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 $\mathcal{O}(1/(nS^2))$ and optimal gradient complexity $\mathcal{O}(n+\sqrt{nL/\epsilon})$, 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 $\mathcal{O}(1/S^2)$ to $\mathcal{O}(1/(nS^2))$. 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 $\mathcal{O}(1/(nS))$ and optimal oracle complexity $\mathcal{O}\big(n + L/\epsilon\big)$ 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

$$\min_{x\in\mathbb{R}^d} \Big\{F(x) \triangleq \frac{1}{n}\sum\nolimits_{i=1}^{n} f_i(x) + h(x)\Big\}, \qquad (1)$$

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 ϵ-suboptimal solution) of the stochastic variance reduction methods mentioned above is $\mathcal{O}((n+\kappa)\log(1/\epsilon))$, while the complexity of accelerated deterministic methods including AGD (Nesterov, 1983) and APG (Beck & Teboulle, 2009) is $\mathcal{O}(n\sqrt{\kappa}\log(1/\epsilon))$, where κ is the condition number. Obviously, the complexities show that the variance reduction methods always converge

[1] Key Lab. of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, China [2] School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China [3] Peng Cheng Laboratory [4] Key Lab. of Machine Perception (MoE), School of Artificial Intelligence, Peking University [5] Institute for Artificial Intelligence, Peking University. Correspondence to: Fanhua Shang, Hongying Liu and Yuanyuan Liu <fhshang@foxmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

Table 1. Comparison of oracle complexities (i.e., the number of first-order oracle calls and proximal oracle calls (Lan, 2020; Xie et al., 2020)) and convergence rates of some stochastic methods for non-SC problems, where $S_0 := \lfloor \log_2(n) \rfloor + 1$. Note that we regard using reductions or proximal point variants as "Indirect" acceleration, such as Catalyst and Katyusha with reduction techniques.

Algorithms	SAGA (Defazio et al., 2014) SVRG (Johnson & Zhang, 2013)	Catalyst (Lin et al., 2015a)	Katyusha$^{\text{ns}}$ (Allen-Zhu, 2018)	Katyusha (Allen-Zhu, 2018)
Convergence rates	$\mathcal{O}\Big(\frac{1}{S}\Big)$	$\mathcal{O}\Big(\frac{\log^4(ns)}{nS^2}\Big)$	$\mathcal{O}\Big(\frac{1}{S^2}\Big)$	NA
Oracle complexities	$\mathcal{O}\Big(\frac{n}{\epsilon} + \frac{L}{\epsilon}\Big)$	$\mathcal{O}\Big((n+\sqrt{\frac{nL}{\epsilon}}) \log^2(\frac{1}{\epsilon})\Big)$	$\mathcal{O}\Big(\frac{n}{\sqrt{\epsilon}} + \sqrt{\frac{nL}{\epsilon}}\Big)$	$\mathcal{O}\Big(n \log(\frac{1}{\epsilon}) + \sqrt{\frac{nL}{\epsilon}}\Big)$
Direct	Yes	No	Yes	No
Algorithms	Varag (Lan et al., 2019)	VRADA (Song et al., 2020)	**DAVIS** **This paper**	Lower Bound (Woodworth & Srebro, 2016)
Convergence rates	$\mathcal{O}\Big(\frac{1}{n(S-S_0+4)^2}\Big)$	NA	$\mathcal{O}\Big(\frac{1}{nS^2}\Big)$	$\mathcal{O}\Big(\frac{1}{nS^2}\Big)$
Oracle complexities	$\mathcal{O}\Big(n \log_2(n) + \sqrt{\frac{nL}{\epsilon}}\Big)$	$\mathcal{O}\Big(n \log_2 \log_2(n) + \sqrt{\frac{nL}{\epsilon}}\Big)$	$\mathcal{O}\Big(n + \sqrt{\frac{nL}{\epsilon}}\Big)$	$\mathcal{O}\Big(n + \sqrt{\frac{nL}{\epsilon}}\Big)$
Direct	Yes	Yes	Yes	–

faster than accelerated batch methods as long as $\kappa \le \mathcal{O}(n^2)$. 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 $\mathcal{O}((n+\sqrt{n\kappa})\log(1/\epsilon))$ 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$^{\text{ns}}$) 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 $\log_2$ 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,

$$\min_{x\in\mathbb{R}^d, w\in\mathbb{R}^{d_1}} \big\{f(x)+h(w),\ \text{s.t.},\ Ax=w\big\}, \tag{2}$$

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 $\mathcal{O}(1/S)$ 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).

Algorithms	SAGA-ADMM (Zhong & Kwok, 2014)	SVRG-ADMM (Zheng & Kwok, 2016)	ASVRG-ADMM (Liu et al., 2021)	**DAVIS-ADMM** **This paper**	Lower Bound (Xie et al., 2020)
Convergence rates	$\mathcal{O}(\frac{1}{S})$	$\mathcal{O}(\frac{1}{S})$	$\mathcal{O}(\frac{1}{S^2})$	$\mathcal{O}(\frac{1}{nS})$	$\mathcal{O}(\frac{1}{nS})$
Oracle complexities	$\mathcal{O}\left(\frac{n}{\epsilon}+\frac{L}{\epsilon}\right)$	$\mathcal{O}\left(\frac{n}{\epsilon}+\frac{L}{\epsilon}\right)$	$\mathcal{O}\left(\frac{n}{\sqrt{\epsilon}}+\sqrt{\frac{nL}{\epsilon}}\right)$	$\mathcal{O}\left(n+\frac{L}{\epsilon}\right)$	$\mathcal{O}\left(n+\frac{L}{\epsilon}\right)$
Boundedness assumption	No	No	Yes	No	–

ing a four-composite minimization problem. However, there also exist some similar gaps between the convergence rates, as well as the oracle complexities, of the existing methods and the optimal convergence rate.

Motivations: (**I**) For solving the non-SC problem (1) (e.g., ℓ_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 $\mathcal{O}(n\log_2(n)+\sqrt{nL/\epsilon})$, 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[1].

(**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 $\mathcal{O}\big(n+L/\epsilon\big)$, 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., $\widetilde{\nabla} f_{i_k}(x_k)$) in (Johnson & Zhang,

[1] A work by Li (Li, 2021) appeared on arXiv, and can match the lower bound in (Woodworth & Srebro, 2016) for non-SC problems for a very wide range of ϵ.

2013; Zhang et al., 2013) uses the gradient at the snapshot $\widetilde{x}$ to progressively reduce the variance of the SGD estimator, $\nabla f_{i_k}(x_k)$, where i_k is chosen uniformly at random from $\{1, 2, \ldots, n\}$. Therefore, the SVRG estimator has been widely used in most stochastic variance reduced methods including Katyusha (Allen-Zhu, 2018) and Varag (Lan et al., 2019). More recently, there is a surge of interest in accelerating stochastic variance reduced methods such as Katyusha and Varag for solving unconstrained optimization problems. By using negative momentum and both proximal descent and mirror descent steps, Katyusha (Allen-Zhu, 2018) can obtain improved convergence rates. In contrast, Varag (Lan et al., 2019) only requires the solution of one sub-problem instead of the two for Katyusha. More recently, VRADA (Song et al., 2020) can match the lower bound for non-SC problems up to a $\log_2\log_2(n)$ factor.

2.2. Stochastic Optimization for Structured Regularization Problems

To solve the equality-constrained problem (2), the alternating direction method of multipliers (ADMM) is an efficient optimization method. However, ADMM and its deterministic variants suffer from a high per-iteration computational cost for large-scale problems. Thus, several stochastic ADMMs such as (Wang & Banerjee, 2012; Ouyang et al., 2013) have recently been proposed. The formulation (2) with the constraint $Ax + By = c$ is the general form of the ADMM (Boyd et al., 2011). With the constraint $Ax = w$, Problem (2) is a simpler optimization problem, e.g., generalized Lasso (Tibshirani & Taylor, 2011), which is called the special ADMM form. Together with the dual variable λ, the update steps of stochastic ADMMs are

$$w_k = \arg\min_{w}\left\{h(w) + \phi_k(x_{k-1}, w)\right\},$$

$$x_k = \arg\min_{x}\left\{x^T\widetilde{\nabla} f_{i_k}(x_{k-1}) + \frac{\|x - x_{k-1}\|_Q^2}{2\eta_k} + \phi_k(x, w_k)\right\},$$

$$\lambda_k = \lambda_{k-1} + Ax_k - w_k,$$

where $\phi_k(x, w) = \frac{\beta}{2}\|Ax - w + \lambda_{k-1}\|^2$, $\beta > 0$ is a penalty parameter, $\eta_k > 0$ is a parameter, and $\|x\|_Q^2 = x^TQx$ 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).

3. A Direct Optimal Stochastic Variance Reduction Algorithm

In this section, we propose a directly accelerated stochastic variance reduction gradient (DAVIS) algorithm for solving the non-SC problem (1). We first present a novel double snapshot acceleration framework for stochastic optimization, in which we need to compute the full gradient at the first snapshot and define a new update rule of the second snapshot in each outer loop. Moreover, we design a new stochastic variance reduction gradient estimator for each inner loop of our accelerated algorithm. Finally, we analyze the convergence properties of DAVIS, which show that it attains the optimal convergence rate $\mathcal{O}(1/(nS^2))$ and the optimal oracle complexity $\mathcal{O}(n+\sqrt{nL/\epsilon})$.

