right) [F(x^*_1) - F(x^*)] + \frac{\theta_s^2}{2\eta} \left( \|x^* - z^*_k - 1\|_2^2 - \|x^* - z^*_k\|_2^2 \right) \\
+ \frac{\theta_s^2}{2\eta} \left( \|x^* - p^*_k - 2(\bar{x}^{s-1} - \bar{x}^{s-1})\|_2^2 - \|x^* - z^*_k\|_2^2 \right) \tag{19}
\]

This completes the proof.

**Proof of Lemma 3.5 (Upper Bound of One-epoch):**

*Proof.* Using the one-iteration upper bound in Lemma 3.4, we have

\[
\mathbb{E}[F(x^*_k) - F(x^*)] \\
\leq \left( 1 - \frac{\theta_s}{m} \right) [F(\bar{x}^{s-1}) - F(x^*)] + C^s + \frac{\theta_s^2}{2\eta} \left( \|x^* - z^*_k - 1\|_2^2 - \|x^* - z^*_k\|_2^2 \right).
\]

Summing the above inequality over $k = 1, \cdots, m$, and using $\bar{x}^s = \frac{1}{m} \sum_{k=1}^m x^*_k$ and $F(\bar{x}^s) \leq \frac{1}{m} \sum_{k=1}^m F(x^*_k)$, we have

\[
\mathbb{E}[F(\bar{x}^s) - F(x^*)] \leq \left( 1 - \frac{\theta_s}{m} \right) [F(\bar{x}^{s-1}) - F(x^*)] + C^s \\
+ \frac{\theta_s^2}{2m\eta} \left( \|x^* - z^*_0\|_2^2 - \|x^* - z^*_m\|_2^2 \right). \tag{20}
\]

Furthermore, using Lemma 3.3, we have

\[
F(\bar{x}^{s-1}) - F(x^*) \\
\leq (1 - \theta_s)(F(\bar{x}^{s-1}) - F(x^*)) + \frac{\theta_s^2}{2m\eta} \left( \|x^* - \bar{x}^{s-1}\|_2^2 - \|x^* - \bar{x}^{s-1}\|_2^2 \right) + \mathcal{R}^s. \tag{21}
\]

By the above analysis, the upper bound of one-epoch can be obtained as follows:

\[
\mathbb{E}[F(\bar{x}^s) - F(x^*)] \\
\leq (1 - \theta_s)[F(\bar{x}^{s-1}) - F(x^*)] + \frac{\theta_s^2}{2m\eta} \left( \|x^* - \bar{x}^{s-1}\|_2^2 - \|x^* - \bar{x}^s\|_2^2 \right). \tag{22}
\]

This completes the proof.
Proof of Theorem 3.6

In this subsection, we prove the convergence property of DA VIS (i.e., Algorithm 1). Theorem 3.6 shows that DA VIS improves the convergence rate of some accelerated methods (e.g., Katyusha) from $O(1/S^2)$ to $O(1/(nS^2))$ for the non-SC problem (1). That is, the result shows that DA VIS has both the optimal oracle complexity, $O(n + \sqrt{nL/\epsilon})$, and the optimal convergence rate, $O(1/(nS^2))$.

Proof of Theorem 3.6:

Proof. Using the update rule of $\theta_s$ (i.e., $\theta_s = \frac{2}{s+1}$) for Algorithm 1, we have $\frac{1}{\theta_{s-1}} \geq \frac{1-\theta_s}{\theta_s}$. Therefore, we telescope the inequality (22) in Lemma 3.5 for all $s = 1, 2, \ldots, S$, and we have

$$\frac{1}{\theta_S^2} \mathbb{E} [F(\bar{x}^S) - F(x^*)] \leq \frac{1}{\theta_1^2} \mathbb{E} [F(\bar{x}^0) - F(x^*)] + \frac{\theta_1^2}{2m\eta} \|x^* - \bar{x}^0\|^2.$$

Since $\theta_1 = 1$,

$$\frac{1}{\theta_S^2} \mathbb{E} [F(\bar{x}^S) - F(x^*)] \leq \frac{1}{2m\eta} \|x^* - \bar{x}^0\|^2.$$

(23)

Since $\theta_s = \frac{2}{s+1}$, we have

$$\mathbb{E} [F(\bar{x}^S) - F(x^*)] \leq O \left( \frac{\|x^* - \bar{x}^0\|^2}{mS^2\eta} \right).$$

(24)

In other words, by choosing $m = \Theta(n)$, the total oracle complexity of our algorithm is $O(n + \sqrt{nL/\epsilon})$.

This completes the proof. □

Appendix C: Theoretical Analysis for DA VIS-ADMM

In this section, we analyze the convergence properties of the proposed DA VIS-ADMM algorithm (i.e., Algorithm 2). Similar to Theorem 3.6, the proof of Theorem 4.3 for DA VIS-ADMM relies on the one-epoch inequality in Lemma 10 below. To prove Lemma 10, we first give the upper bound in Lemma 6 below by using our snapshot scheme in Algorithm 2. Furthermore, by using our stochastic momentum iteration rules in Algorithm 2, we can obtain the upper bounds in Lemmas 8 and 9 below. Thus, we can obtain the upper bound of one-epoch in Lemma 10 by using Lemmas 6, 8 and 9.