3.1. Main Ideas of DAVIS

As discussed above, there still exist some gaps between the oracle complexities of the directly accelerated algorithms (e.g., Katyusha and Varag) and the lower bound in (Woodworth & Srebro, 2016). Let x^* be an optimal solution of Problem (1) and $\widetilde{x}^0$ be a given starting vector, the convergence result of the directly accelerated version (i.e., Katyusha$^{\text{ns}}$) of Katyusha (Allen-Zhu, 2018) is $\mathcal{O}(\frac{F(\widetilde{x}^0)-F(x^*)}{S^2} + \frac{L\|x^*-\widetilde{x}^0\|^2}{nS^2})$. In order to achieve the optimal convergence rate $\mathcal{O}(1/(nS^2))$, we need to accelerate the rate of the first term from $\mathcal{O}(1/S^2)$ to $\mathcal{O}(1/(nS^2))$, and thus Algorithm 1 mainly includes a new extra snapshot and its update rule in each outer loop, one new stochastic gradient estimator and new momentum acceleration rules in each inner loop. Therefore, our accelerated algorithm attains the optimal convergence rate $\mathcal{O}(1/(nS^2))$ without restarting and without using any reduction techniques.

As most stochastic variance reduction methods including Katyusha (Allen-Zhu, 2018) and Varag (Lan et al., 2019), each epoch of our algorithms including Algorithm 1 consists of m inner-iterations, e.g., $m = 2n$ as suggested in (Johnson & Zhang, 2013). In each outer loop of the proposed algorithm, we need to compute the full gradient at the first snapshot point, and design a new update rule for the second snapshot point. In each inner loop of our algorithm, we define a new compensated stochastic variance reduction gradient estimator, and then present a new momentum acceleration scheme. More details are given below.

3.2. New Scheme of Double Snapshots in Outer Loop

In the s-th outer loop of Algorithm 1, we design two snapshot points $\widetilde{x}^{s-1}$ and $\overline{x}^{s-1}$, both of which remain unchanged in all the inner loops inside the same outer loop. The first snapshot point $\widetilde{x}^{s-1}$ takes the same role as in most variance reduction methods such as SVRG and Katyusha. That is, we need to compute the full gradient of $f(\cdot)$ at $\widetilde{x}^{s-1}$ in each

Algorithm 1 DAVIS for Problem (1)

```
Input: The number $S$ of epochs, the number $m$ of iterations per epoch.
Initialize: $\widetilde{x}^0$, $z_0^1=0$, $\theta_1=1$, and $\eta$.
for $s=1,2,\ldots,S$ do
    $\overline{z}^{s-1}=\mathrm{prox}_h^{\frac{\eta}{\theta_s}}(\widetilde{x}^{s-1}-\frac{m\eta}{\theta_s}\nabla f(\widetilde{x}^{s-1}))$;
    $\overline{x}^{s-1}=\theta_s\overline{z}^{s-1}+(1-\theta_s)\widetilde{x}^{s-1}$;// The second snapshot
    Compute the full gradient at the second snapshot point (i.e., $\overline{x}^{s-1}$), $\nabla f(\overline{x}^{s-1})=\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\overline{x}^{s-1})$;
    for $k=1,2,\ldots,m$ do
        Update $y_k^s$ via (4);
        Pick $i_k$ uniformly at random from $\{1,2,\ldots,n\}$;
        $\widetilde{\nabla}_{i_k}(y_k^s)=\nabla f_{i_k}(y_k^s)-\nabla f_{i_k}(\overline{x}^{s-1})+\nabla f(\overline{x}^{s-1})+\frac{m\theta_s}{\eta}\left(\overline{z}^{s-1}-\widetilde{x}^{s-1}\right)$;
        $z_k^s=\text{Prox-SGrad}(y_k^s)$;
        $x_k^s=\frac{\theta_s}{m}(z_k^s-p_k^s)+y_k^s$;
    end for
    $\widetilde{x}^s=\frac{1}{m}\sum_{k=1}^{m}x_k^s$, $\theta_{s+1}=\frac{2}{s+2}$, $z_0^{s+1}=z_m^s$;
end for
Output: $\widetilde{x}^S$.
```

outer loop, which is used to gradually reduce the variance of the SVRG estimator. Moreover, we design a novel update rule for the second snapshot point $\overline{x}^{s-1}$ as follows:

$$\overline{x}^{s-1}=\theta_s\overline{z}^{s-1}+(1-\theta_s)\widetilde{x}^{s-1}, \tag{3}$$

where θ_s is a parameter (e.g., $\theta_s=\frac{2}{s+1}$), and the auxiliary variable z^{s-1} is obtained by solving the following problem:

$$\overline{z}^{s-1}=\operatorname*{argmin}_z\left\{h(z)+\left\langle\nabla f(\widetilde{x}^{s-1}),z\right\rangle+\frac{\theta_s}{2m\eta}\|z-\widetilde{x}^{s-1}\|^2\right\}.$$

Here, η is a learning rate. Clearly, $\overline{z}^{s-1}$ is obtained by performing one deterministic gradient descent step from the snapshot point $\widetilde{x}^{s-1}$, which requires no gradient calculations. For the second snapshot point $\overline{x}^{s-1}$, we give its upper bound in Lemma 3.3 in Section 3.4.

3.3. New Stochastic Update Schemes in Inner Loop

In this subsection, we first define a new compensated stochastic variance reduction gradient estimator for the proposed algorithm, and then we design a new momentum acceleration update rule.

3.3.1. Compensated Stochastic Gradient Estimator

Before giving our new stochastic momentum acceleration scheme for our algorithm, we first define a new compensated gradient estimator.

Definition 3.1 (Compensated stochastic gradient estimator)**.** We define a new compensated stochastic variance reduction gradient estimator for our DAVIS algorithm as follows:

$$\widetilde{\nabla}_{i_k}(x)=\underbrace{\nabla f_{i_k}(x)-\nabla f_{i_k}(\overline{x}^{s-1})+\nabla f(\overline{x}^{s-1})}_{\text{SVRG estimator}}+\underbrace{m\theta_s(\overline{z}^{s-1}-\widetilde{x}^{s-1})/\eta}_{\text{Compensated estimator}}.$$

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 $\widetilde{\nabla}_{i_k}(x)$, 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:

$$\begin{aligned}&\text{Prox-SGrad}(x)\\ &\triangleq\operatorname*{argmin}_z\left\{h(z)+\langle\widetilde{\nabla}_{i_k}(x),z\rangle+\frac{m\theta_s}{2\eta}\|z-\delta_k^s\|^2\right\}\\ &=\mathrm{prox}_h^{\frac{\eta}{m\theta_s}}\left(\delta_k^s-\frac{\eta}{m\theta_s}\widetilde{\nabla}_{i_k}(x)\right),\end{aligned}$$

where $\delta_k^s=p_k^s+2(\overline{z}^{s-1}-\widetilde{x}^{s-1})$, and p_k^s is defined below, and $\mathrm{prox}_h^{\frac{\eta}{m\theta_s}}(\cdot)$ 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:

$$y_k^s=\frac{\theta_s}{m}(z_k^s-p_k^s)+\big(1-\frac{\theta_s}{m}\big)\overline{x}^{s-1}. \tag{4}$$

Here $p_k^s=z_{k-1}^s$ is designed to get a well-structured recursive form (i.e., $\|x^*-z_{k-1}^s\|^2-\|x^*-z_k^s\|^2$) 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:

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

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 $\mathcal{O}(1/S^2)$ to $\mathcal{O}(1/(nS^2))$ for the non-SC problem (1). We also suppose that the distance between an initial point $\widetilde{x}^0$ and an optimal solution x^* can be bounded by a constant c, i.e., $\|x^* - \widetilde{x}^0\| \leq c$, 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 $\{\overline{x}^s\}$ be the sequence generated by our double snapshot scheme in Algorithm 1, for a given v_k^s, we have*

$$F(\overline{x}^{s-1}) - F(x^*) \leq (1-\theta_s)(F(\widetilde{x}^{s-1}) - F(x^*)) + \mathcal{R}^s + \frac{\theta_s^2}{2m\eta}\left(\left\|x^* - \widetilde{x}^{s-1}\right\|^2 - \left\|x^* - \overline{z}^{s-1}\right\|^2\right),$$

where $\mathcal{R}^s = \left(\frac{\theta_s^2}{2\eta} - \frac{\theta_s^2}{2m\eta}\right)\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2$.

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 $\{x_k^s, z_k^s\}$ be the sequence generated by Algorithm 1. then*

$$\mathbb{E}[F(x_k^s) - F(x^*)] \leq \frac{\theta_s^2}{2\eta}\left(\left\|x^* - z_{k-1}^s\right\|^2 - \|x^* - z_k^s\|^2\right) + \mathcal{C}^s + \mathbb{E}\left[\left(1 - \frac{\theta_s}{m}\right)\left(F(\overline{x}^{s-1}) - F(x^*)\right)\right],$$

where $\mathcal{C}^s = \left(\frac{\theta_s^2}{2m\eta} - \frac{\theta_s^2}{2\eta}\right)\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2$.