Upper bound of double snapshot update for DA VIS-ADMM

Before giving the proof of Lemma 6, we first present the following lemma (Zheng & Kwok, 2016).

Lemma 5. Let $\varphi_k = \beta(\lambda_k - \lambda^*)$, any $\varphi = \beta\lambda$, and $\lambda_k = \lambda_{k-1} + Ax_k - w_k$, then

$$\mathbb{E} [-(Ax_k - w_k)^T (\varphi_k - \varphi)] = \beta \frac{1}{2} \mathbb{E} [\|\lambda_{k-1} - \lambda^* - \lambda\|^2 - \|\lambda_k - \lambda^* - \lambda\|^2 - \|\lambda_k - \lambda_{k-1}\|^2].$$
Lemma 6 (Upper bound of double snapshot update). Suppose that Assumption 1 holds. Let \( \{\bar{x}, \bar{u}, \bar{z}\} \) be the sequence generated by our deterministic gradient descent step in Algorithm 2, then we have

\[
\mathbb{E}\left[P(\bar{x}^{s-1}, \bar{u}^{s-1}) - \frac{\theta_s}{m} \langle A\bar{z}^{s-1} - \bar{p}^{s-1}, \phi \rangle\right] \\
\leq (1 - \theta_s) P(\bar{x}^{s-1}, \bar{u}^{s-1}) + \mathcal{R}^s \\
+ \frac{\beta\theta_s}{2m} \mathbb{E}\left[\|x^{s-2} - \lambda^* - \lambda\|^2 - \|x^{s-1} - \lambda^* - \lambda\|^2\right] \\
+ \frac{\beta\theta_s}{2m} \mathbb{E}\left[\|A\bar{z}^{s-2} - \bar{w}^s\|^2 - \|A\bar{z}^{s-1} - \bar{w}^s\|^2\right] \\
+ \frac{\theta_s^2}{2mn} \left(\|x^* - \bar{x}^{s-1}\|^2_{Q_s} - \|x^* - \bar{x}^{s-1}\|^2_{\tilde{Q}_s}\right).
\]

where \( \mathcal{R}^s = \left(\frac{\theta_s^2}{2m} - \frac{\theta_s^2}{2mn}\right) \|\bar{x}^{s-1} - \bar{z}^{s-1}\|^2_{Q_s} \).

Proof of Lemma 6:

Proof. We first recall the following iteration scheme of our deterministic gradient descent step,

\[
\bar{x}^{s-1} = \arg\min_z \left\{ \langle \nabla f(\bar{x}^{s-1}), z \rangle + \frac{\theta_s}{2mn} \|z - \bar{x}^{s-1}\|^2_{Q_s} + \frac{\beta}{2m} \|Az - \bar{p}^{s-1} + \bar{\lambda}^{s-2}\|^2 \right\},
\]

and \( \bar{x}^{s-1} \) is required to satisfy the following optimal condition,

\[
\nabla f(\bar{x}^{s-1}) + \frac{\theta_s}{mn} Q_s (\bar{x}^{s-1} - \bar{x}^{s-1}) + \frac{\beta}{m} A^T (A\bar{z}^{s-1} - \bar{p}^{s-1} + \bar{\lambda}^{s-2}) = 0. \tag{25}
\]

With \( \bar{\lambda}^{s-1} = A\bar{z}^{s-1} - \bar{p}^{s-1} + \bar{\lambda}^{s-2} \), we have

\[
\nabla f(\bar{x}^{s-1}) + \frac{\theta_s}{mn} Q_s (\bar{x}^{s-1} - \bar{x}^{s-1}) + \frac{\beta}{m} A^T \bar{\lambda}^{s-1} = 0. \tag{26}
\]

Since \( f(\cdot) \) is \( L \)-smooth and using the update rule of \( \bar{x}^{s-1} = \theta_s \bar{x}^{s-1} + (1 - \theta_s) \bar{z}^{s-1} \), the following inequality holds

\[
f(\bar{x}^{s-1}) \\
\leq f(\bar{x}^{s-1}) + \langle \nabla f(\bar{x}^{s-1}), \bar{x}^{s-1} - x^* \rangle + \frac{L}{2} \|\bar{x}^{s-1} - x^*\|^2 \\
\leq f(\bar{x}^{s-1}) + \theta_s \langle \nabla f(\bar{x}^{s-1}), \bar{x}^{s-1} - x^* \rangle + \frac{\theta_s^2}{2m} \|\bar{x}^{s-1} - x^*\|^2_{Q_s}, \tag{27}
\]

where the first inequality holds due to the smoothness of \( f(\cdot) \), and the second inequality follows from our choice of \( \eta \leq \frac{1}{L} \) and the fact that \( Q_s > I \). Furthermore, using the optimal condition in (26), we have

\[
f(\bar{x}^{s-1}) \\
\leq f(\bar{x}^{s-1}) + \frac{\theta_s^2}{2m} \|\bar{x}^{s-1} - x^*\|^2_{Q_s} \\
+ \theta_s \langle \nabla f(\bar{x}^{s-1}), \bar{x}^{s-1} - x^* \rangle + \theta_s \langle \nabla f(\bar{x}^{s-1}), x^* - \bar{x}^{s-1} \rangle \\
\leq f(\bar{x}^{s-1}) + \frac{\theta_s^2}{2m} \|\bar{x}^{s-1} - x^*\|^2_{Q_s} \tag{28}
\]

\[
+ \left(\frac{\theta_s}{mn} Q_s (\bar{x}^{s-1} - \bar{x}^{s-1}) + \frac{\beta\theta_s}{m} A^T \bar{\lambda}^{s-1}, x^* - \bar{x}^{s-1}\right) + \theta_s \langle \nabla f(\bar{x}^{s-1}), x^* - \bar{x}^{s-1} \rangle \\
\leq \theta_s f(x^*) + (1 - \theta_s) f(\bar{x}^{s-1}) + \frac{\theta_s^2}{2mn} \|x^* - \bar{x}^{s-1}\|^2_{Q_s} - \|x^* - \bar{x}^{s-1}\|^2_{Q_s} + \frac{\beta\theta_s}{m} \langle A^T \bar{\lambda}^{s-1}, x^* - \bar{x}^{s-1}\rangle + \mathcal{R}^s.
\]
where the last inequality follows from the convexities of $f(\cdot)$ and Property 1, i.e.,

$$
\frac{\theta^2}{2mn} \left\langle Q_s(z_s - \bar{x}_s), x^* - z_s \right\rangle = \frac{\theta^2}{2mn} \left( \|x^* - \bar{x}_s - z_s\|_{Q_s}^2 - \|x^* - z_s\|_{Q_s}^2 - \|z_s - \bar{x}_s\|_{Q_s}^2 \right).
$$

Using the optimality condition of Problem (2), i.e., $\nabla f(x^*) + \beta A^T \lambda^* = 0$, and let $\varphi_s = \beta (x_s - \lambda^*)$, then the following result holds

$$
\left\langle \beta A^T \bar{x}_s, x^* - \bar{x}_s \right\rangle = \langle \nabla f(x^*), \bar{x}_s - x^* \rangle + \langle \beta A^T \lambda^*, \bar{x}_s - x^* \rangle \\
+ \langle \beta A^T \bar{x}_s, x^* - \bar{x}_s \rangle \\
= \langle \nabla f(x^*), \bar{x}_s - x^* \rangle + \langle A^T \varphi_s, x^* - \bar{x}_s \rangle,
$$

(29)

Substituting the above equality into (28) and $\bar{x}_s = \theta_s z_s + (1 - \theta_s) \bar{x}_s$. Then we have

$$
f(\bar{x}_s) - f(x^*) + \langle \nabla f(x^*), \bar{x}_s - x^* \rangle - \frac{\theta_s}{m} \langle A^T \varphi_s, x^* - \bar{x}_s \rangle \\
\leq (1 - \theta_s) \left( f(\bar{x}_s) - F(x^*) + \langle \nabla f(x^*), x^* - \bar{x}_s \rangle \right) \\
+ \frac{\theta^2}{2mn} \left( \|x^* - \bar{x}_s\|_{Q_s}^2 - \|x^* - \bar{x}_s\|_{Q_s}^2 \right) + R^s.
$$

(30)

With $\bar{x}_s = A \bar{x}_s - \vec{p}_s + \bar{\lambda}_{s-2}$ and the update rules in Eq. (6), and using Lemma 3 in (Zheng & Kwok, 2016), we obtain

$$
h(\bar{p}_s) - h(w^*) - \nabla h(w^*)^T (\bar{p}_s - w^*) - \langle \varphi^{s-1}, w^* - \bar{p}_s \rangle \\
\leq \frac{\beta}{2m} \left[ \|A\bar{p}_s - w_s\|^2 - \|A\bar{p}_s - w^*\|^2 + \|\bar{\lambda}_{s-2}\| \right].
$$

(31)

Furthermore, using the update rule of $\bar{p}_s = \theta_s p^{s-1} + (1 - \theta_s) \bar{p}_s$, the optimal condition (i.e., $\nabla h(w^*) + \beta \lambda^* = 0$) and the result in (31), we have

$$
h(\bar{w}_s) - h(w^*) + \langle \nabla h(w^*), w^* - \bar{w}_s \rangle - \frac{\theta_s}{m} \langle \varphi^{s-1}, w^* - \bar{p}_s \rangle \\
\leq (1 - \theta_s) \left( h(\bar{w}_s) - h(w^*) + \langle \nabla h(w^*), w^* - \bar{w}_s \rangle - \langle \varphi^{s-1}, w^* \rangle \right) \\
+ \frac{\beta}{2m} \left[ \|A\bar{p}_s - w_s\|^2 - \|A\bar{p}_s - w^*\|^2 + \|\bar{\lambda}_{s-2}\| \right].
$$

(32)

For any $\varphi = \beta \lambda$ and $Ax^* - w^* = 0$, we have

$$
\langle A^T \varphi_s, x^* - \bar{x}_s \rangle + \langle \varphi_s, w^* - \bar{p}_s \rangle + \langle A\bar{x}_s - \bar{p}_s, \varphi_s - \bar{\varphi} \rangle \\
= -\langle A\bar{x}_s - \bar{p}_s, \varphi \rangle.
$$