Remark 1. In Lemma 3.4, the term $\mathcal{C}_k^s$ is produced by our gradient estimator, and is used to compensate the additional term $\mathcal{R}^s$ in Lemma 3.3 (i.e., $\mathcal{R}^s + \mathcal{C}^s = 0$). As we expected, the designed stochastic descent step can be used to counteract 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 $\{x_k^s\}$ be the sequence generated by Algorithm 1. Then we have*

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

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 $\widetilde{x}^s = \frac{1}{m}\sum_{k=1}^m x_k^s$ (i.e., the average point of the previous epoch), then the following result holds*

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

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

Remark 2. Theorem 3.6 shows that DAVIS achieves the optimal convergence rate $\mathcal{O}(1/(nS^2))$, while most existing directly accelerated methods including (Allen-Zhu, 2018; Zhou et al., 2018) attain the rate $\mathcal{O}(1/S^2)$. In particular, Algorithm 1 also attains the optimal oracle complexity $\mathcal{O}(n + \sqrt{nL/\epsilon})$, 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 $\mathcal{O}(n/\sqrt{\epsilon} + \sqrt{nL/\epsilon})$, $\mathcal{O}(n\log_2(n) + \sqrt{nL/\epsilon})$ and $\mathcal{O}(n\log_2\log_2(n) + \sqrt{nL/\epsilon})$, 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., $\widetilde{x}^s$ and $\overline{x}^s$) are introduced into our update rules in (4) and (5), while only one snapshot $\widetilde{x}^s$ 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

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 θ_s can be defined as: $\theta_s = \frac{2}{s+1}$ satisfying $\frac{1}{\theta_{s-1}^2} \geq \frac{1-\theta_s}{\theta_s^2}$. Different from Katyusha$^{\text{ns}}$, the coefficient of our momentum is θ_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. 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 $Q_s = \gamma I - \frac{\eta\beta}{\theta_s} A^T A$ with $\gamma \geq \eta\beta\|A^T A\|_2 + 1$ to ensure that $Q_s \succeq I$, 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 $\widetilde{x}^s$, $\widetilde{w}^s$ and $\widetilde{\lambda}^s$, while three additional snapshots $\overline{x}^s$, $\overline{w}^s$ and $\overline{\lambda}^s$ 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: $\widetilde{x}^s = \frac{1}{m}\sum_{k=1}^{m} x_k^s$, $\widetilde{w}^s = \frac{\theta_s}{m^2}\sum_{k=1}^{m} w_k^s + \left(1 - \frac{\theta_s}{m}\right)\overline{w}^{s-1}$, and $\widetilde{\lambda}^s = \frac{1}{m}\sum_{k=1}^{m}\lambda_k^s$. And the update rules of the three new snapshots in our DAVIS-ADMM are given as follows:

$$\begin{aligned}
\overline{x}^{s-1} &= \theta_s \overline{z}^{s-1} + (1-\theta_s)\widetilde{x}^{s-1}, \\
\overline{w}^{s-1} &= \theta_s \overline{p}^{s-1} + (1-\theta_s)\widetilde{w}^{s-1}, \\
\overline{\lambda}^{s-1} &= A\overline{z}^{s-1} - \overline{p}^{s-1} + \overline{\lambda}^{s-2}.
\end{aligned} \tag{6}$$

Let $\phi^s(z,p) = \frac{1}{2m}\|Az - p + \overline{\lambda}^{s-2}\|^2$, the auxiliary variables are defined as follows: $\overline{p}^{s-1} = \arg\min_p \left\{ h(p) + \frac{\beta}{2}\phi^s(\overline{z}^{s-2}, p) \right\}$, and $\overline{z}^{s-1} = \arg\min_z \left\{ \langle \nabla f(\widetilde{x}^{s-1}), z \rangle + \frac{\theta_s}{2m\eta}\|z - \widetilde{x}^{s-1}\|_{Q_s}^2 + \phi^s(z, \overline{p}^{s-1}) \right\}$.

Algorithm 2 DAVIS-ADMM for Problem (2)

```
Input: $S$ and $m$.
Initialize: $\widetilde{x}^0$, $\widetilde{w}^0$, $\overline{\lambda}^0$, $\theta_1 = 1$, and $\eta$.
for $s = 1, 2, \ldots, S$ do
    Update the snapshots $\overline{x}^{s-1}$, $\overline{w}^{s-1}$ and $\overline{\lambda}^{s-1}$ via (6);
    Compute the full gradient at the snapshot $\overline{x}^{s-1}$, $\nabla f(\overline{x}^{s-1}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\overline{x}^{s-1})$;
    for $k = 1, 2, \ldots, m$ do
        $w_k^s = \arg\min_w \left\{ h(w) + \frac{\beta}{2}\|Az_{k-1}^s - w + \lambda_{k-1}^s\|^2 \right\}$;
        $y_k^s = \frac{\theta_s}{m} w_k^s + \left(1 - \frac{\theta_s}{m}\right)\overline{w}^{s-1}$;
        Pick $i_k$ uniformly at random from $\{1, 2, \ldots, n\}$;
        $\widehat{\nabla}_{I_k}(y_k^s) = g_{I_k}(y_k^s) + \frac{m\theta_s}{\eta} Q_s(\overline{z}^{s-1} - \widetilde{x}^{s-1})$;
        $z_k^s = \arg\min_z \Big\{ \langle \widehat{\nabla}_{I_k}(y_k^s),\ z \rangle + \phi_k^s(z, w_k^s)$
                  $+ \frac{m\theta_s}{2\eta}\|z - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|_{Q_s}^2 \Big\}$;
        $x_k^s = \frac{\theta_s}{m} z_k^s + (1 - \frac{\theta_s}{m})\overline{x}^{s-1}$, $\lambda_k^s = Az_k^s - w_k^s + \lambda_{k-1}^s$;
    end for
    $\widetilde{\lambda}^s = \frac{1}{m}\sum_{k=1}^{m} \lambda_k^s$, $\theta_s = \frac{\sqrt{\theta_{s-1}^4 + 4\theta_{s-1}^2} - \theta_{s-1}^2}{2}$,
    $\widetilde{x}^s = \frac{1}{m}\sum_{k=1}^{m} x_k^s$, $\widetilde{w}^s = \frac{\theta_s}{m^2}\sum_{k=1}^{m} w_k^s + \left(1 - \frac{\theta_s}{m}\right)\overline{w}^{s-1}$.
end for
Output: $\widetilde{x}^S$, $\widetilde{y}^S$.
```

4.2. Stochastic Update Rules in Inner Loop

Like Algorithm 1, DAVIS-ADMM uses the same momentum accelerated rules in (4) and (5) for y_k^s and x_k^s. Unlike existing accelerated algorithms, the weight in DAVIS-ADMM is θ_s/m. Then we can remove the constraint (i.e., $\theta_s \leq 1 - \frac{L\eta}{1-L\eta}$) for θ_s in (Liu et al., 2021). That is, the initial value is set to $\theta_1 = 1$ for our DAVIS-ADMM, while that in (Liu et al., 2021) requires to satisfy the condition $\theta_1 \leq 1 - \frac{L\eta}{1-L\eta}$, which is a reason that DAVIS-ADMM improves the convergence rate from $O(1/S)$ to $O(1/nS)$, as shown in Table 2. The update rule of θ_s is $\theta_s = (\sqrt{\theta_{s-1}^4 + 4\theta_{s-1}^2} - \theta_{s-1}^2)/2$, and let $\phi_k^s(z,w) = \frac{\beta}{2}\|Az - w + \lambda_{k-1}^s\|^2$. For the equality-constrained problem (2), the mini-batch compensated stochastic gradient estimator $\widehat{\nabla}_{I_k}(x)$ is defined as follows.

Definition 4.1 (Mini-batch compensated stochastic gradient estimator for Problem (2))**.**

$$\widehat{\nabla}_{I_k}(x) = g_{I_k}(x) + \frac{m\theta_s}{\eta} Q_s(\overline{z}^{s-1} - \widetilde{x}^{s-1}),$$

where $g_{I_k}(x) = \nabla f_{I_k}(x) - \nabla f_{I_k}(\overline{x}^{s-1}) + \nabla f(\overline{x}^{s-1})$, and $I_k \subset \{1, 2, \ldots, n\}$ is a randomly chosen mini-batch of size b. Compared with the gradient estimator in Definition 3.1, an additional matrix Q_s is introduced into the estimator. When $b = 1$, the estimator becomes $\widehat{\nabla}_{i_k}(x)$.