(33)

Using Lemma 5 and the update rule of $\bar{\lambda}_{s-1}$ in Algorithm 2, we have

$$
-\langle A\bar{x}_s - \bar{p}_s, \varphi \rangle \\
= \frac{\beta}{2} \left( \|\bar{\lambda}_{s-2} - \lambda^* - \bar{\lambda}_{s-2} - \lambda^*\|^2 \right),
$$

(34)

where the equality holds due to Property 1.
Using the results in (30), (32) and (34), the definition of $P(x, y)$ (i.e., $P(x, w) = f(x) - f(x^*) - \nabla f(x^*)^T(x - x^*) + h(w) - h(w^*) - \nabla h(w^*)^T(w - w^*)$) and the update rules in Algorithm 2, we have

$$
\mathbb{E}\left[P(\overbar{\sigma}^{s-1}, \overbar{\sigma}^{s-1}) - \frac{\theta_s}{m} \langle A\overbar{\sigma}^{s-1} - \overbar{\theta}^{s-1}, \varphi \rangle \right] \\
\leq (1 - \theta_s) P(\overbar{\sigma}^{s-1}, \overbar{\sigma}^{s-1}) + \mathcal{R}^s \\
+ \frac{\beta \theta_s}{2m} \mathbb{E}\left[\|\overbar{\sigma}^{s-2} - \lambda^* - \lambda\|^2 - \|\overbar{\sigma}^{s-1} - \lambda^* - \lambda\|^2 \right] \\
+ \frac{\beta \theta_s}{2m} \mathbb{E}\left[\|A\overbar{\sigma}^{s-2} - w^*\|^2 - \|A\overbar{\sigma}^{s-1} - w^*\|^2 \right] \\
+ \frac{\theta^2}{2m\eta} (\|x^* - \overbar{\sigma}^{s-1}\|_{Q_x}^2 - \|x^* - \overbar{\sigma}^{s-1}\|_{Q_x}^2). 
$$

(35)

This completes the proof. \qed

**Upper bound of one-iteration in our stochastic gradient descent step**

Before giving the proofs of Lemmas 8 and 9, we first present the following lemma.

**Lemma 7.** Let $g_k^s = \nabla f_{I_k}(y_k^s) - \nabla f_{I_k}(\overbar{\sigma}^{s-1}) + \nabla f(\overbar{\sigma}^{s-1})$, and $b$ be the size of mini-batch $I_k$. Then

$$
\mathbb{E}\left[\|\nabla f(y_k^s) - g_k^s\|^2 \right] \\
\leq \frac{2L(n-b)}{b(n-1)} \left[f(\overbar{\sigma}^{s-1}) - f(y_k^s) + \langle \nabla f(y_k^s), y_k^s - \overbar{\sigma}^{s-1}\rangle \right].
$$

**Lemma 8** (Upper bound of one-iteration of $f$). Suppose that Assumption 1 holds. Let $\{\overbar{\sigma}, \overbar{\omega}\}$ be sequence generated by Algorithm 2, then we have

$$
\mathbb{E}\left[f(\overbar{\sigma}) - f(x^*) + \langle \nabla f(x^*), x^* - \overbar{\sigma}\rangle - \frac{\theta_s}{m} \sum_{k=1}^{m} \langle A^T \varphi_k^*, x^* - x^{s-1} + v_k^s - z_k^s \rangle \right] \\
\leq \mathbb{E}\left[(1 - \frac{\theta_s}{m}) \left[f(\overbar{\sigma}^{s-1}) - f(x^*) + \langle \nabla f(x^*), x^* - \overbar{\sigma}^{s-1}\rangle \right] \right] \\
+ \mathbb{E}\left[\frac{\theta^2}{2m\eta} (\|x^* - z_0^s\|_{Q_x}^2 - \|x^* - z_m^s\|_{Q_x}^2) \right] + C^s,
$$

(36)

where $C^s = \left(\frac{\theta^2}{2m\eta} - \frac{\theta^2}{2\eta}\right) \|z_0^s - \overbar{\sigma}^{s-1}\|_{Q_x}^2$ and $\varphi_k^* = \beta(\lambda_k^* - \lambda^*)$.

**Proof of Lemma 8:**

We give the upper bound for our stochastic gradient descent step in this lemma. Using the similar proof in Lemma 3.4 in Appendix B, we have the following result for our stochastic gradient descent step.

Let $g_{I_k}(y_k^s) = \nabla f_{I_k}(y_k^s) - \nabla f_{I_k}(\overbar{\sigma}^{s-1}) + \nabla f(\overbar{\sigma}^{s-1})$, we have $\overbar{\nabla}_{I_k}(y_k^s) = g_{I_k}(y_k^s) + \frac{m\theta_s}{\eta} Q_x(\overbar{\sigma}^{s-1} - \overbar{\sigma}^{s-1})$. Since the
function $f(\cdot)$ is $L$-smooth, and by using the update rule of $x_k^*$ and the similar derivation as in Lemma 3.4, we have

$$f(x_k^*) \leq f(y_k^*) + \left\langle \nabla f(y_k^*), \frac{\theta_m}{m} (z_k^* - p_k^*) \right\rangle + \frac{L \theta^2}{2m^2} \|z_k^* - p_k^*\|^2$$

$$= f(y_k^*) + \frac{\theta_m}{m} \left\langle \nabla I_{t_k}(y_k^*), x^* - p_k^* \right\rangle + \frac{\theta_m}{m} \left\langle \frac{m \theta_s}{\eta} Q_s(z_k^* - p_k^*) + \beta A^T \lambda_k^*, x^* - z_k^* \right\rangle$$

$$+ \left\langle \nabla f(y_k^*) - \nabla I_{t_k}(y_k^*), \frac{\theta_s}{m} (z_k^* - p_k^*) \right\rangle + \frac{L \theta^2}{2m^2} \|z_k^* - p_k^*\|^2$$

$$= f(y_k^*) + \frac{\theta_m}{m} \left\langle \nabla I_{t_k}(y_k^*), x^* - p_k^* \right\rangle + \frac{\theta_m}{m} \left\langle \beta A^T \lambda_k^*, x^* - z_k^* \right\rangle$$

$$+ \left\langle \nabla f(y_k^*) - \nabla I_{t_k}(y_k^*), \frac{\theta_s}{m} (z_k^* - p_k^*) \right\rangle + \frac{L \theta^2}{2m^2} \|z_k^* - p_k^*\|^2$$

$$+ \frac{\theta^2}{2\eta} (\|x^* - p_k^* - 2(\bar{x}^{s-1} - \bar{x}^{s-1})\|^2 - \|x^* - z_k^*\|^2 - \|z_k^* - p_k^* - 2(\bar{x}^{s-1} - \bar{x}^{s-1})\|^2),$$

where the equality $\overset{\Delta}{=} = 0$. Using the optimality condition of Problem (2) and the update rule of $x_k^*$ in Algorithm 2, i.e.,

$$\nabla f(x_k^*) + \beta A^T \lambda_k^* = 0,$$

and the similar derivation as in Lemma 3.4 in this main paper and Lemma 6, the inequality (37) can be rewritten as follows:

$$\mathbb{E}[f(x_k^*)] \leq \mathbb{E} \left[ f(y_k^*) \right] + \frac{\theta_m}{m} \mathbb{E} \left[ \left\langle \nabla I_{t_k}(y_k^*), x^* - p_k^* \right\rangle \right]$$

$$+ \frac{\theta_m}{m} \mathbb{E} \left[ \left\langle \beta A^T \lambda_k^*, x^* - z_k^* \right\rangle \right]$$

$$+ \frac{L \theta^2}{2m^2} \mathbb{E} [\|z_k^* - p_k^*\|^2] + \frac{\theta^2}{2\eta} \mathbb{E} \left[ \|x^* - p_k^* - 2(\bar{x}^{s-1} - \bar{x}^{s-1})\|^2 - \|x^* - z_k^*\|^2 - \|z_k^* - p_k^* - 2(\bar{x}^{s-1} - \bar{x}^{s-1})\|^2 \right].$$

Using the inequality (39) and the similar derivation as in Lemma 3.4 in this main paper and Lemma 6, the following result holds

$$\mathbb{E}[f(x_k^*) - f(x^*)]$$

$$\leq \mathbb{E} \left[ \left( 1 - \frac{\theta_m}{m} \right) \left[ f(\bar{x}^{s-1}) - f(x^*) \right] \right] + \mathbb{E} \left[ \frac{\theta_m}{m} \left\langle \beta A^T \lambda_k^*, x^* - z_k^* \right\rangle \right]$$

$$+ \frac{\theta^2}{2\eta} \mathbb{E} \left[ \|x^* - z_{k-1}^*\|^2_{Q_s} - \|x^* - z_k^*\|^2_{Q_s} \right] + C^s.$$
Since \( f(x) - f(x^*) + \nabla f(x^*)^T (x^* - x) \geq 0 \), using the update rules in Algorithm 2 and summing up the inequality (41) for all the iterations \( k = 1, 2, \ldots, m \), and dividing both side of the resulting inequality by \( m \), and using the update rules of \( \bar{x}^s = \frac{1}{m} \sum_{k=1}^{m} x^s_k \), \( f(\bar{x}^s) \leq \frac{1}{m} \sum_{k=1}^{m} f(x^s_k) \), and \( x^0_0 = \bar{x}^{s-1} \), we have

\[
\mathbb{E} \left[ f(\bar{x}^s) - f(x^*) + (\nabla f(x^*) \cdot x^* - \bar{x}^s) \right] - \frac{\theta_s}{m^2} \sum_{k=1}^{m} \langle A^T \varphi_k^s, x^* - z_k^s \rangle \\
\leq \mathbb{E} \left[ (1 - \frac{\theta_s}{m}) \left[ f(\bar{x}^{s-1}) - f(x^*) + (\nabla f(x^*) \cdot x^* - \bar{x}^{s-1}) \right] \\
+ \mathbb{E} \left[ \frac{\theta^2_s}{2m^2} \left( \|x^* - z_0^s\|^2_{Q_*} - \|x^* - z_m^s\|^2_{Q_*} \right) \right] + C^s.
\]