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 λ^* 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: $\phi(\widetilde{x}^S, \widetilde{w}^S) = P(\widetilde{x}^S, \widetilde{w}^S) + \delta\|A\widetilde{x}^S - \widetilde{w}^S\| \geq 0$, where $P(x, w) = f(x) - f(x^*) - \nabla f(x^*)^T(x - x^*) + h(w) - h(w^*) - \widehat{\nabla} h(w^*)^T(w - w^*)$, and $\widehat{\nabla} h(w)$ 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 Λ are compact convex sets with diameters $D_{\mathcal{X}} = \sup_{x_1, x_2 \in \mathcal{X}} \|x_1 - x_2\|$ 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 $\mathcal{O}(1/S^2)$ 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* $c_1 = 2\|A^TA\|_2\|x^* - \widetilde{x}^0\|^2 + 2\|\lambda^* - \widetilde{\lambda}^0\|^2 + 8\delta^2 + 10\|\lambda^*\|^2$ *and choose* $m = \Theta(n)$*, then*

$$\mathbb{E}\big[\phi(\widetilde{x}^S, \widetilde{w}^S)\big] \leq \mathcal{O}\Big(\frac{2\phi(\widetilde{x}^0, \widetilde{w}^0) + \|x^* - \widetilde{x}^0\|^2_{Q_1}/\eta}{n(S+1)} + \frac{c_1\beta}{n(S+1)}\Big).$$

Remark 3. Theorem 4.3 shows that without the boundedness assumption, the convergence rate of DAVIS-ADMM is $\mathcal{O}(1/(nS))$, while the best-known convergence result as in (Liu et al., 2021) with a strong boundedness assumption is $\mathcal{O}(1/S^2)$, as shown in Table 2. That is, DAVIS-ADMM improves the best-known convergence rate from $\mathcal{O}(1/S^2)$ to $\mathcal{O}(1/(nS))$ 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 Λ are bounded with the diameters $D_{\mathcal{X}}$ and D_{Λ}. 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 $\mathcal{O}(1/S)$, which is consistent with the optimal convergence rate of accelerated deterministic methods such as (Ouyang & Xu, 2020).

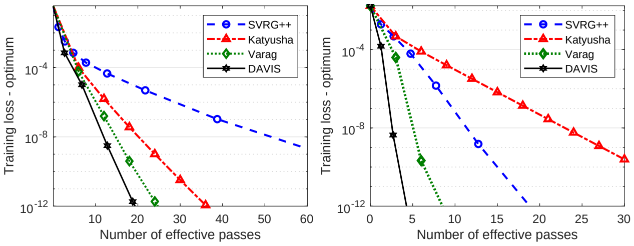

Figure 1. Comparison of all the methods for solving ℓ_1-norm regularized logistic regression problems on Adult and Covtype.

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. ℓ_1-Norm Regularized Logistic Regression

We first apply Algorithm 1 to solve the binary non-strongly convex ℓ_1-norm regularized logistic regression problem,

$$\min_x \Big\{\frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(-b_i a_i^T x\right)\right) + \lambda\|x\|_1\Big\},$$

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:

$$\min_x \Big\{\frac{1}{n}\sum_{i=1}^{n} f_i(x) + \lambda\|Ax\|_1\Big\},$$

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.

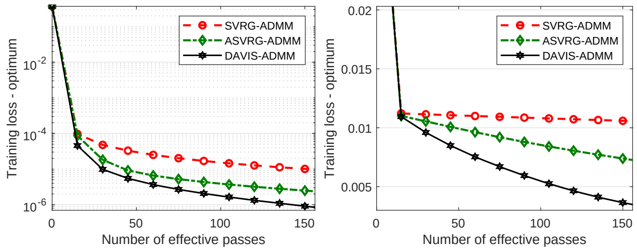

Figure 2. Comparison of all the methods for solving graph-guided fused Lasso problems on Adult and Covtype, where the regularization parameter is $\lambda = 10^{-5}$.

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., $\mathcal{O}(1/nS^2)$ vs. $\mathcal{O}(1/S^2)$, and a factor $\frac{n}{S}$ faster than accelerated stochastic ADMM algorithms for Problem (2), i.e., $\mathcal{O}(1/nS)$ vs. $\mathcal{O}(1/S^2)$.

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

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. *J. Mach. Learn. Res.*, 18: 1–51, 2018.

Allen-Zhu, Z. and Yuan, Y. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *ICML*, pp. 1080–1089, 2016.

Azadi, S. and Sra, S. Towards an optimal stochastic alternating direction method of multipliers. In *ICML*, pp. 620–628, 2014.

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516, 2008.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011.

Bubeck, S. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8:231–358, 2015.

Defazio, A. A simple practical accelerated method for finite sums. In *NIPS*, pp. 676–684, 2016.

Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pp. 1646–1654, 2014.

Frostig, R., Ge, R., Kakade, S. M., and Sidford, A. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *ICML*, pp. 2540–2548, 2015.

Goldstein, T., Li, M., and Yuan, X. Adaptive primal-dual splitting methods for statistical learning and image processing. In *NIPS*, pp. 2089–2097, 2015.

He, B. and Yuan, X. Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. *SIAM J. Imaging Sciences*, 5(1):119–149, 2012.

Hien, L., Lu, C., Xu, H., and Feng, J. Accelerated stochastic mirror descent algorithms for composite non-strongly convex optimization. *J. Optimiz. Theory App.*, 2019.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pp. 315–323, 2013.

Kim, D. and Fessler, J. A. Optimized first-order methods for smooth convex minimization. *Math. Program.*, 159: 81–107, 2016.

Kim, S., Sohn, K. A., and Xing, E. P. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, 25:204–212, 2009.

Lan, G. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer, Cham, 1st edition, 2020.

Lan, G. and Zhou, Y. Random gradient extrapolation for distributed and stochastic optimization. *SIAM J. Optim.*, 28(4):2753–2782, 2018a.

Lan, G. and Zhou, Y. An optimal randomized incremental gradient method. *Math. Program.*, 171:167–215, 2018b.

Lan, G., Li, Z., and Zhou, Y. A unified variance-reduced accelerated gradient method for convex optimization. In *NeurIPS*, 2019.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, S-tanford, CA, 2000. Morgan Kaufmann.

Li, H. and Lin, Z. Faster and non-ergodic $o(1/k)$ stochastic alternating direction method of multipliers. In *NIPS*, pp. 4479–4488, 2017.

Li, Z. ANITA: An optimal loopless accelerated variance-reduced gradient method. 2021.

Lin, H., Mairal, J., and Harchaoui, Z. A universal catalyst for first-order optimization. In *NIPS*, pp. 3366–3374, 2015a.

Lin, Q., Lu, Z., and Xiao, L. An accelerated proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM J. Optim.*, 25(4): 2244–2273, 2015b.

Liu, Y., Shang, F., Liu, H., Kong, L., Jiao, L., and Lin, Z. Accelerated variance reduction stochastic ADMM for large-scale machine learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4242–4255, 2021.

Mahdavi, M., Zhang, L., and Jin, R. Mixed optimization for smooth functions. In *NIPS*, pp. 674–682, 2013.

Murata, T. and Suzuki, T. Doubly accelerated stochastic variance reduced dual averaging method for regularized empirical risk minimization. In *NIPS*, pp. 608–617, 2017.

Nesterov, Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Math. Doklady*, 27:372–376, 1983.

Nesterov, Y. Gradient methods for minimizing composite functions. *Math. Program.*, 140:125–161, 2013.

Nitanda, A. Stochastic proximal gradient descent with acceleration techniques. In *NIPS*, pp. 1574–1582, 2014.

Ouyang, H., He, N., Tran, L. Q., and Gray, A. Stochastic alternating direction method of multipliers. In *ICML*, pp. 80–88, 2013.

Ouyang, Y. and Xu, Y. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Math. Program.*, 2020.

Polyak, B. Some methods of speeding up the convergence of iteration methods. *USSR Comp. Math. Math. Phys.*, 4 (5):1–17, 1964.

Robbins, H. and Monro, S. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 1951.

Roux, N. L., Schmidt, M., and Bach, F. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pp. 2672–2680, 2012.

Shalev-Shwartz, S. and Zhang, T. Stochastic dual coordinate ascent methods for regularized loss minimization. *J. Mach. Learn. Res.*, 14:567–599, 2013.

Song, C., Jiang, Y., and Ma, Y. Variance reduction via accelerated dual averaging for finite-sum optimization. In *NeurIPS*, pp. 315–323, 2020.

Suzuki, T. Stochastic dual coordinate ascent with alternating direction method of multipliers. In *ICML*, pp. 736–744, 2014.

Tibshirani, R. J. and Taylor, J. The solution path of the generalized lasso. *Annals of Statistics*, 39(3):1335–1371, 2011.

Wang, H. and Banerjee, A. Online alternating direction method. In *ICML*, pp. 1119–1126, 2012.

Wang, M., Liu, J., and Fang, E. X. Accelerating stochastic composition optimization. *J. Mach. Learn. Res.*, 18:1–23, 2017.

Woodworth, B. and Srebro, N. Tight complexity bounds for optimizing composite objectives. In *NIPS*, 2016.

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

Xie, G., Luo, L., Lian, Y., and Zhang, Z. Lower complexity bounds for finite-sum convex-concave minimax optimization problems. In *ICML*, pp. 10504–10513, 2020.

Xu, Y., Liu, M., Lin, Q., and Yang, T. ADMM without a fixed penalty parameter: Faster convergence with new adaptive penalization. In *NIPS*, pp. 1267–1277, 2017.

Zhang, L., Mahdavi, M., and Jin, R. Linear convergence with condition number independent access of full gradients. In *NIPS*, pp. 980–988, 2013.