This completes the proof. □

**Lemma 9** (Upper bound of one-iteration of \( h \)). Using the same notation as in Lemma 8, we have

\[
\mathbb{E} \left[ h(\bar{w}^s) - h(w^*) + \tilde{\nabla} h(w^*)^T (w^* - \bar{w}^s) \right] - \frac{\theta_s}{m^2} \sum_{k=1}^{m} \langle \varphi_k^s, w^* - w_k^s \rangle \\
\leq (1 - \frac{\theta_s}{m}) \left[ h(\bar{w}^{s-1}) - h(w^*) + \tilde{\nabla} h(w^*)^T (w^* - \bar{w}^{s-1}) \right] \\
+ \frac{\beta \theta_s}{2m^2} \mathbb{E} \left[ \|Az_0^s - w^*\|^2 - \|Az_m^s - w^*\|^2 + \sum_{k=1}^{m} \|\lambda_k - \lambda_{k-1}\|^2 \right].
\]

**Proof.** The optimal condition with respect to \( w_k^s \) is \( \zeta_k^s - \beta (Az_{k-1}^s - w_k^s + \lambda_k^s) = 0 \), where \( \zeta_k^s \) is a subgradient at \( w_k^s \). Using the similar derivation as in Lemma 3 in (Zheng & Kwok, 2016) and the update rules in Algorithm 2, we obtain

\[
\mathbb{E} \left[ h(w_k^s) - h(w^*) - \tilde{\nabla} h(w^*)^T (w_k^s - w^*) \right] - \frac{\theta_s}{m} \langle \varphi_k^s, w^* - w_k^s \rangle \\
\leq \frac{\beta}{2} \mathbb{E} \left[ \|Az_{k-1}^s - w^*\|^2 - \|Az_k^s - w^*\|^2 + \|\lambda_k^s - \lambda_{k-1}\|^2 \right].
\]

Since \( h(x) - h(w^*) + \tilde{\nabla} h(x^*)^T (w^* - w) \geq 0 \), using the update rules in Algorithm 2 and summing up the above inequality for all the iterations \( k = 1, 2, \ldots, m \), and using the update rules in Algorithm 2, we can obtain the result of (42).

This completes the proof. □

**Lemma 10** (Upper bound of one-epoch). Using the same notation as in Lemma 8, we have

\[
\mathbb{E} \left[ P(\bar{x}^*, \bar{w}^*) - \langle A\bar{x}^* - \bar{w}^*, \varphi \rangle \right] \\
\leq (1 - \theta_s) \left[ P(\bar{x}^{s-1}, \bar{w}^{s-1}) - \langle A\bar{x}^{s-1} - \bar{w}^{s-1}, \varphi \rangle \right] \\
+ \frac{\theta^2_s}{2m^2} \mathbb{E} \left[ \|x^* - z_0^s\|^2_{Q_*} - \|x^* - z_m^s\|^2_{Q_*} \right] \\
+ \frac{\beta \theta_s}{2m^2} \mathbb{E} \left[ \|\lambda_0^s - \lambda^* - \lambda\|^2 - \|\lambda_m^s - \lambda^* - \lambda\|^2 \right] \\
+ \frac{\beta \theta_s}{2m} \mathbb{E} \left[ \|A\bar{x}^{s-2} - w^*\|^2 - \|A\bar{x}^{s-1} - w^*\|^2 \right] \\
+ \frac{\beta \theta_s}{2m} \mathbb{E} \left[ \|\lambda^{s-2} - \lambda^* - \lambda\|^2 - \|\lambda^{s-1} - \lambda^* - \lambda\|^2 \right].
\]
Proof of Lemma 10:

Proof. Using the definition of $P(x, w)$ and combining the inequality (36) in Lemma 8 and the inequality (42) in Lemma 9, we have

$$
\mathbb{E}
\left[
\frac{1}{m} P(x', w') - \frac{\theta_s}{m^2} \sum_{k=1}^{m} (A^T \varphi_k^s, x^* - z_k^s) - \frac{\theta_s}{m^2} \sum_{k=1}^{m} (\varphi_k^s, w^* - w_k^s)
\right]
\leq
\left(1 - \frac{\theta_s}{m}\right) P(x^{s-1}, w^{s-1})
+ \frac{\theta_s^2}{2m\eta} \mathbb{E} \left[\|x^* - z_0^s\|_{Q_s}^2 - \|x^* - z_m^s\|_{Q_s}^2\right] + C^s
+ \frac{\beta\theta_s}{2m^2} \mathbb{E} \left[\|A z_0^s - w^s\|^2 - \|A z_m^s - w^s\|^2 + \sum_{k=1}^{m} \|\lambda_{k-1}^s - \lambda_k^s\|_2^2\right].
\tag{44}
$$

Using Lemma 5 with the update rule of $\lambda_k^s$ in Algorithm 2, $\varphi_k^s = \beta(\lambda_k^s - \lambda^s)$ and $\varphi = \beta \lambda$, we have

$$
- \frac{\theta_s}{m^2} \sum_{k=1}^{m} (A z_k^s - w_k^s, \varphi_k^s - \varphi)
= \frac{\beta\theta_s}{2m^2} \left(\|\lambda_0^s - \lambda^s - \lambda\|^2 - \|\lambda_m^s - \lambda^s - \lambda\|^2\right) - \frac{\beta\theta_s}{2m^2} \sum_{k=1}^{m} \|\lambda_{k-1}^s - \lambda_k^s\|_2^2,
\tag{45}
$$

where the equality holds due to Property 1.

Adding both sides of the inequalities (44) and (45) and the similar derivation as in Lemma 6 with the update rule of $x_k^s$ in Algorithm 2, we have

$$
\mathbb{E}[P(x^s, w^s) - \langle A x^s - w^s, \varphi \rangle]
\leq
\left(1 - \frac{\theta_s}{m}\right) [P(x^{s-1}, w^{s-1}) - \langle A x^{s-1} - w^{s-1}, \varphi \rangle]
+ \frac{\theta_s^2}{2m\eta} \mathbb{E} \left[\|x^* - z_0^s\|_{Q_s}^2 - \|x^* - z_m^s\|_{Q_s}^2\right] + C^s
+ \frac{\beta\theta_s}{2m^2} \mathbb{E} \left[\|A z_0^s - w^s\|^2 - \|A z_m^s - w^s\|^2 + \|\lambda_0^s - \lambda^s - \lambda\|^2 - \|\lambda_m^s - \lambda^s - \lambda\|^2\right].
\tag{46}
$$

Furthermore, by using Lemma 6 for the upper bound of our new snapshot point, we have

$$
\mathbb{E}[P(x^{s-1}, w^{s-1}) - \langle A x^{s-1} - w^{s-1}, \varphi \rangle]
\leq
\left(1 - \theta_s\right) [P(x^{s-1}, w^{s-1}) - \langle A x^{s-1} - w^{s-1}, \varphi \rangle]
+ \mathcal{R}^s
+ \frac{\beta\theta_s}{2m} \mathbb{E} \left[\|x^s - \bar{x}^{s-2} - \lambda^s - \lambda\|^2 - \|\bar{x}^{s-1} - \lambda^s - \lambda\|^2\right]
+ \frac{\beta\theta_s}{2m} \mathbb{E} \left[\|A z^{s-2} - w^s\|^2 - \|A z^{s-1} - w^s\|^2\right]
+ \frac{\theta_s^2}{2m\eta} \left(\|x^s - \bar{x}^{s-1}\|_{Q_s}^2 - \|x^s - \bar{x}^{s-1}\|_{Q_s}^2\right).
$$

By adding up the above two inequalities, the result of Lemma 10 holds.

This completes the proof.
Proof of Theorem 4.3 (i.e., Convergence Analysis of DAVIS-ADMM)

Proof. Using the upper bound of the $s$-th epoch in Lemma 10 and dividing both sides of the inequality (43) by $\theta_s$ instead of $\theta^2_s$ in Theorem 4.3. According to the update rule of $\theta_s$, and summing up the inequality (43) for all the stages ($s = 1, 2, \cdots, S$) with $w^* = Ax^*$ and $x^{-1} = x^0$, we have

$$
\frac{1}{\theta_S} \mathbb{E} \left[ P(\bar{x}^S, \bar{w}^S) - \frac{1}{S} \sum_{s=1}^{S} \langle \varphi, A\bar{x}^s - \bar{w}^s \rangle \right] 
\leq \left[ \frac{(1 - \theta_1)}{\theta_1} \right] \left[ P(\bar{x}^0, \bar{w}^0) - \langle \varphi, A\bar{x}^0 - \bar{w}^0 \rangle \right] 
+ \frac{\theta_1}{2m^{\eta}} \|x^* - \bar{x}^0\|^2_{Q_1}
+ \frac{\theta_1}{2m^{\eta}} \|\bar{x}^0 - w^*\|^2 
+ \frac{\beta}{2m} \mathbb{E} \left[ \|A\bar{x}^0 - w^*\|^2 + \|\lambda^0 - \lambda^*\|^2 \right] 
+ \frac{\beta}{2m} \mathbb{E} \left[ \|\lambda^* - \lambda\|^2 + \|A\bar{x}^0 - w^*\|^2 \right].
$$

(47)

With $\theta_1 \leq 2/(s + 1)$ and $\theta_1 = 1$, using the update rules of Algorithm 2, and multiplying both sides of the above inequality by $2/(S + 1)$, we have

$$
\mathbb{E} \left[ P(\bar{x}^S, \bar{w}^S) - \langle \varphi, \frac{1}{S} \sum_{s=1}^{S} (A\bar{x}^s - \bar{w}^s) \rangle \right]
\leq \frac{2}{m(S + 1)} \left[ P(\bar{x}^0, \bar{w}^0) - \langle \varphi, A\bar{x}^0 - \bar{w}^0 \rangle \right]
+ \frac{1}{m^{\eta}(S + 1)} \|x^* - \bar{x}^0\|^2_{Q_1}
+ \frac{\beta}{m(S + 1)} \left[ 2\|ATAx\|_2^2 \|x^* - \bar{x}^0\|^2 + \|\lambda^0 - \lambda^*\|^2 + \|\lambda^* - \lambda\|^2 \right].
$$