Zhang, Y. and Xiao, L. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *ICML*, pp. 353–361, 2015.

Zheng, S. and Kwok, J. T. Fast-and-light stochastic ADMM. In *IJCAI*, pp. 2407–2613, 2016.

Zhong, L. W. and Kwok, J. T. Fast stochastic alternating direction method of multipliers. In *ICML*, pp. 46–54, 2014.

Zhou, K., Shang, F., and Cheng, J. A simple stochastic variance reduced algorithm with fast convergence rates. In *ICML*, pp. 5975–5984, 2018.

Zhu, M. and Chan, T. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. Technical report, 2008.

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:*

$$\widetilde{\nabla} f_{i_k}(x_k^s) = \nabla f_{i_k}(x_k^s) - \nabla f_{i_k}(\overline{x}^{s-1}) + \nabla f(\overline{x}^{s-1}).$$

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

$$\begin{aligned}&\mathbb{E}\left[\left\|\widetilde{\nabla} f_{i_k}(x_k^s) - \nabla f(x_k^s)\right\|^2\right]\\ &\leq 2L\big[f(\overline{x}^{s-1}) - f(x_k^s) + \langle \nabla f(x_k^s),\ x_k^s - \overline{x}^{s-1}\rangle\big]\,.\end{aligned} \tag{7}$$

The convergence analysis for the proposed algorithms requires the above upper bound on the term $\mathbb{E}[\|\widetilde{\nabla} f_{i_k}(x_k^s) - \nabla f(x_k^s)\|^2]$ 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*

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

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

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 $\mathcal{R}$ 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 $\mathcal{C}$ is introduced in the upper bound in Lemma 3.4, which can be used to offset by the residual term $\mathcal{R}$ 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,

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

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

$$\nabla f(\widetilde{x}^{s-1}) + \xi + \frac{\theta_s}{m\eta}(\overline{z}^{s-1} - \widetilde{x}^{s-1}) = 0, \tag{8}$$

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

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

$$\begin{aligned} F(\overline{x}^{s-1}) &= h(\overline{x}^{s-1}) + f(\overline{x}^{s-1}) \\ &\leq h(\overline{x}^{s-1}) + f(\widetilde{x}^{s-1}) + \left\langle \nabla f(\widetilde{x}^{s-1}),\, \overline{x}^{s-1} - \widetilde{x}^{s-1} \right\rangle + \frac{L}{2}\|\overline{x}^{s-1} - \widetilde{x}^{s-1}\|^2 \\ &\leq h(\overline{x}^{s-1}) + f(\widetilde{x}^{s-1}) + \theta_s \left\langle \nabla f(\widetilde{x}^{s-1}),\, x^* - \widetilde{x}^{s-1} \right\rangle + \frac{\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2 \\ &\quad - \theta_s \left\langle \nabla f(\widetilde{x}^{s-1}),\, x^* - \overline{z}^{s-1} \right\rangle \\ &= h(\overline{x}^{s-1}) + f(\widetilde{x}^{s-1}) + \theta_s \left\langle \nabla f(\widetilde{x}^{s-1}),\, x^* - \widetilde{x}^{s-1} \right\rangle + \frac{\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2 \\ &\quad + \theta_s \left\langle \xi + \frac{\theta_s}{m\eta}(\overline{z}^{s-1} - \widetilde{x}^{s-1}),\, x^* - \overline{z}^{s-1} \right\rangle \\ &\leq \theta_s F(x^*) + (1-\theta_s)F(\widetilde{x}^{s-1}) + \frac{\theta_s^2}{2m\eta}\left(\|x^* - \widetilde{x}^{s-1}\|^2 - \|x^* - \overline{z}^{s-1}\|^2\right) \\ &\quad + \frac{\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2, \end{aligned} \tag{9}$$

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

$$\frac{\theta_s^2}{m\eta}\left\langle \overline{z}^{s-1} - \widetilde{x}^{s-1},\, x^* - \overline{z}^{s-1} \right\rangle = \frac{\theta_s^2}{2m\eta}\left(\|x^* - \widetilde{x}^{s-1}\|^2 - \|x^* - \overline{z}^{s-1}\|^2 - \|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2\right).$$

Note that the above equality holds due to Property 1.

Therefore, we have

$$\begin{aligned} &F(\overline{x}^{s-1}) - F(x^*) \\ \leq\ &(1-\theta_s)(F(\widetilde{x}^{s-1}) - F(x^*)) + \frac{\theta_s^2}{2m\eta}\left(\|x^* - \widetilde{x}^{s-1}\|^2 - \|x^* - \overline{z}^{s-1}\|^2\right) + \mathcal{R}^s. \end{aligned} \tag{10}$$

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 $g_k^s = \nabla f_{i_k}(y_k^s) - \nabla f_{i_k}(\overline{x}^{s-1}) + \nabla f(\overline{x}^{s-1})$, our compensated stochastic variance reduction gradient estimator is rewritten as follows:

$$\widetilde{\nabla}_{i_k}(y_k^s) = g_k^s + \frac{m\theta_s}{\eta}\left(\overline{z}^{s-1} - \widetilde{x}^{s-1}\right).$$

And the update rule of z_k^s is

$$z_k^s \triangleq \arg\min_z \left\{ h(z) + \left\langle \widetilde{\nabla}_{i_k}(y_k^s),\ z \right\rangle + \frac{m\theta_s}{2\eta}\left\| z - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1}) \right\|^2 \right\},$$

which implies that z_k^s is required to satisfy the following optimal condition:

$$\widetilde{\nabla}_{i_k}(y_k^s) + \zeta_k^s + \frac{m\theta_s}{\eta}\left[z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1}) \right] = 0, \tag{11}$$

where $\zeta_k^s \in \partial h(z_k^s)$.

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

$$\begin{aligned} y_k^s &= \frac{\theta_s}{m} p_k^s + \left(1 - \frac{\theta_s}{m}\right)\overline{x}^{s-1}, \\ x_k^s &= \frac{\theta_s}{m}(z_k^s - p_k^s) + y_k^s \\ &= \frac{\theta_s}{m} z_k^s + \left(1 - \frac{\theta_s}{m}\right)\overline{x}^{s-1}. \end{aligned} \tag{12}$$

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

$$
\begin{aligned}
&\mathbb{E}[F(x_k^s)]\\
\le\;& \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]\\
\overset{\text{a}}{=}\;& \mathbb{E}\left[h(x_k^s) + f(y_k^s) + \left\langle \nabla f(y_k^s),\, \frac{\theta_s}{m}\left(z_k^s - p_k^s\right)\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 \widetilde{\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) - \widetilde{\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 \widetilde{\nabla}_{i_k}(y_k^s),\, x^* - p_k^s\right\rangle - \frac{\theta_s}{m}\left\langle \widetilde{\nabla}_{i_k}(y_k^s),\, x^* - z_k^s\right\rangle\right]\\
&+\mathbb{E}\left[\left\langle \nabla f(y_k^s) - \widetilde{\nabla}_{i_k}(y_k^s),\, \frac{\theta_s}{m}\left(z_k^s - p_k^s\right)\right\rangle + \frac{L\theta_s^2}{2m^2}\|z_k^s - p_k^s\|^2\right]\\
\overset{\text{b}}{=}\;& \mathbb{E}\left[h(x_k^s) + f(y_k^s) + \frac{\theta_s}{m}\left\langle \widetilde{\nabla}_{i_k}(y_k^s),\, x^* - p_k^s\right\rangle\right]\\
&+\frac{\theta_s^2}{2\eta}\left(\|x^* - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 - \|x^* - z_k^s\|^2 - \|z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2\right)\\
&+\mathbb{E}\left[\frac{\theta_s}{m}\langle \zeta_k^s,\, x^* - z_k^s\rangle\right] + \mathbb{E}\left[\left\langle \nabla f(y_k^s) - \widetilde{\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[f(y_k^s) + \frac{\theta_s^2}{2\eta}\left(\|x^* - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 - \|x^* - z_k^s\|^2 - \|z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2\right)\right]\\
&+\underbrace{\mathbb{E}\left[h(x_k^s) + \frac{\theta_s}{m}\langle \zeta_k^s,\, x^* - z_k^s\rangle\right]}_{A_1} + \underbrace{\mathbb{E}\left[\frac{\theta_s}{m}\left\langle \widetilde{\nabla}_{i_k}(y_k^s),\, x^* - p_k^s\right\rangle\right]}_{A_2}\\
&+\underbrace{\mathbb{E}\left[\left\langle \nabla f(y_k^s) - \widetilde{\nabla}_{i_k}(y_k^s),\, \frac{\theta_s}{m}(z_k^s - p_k^s)\right\rangle\right]}_{A_3} + \frac{L\theta_s^2}{2m^2}\|z_k^s - p_k^s\|^2,
\end{aligned}
\tag{13}
$$

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

$$
\begin{aligned}
&\frac{\theta_s}{m}\left\langle -\widetilde{\nabla}_{i_k}(y_k^s),\, x^* - z_k^s\right\rangle\\
=\;& \frac{\theta_s^2}{\eta}\left\langle z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1}),\, x^* - z_k^s\right\rangle + \frac{\theta_s}{m}\langle \zeta_k^s,\, x^* - z_k^s\rangle\\
=\;& \frac{\theta_s^2}{2\eta}\left(\left\|x^* - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\right\|^2 - \|x^* - z_k^s\|^2 - \left\|z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\right\|^2\right)\\
&+\frac{\theta_s}{m}\langle \zeta_k^s,\, x^* - z_k^s\rangle.
\end{aligned}
$$

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 rule of x_k^s in Algorithm 1, we have

$$\begin{aligned}
A_1 \;&= \mathbb{E}\left[h(x_k^s) + \frac{\theta_s}{m}\left\langle \zeta_k^s,\, x^* - z_k^s\right\rangle\right]\\
&= \mathbb{E}\left[h\left(\frac{\theta_s}{m}z_k^s + \left(1-\frac{\theta_s}{m}\right)\overline{x}^{s-1}\right) + \frac{\theta_s}{m}\left\langle \zeta_k^s,\, x^* - z_k^s\right\rangle\right]\\
&\leq \mathbb{E}\left[\frac{\theta_s}{m}h(z_k^s) + \left(1-\frac{\theta_s}{m}\right)h(\overline{x}^{s-1})\right]\\
&\quad + \mathbb{E}\left[\left\langle \zeta_k^s,\, \frac{\theta_s}{m}\left(x^* - z_k^s\right)\right\rangle\right]\\
&\leq \mathbb{E}\left[\frac{\theta_s}{m}h(z_k^s) + \left(1-\frac{\theta_s}{m}\right)h(\overline{x}^{s-1})\right]\\
&\quad + \frac{\theta_s}{m}\mathbb{E}[h(x^*) - h(z_k^s)]\\
&= \mathbb{E}\left[\left(1-\frac{\theta_s}{m}\right)h(\overline{x}^{s-1}) + \frac{\theta_s}{m}h(x^*)\right],
\end{aligned} \tag{14}$$

where the first inequality holds due to the convexity of $h(\cdot)$, and the second inequality follows from the facts that $\zeta_k^s \in \partial h(z_k^s)$ and $\langle \zeta_k^s,\ x^* - z_k^s\rangle \leq h(x^*) - h(z_k^s)$.

By the definition of $\widetilde{\nabla}_{i_k}(y_k^s)$ and Property 1, the term A_2 in the inequality (13) is rewritten as follows:

$$\begin{aligned}
A_2 &= \mathbb{E}\left[\frac{\theta_s}{m}\left\langle \widetilde{\nabla}_{i_k}(y_k^s),\ x^* - p_k^s\right\rangle\right]\\
&= \frac{\theta_s}{m}\left\langle \nabla f(y_k^s) + \frac{m\theta_s}{\eta}(\overline{z}^{s-1} - \widetilde{x}^{s-1}),\ x^* - p_k^s\right\rangle\\
&= \frac{\theta_s}{m}\left\langle \nabla f(y_k^s),\ x^* - p_k^s\right\rangle + \frac{\theta_s^2}{2\eta}\left(\|x^* - p_k^s\|^2 - \|x^* - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 + \|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2\right).
\end{aligned} \tag{15}$$

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

$$\begin{aligned}
A_3 &= \mathbb{E}\left[\frac{\theta_s}{m}\left\langle \nabla f(y_k^s) - \widetilde{\nabla}_{i_k}(y_k^s),\ z_k^s - p_k^s\right\rangle\right]\\
&= \frac{\theta_s}{m}\mathbb{E}\left[\left\langle \nabla f(y_k^s) - \widetilde{\nabla}_{i_k}(y_k^s),\ z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\right\rangle\right]\\
&\quad + \frac{\theta_s}{m}\mathbb{E}\left[\left\langle \nabla f(y_k^s) - \widetilde{\nabla}_{i_k}(y_k^s),\ 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\right\rangle\right]\\
&\overset{\text{a}}{=} \frac{\theta_s}{m}\mathbb{E}\left[\left\langle \nabla f(y_k^s) - \widetilde{\nabla}_{i_k}(y_k^s), z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\right\rangle\right] - \frac{2\theta_s^2}{\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2\\
&\overset{\text{b}}{\leq} \frac{\eta}{2m^2}\left\|\nabla f(y_k^s) - \widetilde{\nabla}_{i_k}(y_k^s)\right\|^2 + \frac{\theta_s^2}{2\eta}\|z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 - \frac{2\theta_s^2}{\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2\\
&\overset{\text{c}}{\leq} \frac{1}{m}\left[f(\overline{x}^{s-1}) - f(y_k^s) + \langle \nabla f(y_k^s),\ y_k^s - \overline{x}^{s-1}\rangle\right]\\
&\quad + \frac{\theta_s^2}{2\eta}\|z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 - \frac{3\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2,
\end{aligned} \tag{16}$$

where the equality $\overset{\text{a}}{=}$ holds due to the definition of the gradient estimator in Definition 1, and the facts $\mathbb{E}[\nabla f(y_k^s) - g_k^s] = 0$

and

$$\begin{aligned}
&\frac{\theta_s}{m}\mathbb{E}\Big[\Big\langle \nabla f(y_k^s) - \widetilde{\nabla}_{i_k}(y_k^s),\ 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\Big\rangle\Big]\\
=&\frac{\theta_s}{m}\mathbb{E}\Big[\Big\langle \nabla f(y_k^s) - g_k^s - \frac{\theta_s}{\eta}\left(\overline{z}^{s-1} - \widetilde{x}^{s-1}\right),\ 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\Big\rangle\Big]\\
=&-\frac{2\theta_s^2}{\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2.
\end{aligned}$$

Note that the equality $\overset{\mathrm{b}}{\le}$ in (16) follows from the Young's inequality, and the inequality $\overset{\mathrm{c}}{\le}$ in (16) holds due to the following fact

$$\begin{aligned}
&\frac{\eta}{2m^2}\mathbb{E}\left[\left\|\nabla f(y_k^s) - \widetilde{\nabla}_{i_k}(y_k^s)\right\|^2\right]\\
=&\frac{\eta}{2m^2}\mathbb{E}\left[\|\nabla f(y_k^s) - g_k^s\|^2\right] + \frac{\theta_s^2}{2\eta}\mathbb{E}\left[\left\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\right\|^2\right]\\
&+\frac{\theta_s}{2m}\mathbb{E}\left[\langle \nabla f(y_k^s) - g_k^s,\ \overline{z}^{s-1} - \widetilde{x}^{s-1}\rangle\right]\\
=&\frac{\eta}{2m}\mathbb{E}\left[\|\nabla f(y_k^s) - g_k^s\|^2\right] + \frac{\theta_s^2}{2\eta}\left\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\right\|^2\\
\le&\frac{1}{m}\left[f(\overline{x}^{s-1}) - f(y_k^s) + \langle \nabla f(y_k^s),\ y_k^s - \overline{x}^{s-1}\rangle\right] + \frac{\theta_s^2}{2\eta}\left\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\right\|^2,
\end{aligned}$$

where the first equality holds due to the definition of our gradient operator, the second equality holds due to the fact $\mathbb{E}[\nabla f(y_k^s) - g_k^s] = 0$, and the inequality follows from Lemma 4 with the setting $\eta \le 1/L$, i.e., $L\eta \le 1$.

Combing the equality (15) and the inequality (16), we have

$$\begin{aligned}
&A_2 + A_3\\
\le&\left\langle \nabla f(y_k^s),\ \frac{\theta_s}{m}\left(x^* - p_k^s\right) + \frac{1}{m}(y_k^s - \widetilde{x}^{s-1})\right\rangle\\
&+\frac{\theta_s^2}{2\eta}\|z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 - \frac{\theta_s^2}{\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2\\
&+\frac{\theta_s^2}{2\eta}\left(\|x^* - p_k^s\|^2 - \|x^* - p_k^s - 2(\overline{z}^s - \widetilde{x}^s)\|^2\right)\\
\le&\frac{\theta_s}{m}f(x^*) + \left(1 - \frac{\theta_s}{m}\right)f(\overline{x}^{s-1}) - f(y_k^s)\\
&+\frac{\theta_s^2}{2\eta}\|z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 - \frac{\theta_s^2}{\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2\\
&+\frac{\theta_s^2}{2\eta}\left(\|x^* - p_k^s\|^2 - \|x^* - p_k^s - 2(\overline{z}^s - \widetilde{x}^s)\|^2\right),
\end{aligned}\tag{17}$$

where the first inequality follows from the updated rule of $y_k^s = \frac{\theta_s}{m}p_k^s + \left(1 - \frac{\theta_s}{m}\right)\overline{x}^{s-1}$ and the following fact:

$$\begin{aligned}
&\left\langle \nabla f(y_k^s),\ \frac{\theta_s}{m}\left(x^* - p_k^s\right) + \frac{1}{m}(y_k^s - \overline{x}^{s-1})\right\rangle\\
=&\left\langle \nabla f(y_k^s),\ \frac{\theta_s}{m}x^* + \left(1 - \frac{\theta_s}{m} - \frac{1}{m}\right)\overline{x}^{s-1} + \frac{1}{m}y_k^s - y_k^s\right\rangle + \frac{1}{m}\left[f(\overline{x}^{s-1}) - f(y_k^s)\right]\\
\le&f\left(\frac{\theta_s}{m}x^* + \left(1 - \frac{\theta_s}{m} - \frac{1}{m}\right)\overline{x}^{s-1} + \frac{1}{m}y_k^s\right) - f(y_k^s) + \frac{1}{m}\left[f(\overline{x}^{s-1}) - f(y_k^s)\right]\\
\le&\frac{\theta_s}{m}f(x^*) + \left(1 - \frac{\theta_s}{m}\right)f(\overline{x}^{s-1}) - f(y_k^s).
\end{aligned}\tag{18}$$