(48)

Let $\bar{x} = \frac{1}{S} \sum_{s=1}^{S} \bar{x}^s$ and $\bar{w} = \frac{1}{S} \sum_{s=1}^{S} \bar{w}^s$. Setting $\varphi = \delta \frac{A\bar{x} - \bar{w}}{\|A\bar{x} - \bar{w}\|}$, then the following inequalities hold:

$$
- \langle A\bar{x}^0 - \bar{w}^0, \varphi \rangle \leq \|\varphi\| \|A\bar{x}^0 - \bar{w}^0\| \leq \delta \|A\bar{x}^0 - \bar{w}^0\|.
$$

Therefore, we have

$$
\mathbb{E} \left[ P(\bar{x}^S, \bar{w}^S) + \delta \|A\bar{x}^S - \bar{w}^S\| \right]
\leq \frac{2}{m(S + 1)} \left[ P(\bar{x}^0, \bar{w}^0) + \delta \|A\bar{x}^0 - \bar{w}^0\| \right]
+ \frac{1}{m^{\eta}(S + 1)} \|x^* - \bar{x}^0\|^2_{Q_1}
+ \frac{\beta}{m(S + 1)} \left[ 2\|ATAx\|_2^2 \|x^* - \bar{x}^0\|^2 + \|\lambda^0 - \lambda^*\|^2 + \|\lambda^* - \lambda\|^2 \right]
\leq \frac{2}{m(S + 1)} \left[ P(\bar{x}^0, \bar{w}^0) + \delta \|A\bar{x}^0 - \bar{w}^0\| \right]
+ \frac{1}{m^{\eta}(S + 1)} \|x^* - \bar{x}^0\|^2_{Q_1}
+ \frac{\beta}{m(S + 1)} \left[ 2\|ATAx\|_2^2 \|x^* - \bar{x}^0\|^2 + 2\|\lambda^0\|^2 + 2\|\lambda^*\|^2 + 4\|\lambda\|^2 \right].
$$

(49)
Closing the Convergence Gaps in Non-Strongly Convex Optimization

By choosing \( m = \Theta(n) \), we have

\[
\mathbb{E} \left[ P(\tilde{x}^S, \tilde{w}^S) + \delta \| A\tilde{x}^S - \tilde{w}^S \| \right] \\
\leq \mathcal{O} \left( \frac{2[P(\tilde{x}^0, \tilde{w}^0) + \delta \| A\tilde{x}^0 - \tilde{w}^0 \|]}{n(S + 1)} + \frac{\| x^* - \tilde{x}^0 \|_Q^2}{n(S + 1)} + \frac{c_1 \beta}{n(S + 1)} \right),
\]

(50)

where \( c_1 \) is a constant, i.e., \( c_1 = 2\| A^T A \|^2_2 \| x^* - \tilde{x}^0 \|^2 + 2\| \lambda^0 - \lambda^* \|^2 + 8\delta^2 + 10\| \lambda^* \|^2 \).

This completes the proof. \( \Box \)

Appendix D: Experimental Setup

In this section, we also present detailed experimental setups for solving non-SC problems. All the experimental results were conducted on a PC with an Intel Core i7-7700 3.6GHz and 32GB RAM. All the experimental results show that the proposed DAVIS algorithm consistently converges much faster than other accelerated stochastic algorithms including Katyusha (Allen-Zhu, 2018) and Varag (Lan et al., 2019).

We are optimizing the following binary non-strongly convex problem with \( a_i \in \mathbb{R}^d, b_i \in \{-1, +1\}, i = 1, 2, \ldots n \):

\[
\ell_1\text{-norm Regularized Logistic Regression: } \frac{1}{n} \sum_{i=1}^{n} \log (1 + \exp ( -b_i a_i^T x )) + \lambda \| x \|_1,
\]

(51)

where \( \lambda \) is the regularization parameter.

For the \( \ell_1\)-norm regularized logistic regression problem, we set the length of each epoch to \( m = 2n \), as suggested by (Johnson & Zhang, 2013). For all the algorithms including the stochastic ADMM algorithms, we carefully tune the step-size \( \eta \) from the set \( \{ a \times 10^{-k} : a \in \{1, 2, \ldots, 9\}, k \in \mathbb{Z} \} \) for each plot. For Katyusha, the parameter \( \tau \) is set to 0.5, as suggested by (Allen-Zhu, 2018). In particular, for Varag, the length of each epoch and other parameters are set according to its original paper (Lan et al., 2019).

In the experiment of graph-guided fused Lasso, we set the mini-batch size to \( b = 50 \) for all the stochastic ADMM algorithms such as SVRG-ADMM, ASVRG-ADMM and our DAVIS-ADMM algorithm. Moreover, the length of each epoch in SVRG-ADMM, ASVRG-ADMM and DAVIS-ADMM is set to \( m = \lfloor 2n/b \rfloor \). For ASVRG-ADMM, the parameters \( \eta \) and \( \beta \) are set according to its original papers (Liu et al., 2017; 2021).

References