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

$$\begin{aligned}
&\mathbb{E}[F(x_k^s)-F(x^*)]\\
\leq\; &\left(1-\frac{\theta_s}{m}\right)\left[F(\overline{x}^{s-1})-F(x^*)\right]-\left(\frac{\theta_s^2}{2\eta}-\frac{\theta_s^2}{2m\eta}\right)\|\overline{z}^{s-1}-x^{s-1}\|^2\\
&+\frac{\theta_s^2}{2\eta}\mathbb{E}\left(\|x^*-p_k^s\|^2-\|x^*-p_k^s-2(\overline{z}^s-\widetilde{x}^s)\|^2\right)\\
&+\frac{\theta_s^2}{2\eta}\mathbb{E}\left(\|x^*-p_k^s+2(\overline{z}^s-\widetilde{x}^s)\|^2-\|x^*-z_k^s\|^2\right)\\
=\; &\left(1-\frac{\theta_s}{m}\right)\left[F(\overline{x}^{s-1})-F(x^*)\right]+\mathcal{C}^s\\
&+\frac{\theta_s^2}{2\eta}\mathbb{E}\left(\|x^*-z_{k-1}^s\|^2-\|x^*-z_k^s\|^2\right),
\end{aligned} \tag{19}$$

where $p_k^s=z_{k-1}^s$. This completes the proof. □

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

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

$$\begin{aligned}
&\mathbb{E}[F(x_k^s)-F(x^*)]\\
\leq &\left(1-\frac{\theta_s}{m}\right)\left[F(\overline{x}^{s-1})-F(x^*)\right]+\mathcal{C}^s+\frac{\theta_s^2}{2\eta}\left(\|x^*-z_{k-1}^s\|^2-\|x^*-z_k^s\|^2\right).
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[F(\widetilde{x}^s)-F(x^*)]\leq &\left(1-\frac{\theta_s}{m}\right)\left[F(\overline{x}^{s-1})-F(x^*)\right]+\mathcal{C}^s\\
&+\frac{\theta_s^2}{2m\eta}\left(\|x^*-z_0^s\|^2-\|x^*-z_m^s\|^2\right).
\end{aligned} \tag{20}$$

Furthermore, by using Lemma 3.3, we have

$$\begin{aligned}
&F(\overline{x}^{s-1})-F(x^*)\\
\leq&(1-\theta_s)(F(\widetilde{x}^{s-1})-F(x^*))+\frac{\theta_s^2}{2m\eta}\left(\|x^*-\widetilde{x}^{s-1}\|^2-\|x^*-\overline{z}^{s-1}\|^2\right)+\mathcal{R}^s.
\end{aligned} \tag{21}$$

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

$$\begin{aligned}
&\mathbb{E}[F(\widetilde{x}^s)-F(x^*)]\\
\leq\; &(1-\theta_s)(F(\widetilde{x}^{s-1})-F(x^*))+\frac{\theta_s^2}{2m\eta}\left(\|x^*-\widetilde{x}^{s-1}\|^2-\|x^*-\widetilde{x}^s\|^2\right).
\end{aligned} \tag{22}$$

This completes the proof. □

Proof of Theorem 3.6

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

Proof of Theorem 3.6:

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

$$\begin{aligned}&\frac{1}{\theta_S^2}\mathbb{E}\left[F(\widetilde{x}^S)-F(x^*)\right]\\ \leq\;&\frac{(1-\theta_1)}{\theta_1^2}\left[F(\widetilde{x}^0)-F(x^*)\right]+\frac{\theta_1^2}{2m\eta}\left\|x^*-\widetilde{x}^0\right\|^2.\end{aligned}$$

Since $\theta_1=1$,

$$\frac{1}{\theta_S^2}\mathbb{E}\big[F(\widetilde{x}^S)-F(x^*)\big]\leq\frac{1}{2m\eta}\left\|x^*-\widetilde{x}^0\right\|^2. \tag{23}$$

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

$$\mathbb{E}\big[F(\widetilde{x}^S)-F(x^*)\big]\leq\mathcal{O}\left(\frac{\left\|x^*-\widetilde{x}^0\right\|^2}{mS^2\eta}\right). \tag{24}$$

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

This completes the proof. □

Appendix C: Theoretical Analysis for DAVIS-ADMM in Section 4

In this section, we analyze the convergence properties of the proposed DAVIS-ADMM algorithm (i.e., Algorithm 2). Similar to Theorem 3.6, the proof of Theorem 4.3 for DAVIS-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 DAVIS-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*

$$\begin{aligned}&\mathbb{E}\big[-(Ax_k-w_k)^T(\varphi_k-\varphi)\big]\\ =\;&\frac{\beta}{2}\mathbb{E}\big[\|\lambda_{k-1}-\lambda^*-\lambda\|^2-\|\lambda_k-\lambda^*-\lambda\|^2-\|\lambda_k-\lambda_{k-1}\|^2\big].\end{aligned}$$

Lemma 6 (Upper bound of double snapshot update). *Suppose that Assumption 1 holds. Let $\{\overline{x}^s, \overline{w}^s, \overline{\lambda}^s\}$ be the sequence generated by our deterministic gradient descent step in Algorithm 2, then we have*

$$
\begin{aligned}
&\mathbb{E}\bigg[P(\overline{x}^{s-1}, \overline{w}^{s-1}) - \frac{\theta_s}{m}\langle A\overline{z}^{s-1} - \overline{p}^{s-1},\ \varphi\rangle\bigg]\\
\leq\ & (1-\theta_s)\, P(\widetilde{x}^{s-1}, \widetilde{w}^{s-1}) + \mathcal{R}^s\\
&+ \frac{\beta\theta_s}{2m}\mathbb{E}\Big[\|\overline{\lambda}^{s-2} - \lambda^* - \lambda\|^2 - \|\overline{\lambda}^{s-1} - \lambda^* - \lambda\|^2\Big]\\
&+ \frac{\beta\theta_s}{2m}\mathbb{E}\big[\|A\overline{z}^{s-2} - w^*\|^2 - \|A\overline{z}^{s-1} - w^*\|^2\big]\\
&+ \frac{\theta_s^2}{2m\eta}\left(\|x^* - \widetilde{x}^{s-1}\|^2_{Q_s} - \|x^* - \overline{z}^{s-1}\|^2_{Q_s}\right).
\end{aligned}
$$

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

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

$$
\overline{z}^{s-1} = \arg\min_{z}\left\{\big\langle\nabla f(\widetilde{x}^{s-1}),\ z\big\rangle + \frac{\theta_s}{2m\eta}\big\|z - \widetilde{x}^{s-1}\big\|^2_{Q_s} + \frac{\beta}{2m}\|Az - \overline{p}^{s-1} + \overline{\lambda}^{s-2}\|^2\right\},
$$

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

$$
\nabla f(\widetilde{x}^{s-1}) + \frac{\theta_s}{m\eta}Q_s(\overline{z}^{s-1} - \widetilde{x}^{s-1}) + \frac{\beta}{m}A^T(A\overline{z}^{s-1} - \overline{p}^{s-1} + \overline{\lambda}^{s-2}) = 0. \tag{25}
$$

With $\overline{\lambda}^{s-1} = A\overline{z}^{s-1} - \overline{p}^{s-1} + \overline{\lambda}^{s-2}$, we have

$$
\nabla f(\widetilde{x}^{s-1}) + \frac{\theta_s}{m\eta}Q_s(\overline{z}^{s-1} - \widetilde{x}^{s-1}) + \frac{\beta}{m}A^T\overline{\lambda}^{s-1} = 0. \tag{26}
$$

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

$$
\begin{aligned}
&f(\overline{x}^{s-1})\\
\leq\ & f(\widetilde{x}^{s-1}) + \big\langle\nabla f(\widetilde{x}^{s-1}),\, \overline{x}^{s-1} - \widetilde{x}^{s-1}\big\rangle + \frac{L}{2}\|\overline{x}^{s-1} - \widetilde{x}^{s-1}\|^2\\
\leq\ & f(\widetilde{x}^{s-1}) + \theta_s\big\langle\nabla f(\widetilde{x}^{s-1}),\, \overline{z}^{s-1} - \widetilde{x}^{s-1}\big\rangle + \frac{\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2_{Q_s},
\end{aligned} \tag{27}
$$

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 $Q_s \succ I$. Furthermore, using the optimal condition in (26), we have

$$
\begin{aligned}
&f(\overline{x}^{s-1})\\
\leq\ & f(\widetilde{x}^{s-1}) + \frac{\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2_{Q_s}\\
&+ \theta_s\big\langle\nabla f(\widetilde{x}^{s-1}),\, \overline{z}^{s-1} - x^*\big\rangle + \theta_s\big\langle\nabla f(\widetilde{x}^{s-1}),\, x^* - \widetilde{x}^{s-1}\big\rangle\\
\leq\ & f(\widetilde{x}^{s-1}) + \frac{\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2_{Q_s}\\
&+ \left\langle\frac{\theta_s^2}{m\eta}Q_s(\overline{z}^{s-1} - \widetilde{x}^{s-1}) + \frac{\beta\theta_s}{m}A^T\overline{\lambda}^{s-1},\, x^* - \overline{z}^{s-1}\right\rangle + \theta_s\big\langle\nabla f(\widetilde{x}^{s-1}),\, x^* - \widetilde{x}^{s-1}\big\rangle\\
\leq\ & \theta_s f(x^*) + (1-\theta_s)f(\widetilde{x}^{s-1}) + \frac{\theta_s^2}{2m\eta}\left(\|x^* - \widetilde{x}^{s-1}\|^2_{Q_s} - \|x^* - \overline{z}^{s-1}\|^2_{Q_s}\right) + \frac{\beta\theta_s}{m}\left\langle A^T\overline{\lambda}^{s-1},\, x^* - \overline{z}^{s-1}\right\rangle + \mathcal{R}^s,
\end{aligned} \tag{28}
$$

where the last inequality follows from the convexities of $f(\cdot)$ and Property 1, i.e.,

$$\frac{\theta_s^2}{m\eta}\left\langle Q_s(\overline{z}^{s-1}-\widetilde{x}^{s-1}),\, x^*-\overline{z}^{s-1}\right\rangle = \frac{\theta_s^2}{2m\eta}\left(\|x^*-\widetilde{x}^{s-1}\|_{Q_s}^2 - \|x^*-\overline{z}^{s-1}\|_{Q_s}^2 - \|\overline{z}^{s-1}-\widetilde{x}^{s-1}\|_{Q_s}^2\right).$$

Using the optimality condition of Problem (2), i.e., $\nabla f(x^*)+\beta A^T\lambda^*=0$, and let $\varphi^{s-1}=\beta\left(\overline{\lambda}^{s-1}-\lambda^*\right)$, then the following result holds

$$\begin{aligned}
&\left\langle \beta A^T\overline{\lambda}^{s-1},\, x^*-\overline{z}^{s-1}\right\rangle\\
=&\left\langle \nabla f(x^*),\, \overline{z}^{s-1}-x^*\right\rangle + \left\langle \beta A^T\lambda^*,\, \overline{z}^{s-1}-x^*\right\rangle\\
&+\left\langle \beta A^T\overline{\lambda}^{s-1},\, x^*-\overline{z}^{s-1}\right\rangle\\
=&\left\langle \nabla f(x^*),\, \overline{z}^{s-1}-x^*\right\rangle + \left\langle A^T\varphi^{s-1},\, x^*-\overline{z}^{s-1}\right\rangle,
\end{aligned}\tag{29}$$

Substituting the above equality into (28) and $\overline{x}^{s-1}=\theta_s\overline{z}^{s-1}+(1-\theta_s)\widetilde{x}^{s-1}$. Then we have

$$\begin{aligned}
&f(\overline{x}^{s-1})-f(x^*)+\left\langle \nabla f(x^*),\, x^*-\overline{x}^{s-1}\right\rangle - \frac{\theta_s}{m}\left\langle A^T\varphi^{s-1},\, x^*-\overline{z}^{s-1}\right\rangle\\
\le&\,(1-\theta_s)\big(f(\widetilde{x}^{s-1})-F(x^*)+\left\langle \nabla f(x^*),\, x^*-\widetilde{x}^{s-1}\right\rangle\big)\\
&+\frac{\theta_s^2}{2m\eta}\left(\|x^*-\widetilde{x}^{s-1}\|_{Q_s}^2-\|x^*-\overline{z}^{s-1}\|_{Q_s}^2\right)+\mathcal{R}^s.
\end{aligned}\tag{30}$$

With $\overline{\lambda}^{s-1}=A\overline{z}^{s-1}-\overline{p}^{s-1}+\overline{\lambda}^{s-2}$ and the update rules in Eq.(6), and using Lemma 3 in ((Zheng & Kwok, 2016)), we obtain

$$\begin{aligned}
&h(\overline{p}^{s-1})-h(w^*)-\widehat{\nabla}h(w^*)^T(\overline{p}^{s-1}-w^*)-\left\langle \varphi^{s-1}, w^*-\overline{p}^{s-1}\right\rangle\\
\le&\frac{\beta}{2m}\left[\|A\overline{z}^{s-2}-\overline{p}^{s-1}\|^2-\|A\overline{z}^{s-1}-\overline{p}^{s-1}\|^2+\|\overline{\lambda}^{s-1}-\overline{\lambda}^{s-2}\|^2\right].
\end{aligned}\tag{31}$$

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

$$\begin{aligned}
&h(\overline{w}^{s-1})-h(w^*)+\langle\hat{\nabla}h(w^*),\, w^*-\overline{w}^{s-1}\rangle-\frac{\theta_s}{m}\left\langle \varphi^{s-1},\, w^*-\overline{p}^{s-1}\right\rangle\\
\le&\,(1-\theta_s)\big[h(\widetilde{w}^{s-1})-h(w^*)+\left\langle\hat{\nabla}h(w^*),\, w^*-\widetilde{w}^{s-1}\right\rangle-\left\langle\varphi^{s-1},\, w^*\right\rangle\big]\\
&+\frac{\beta}{2m}\left[\|A\overline{z}^{s-2}-\overline{p}^{s-1}\|^2-\|A\overline{z}^{s-1}-\overline{p}^{s-1}\|^2+\|\overline{\lambda}^{s-1}-\overline{\lambda}^{s-2}\|^2\right].
\end{aligned}\tag{32}$$

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

$$\begin{aligned}
&\langle A^T\varphi^{s-1},\, x^*-\overline{z}^{s-1}\rangle+\langle\varphi^{s-1},\, w^*-\overline{p}^{s-1}\rangle+\langle A\overline{z}^{s-1}-\overline{p}^{s-1},\, \varphi^{s-1}-\varphi\rangle\\
=&\,-\langle A\overline{z}^{s-1}-\overline{p}^{s-1},\, \varphi\rangle.
\end{aligned}\tag{33}$$

Using Lemma 5 and the updated rule of $\overline{\lambda}^{s-1}$ in Algorithm 2, we have

$$\begin{aligned}
&-\langle A\overline{z}^{s-1}-\overline{p}^{s-1},\, \varphi^{s-1}-\varphi\rangle\\
=&\frac{\beta}{2}\left(\|\overline{\lambda}^{s-2}-\lambda^*-\lambda\|^2-\|\overline{\lambda}^{s-1}-\lambda^*-\lambda\|^2-\|\overline{\lambda}^{s-2}-\overline{\lambda}^{s-1}\|^2\right),
\end{aligned}\tag{34}$$

where the equality holds due to Property 1.

Using the results in (30), (32) and (34), the definition of $P(x,y)$ (i.e., $P(x,w)\!=\!f(x)\!-\!f(x^{*})\!-\!\nabla f(x^{*})^{T}(x\!-\!x^{*})\!+\!h(w)\!-\!h(w^{*})\!-\!\hat{\nabla}h(w^{*})^{T}(w\!-\!w^{*})$) and the update rules in Algorithm 2, we have

$$\begin{aligned}
&\mathbb{E}\Big[P(\overline{x}^{s-1},\overline{w}^{s-1})-\frac{\theta_s}{m}\langle A\overline{z}^{s-1}-\overline{p}^{s-1},\ \varphi\rangle\Big]\\
\leq\ &(1-\theta_s)\,P(\widetilde{x}^{s-1},\widetilde{w}^{s-1})+\mathcal{R}^{s}\\
&+\frac{\beta\theta_s}{2m}\mathbb{E}\Big[\|\overline{\lambda}^{s-2}-\lambda^{*}-\lambda\|^{2}-\|\overline{\lambda}^{s-1}-\lambda^{*}-\lambda\|^{2}\Big]\\
&+\frac{\beta\theta_s}{2m}\mathbb{E}\big[\|A\overline{z}^{s-2}-w^{*}\|^{2}-\|A\overline{z}^{s-1}-w^{*}\|^{2}\big]\\
&+\frac{\theta_s^{2}}{2m\eta}\left(\|x^{*}-\widetilde{x}^{s-1}\|_{Q_s}^{2}-\|x^{*}-\overline{z}^{s-1}\|_{Q_s}^{2}\right).
\end{aligned}\tag{35}$$

This completes the proof. □

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

$$\begin{aligned}
&\mathbb{E}\Big[\|\nabla f(y_k^s)-g_k^s\|^{2}\Big]\\
\leq\ &\frac{2L(n\!-\!b)}{b(n\!-\!1)}\left[f(\overline{x}^{s-1})-f(y_k^s)+\big\langle\nabla f(y_k^s),\ y_k^s-\overline{x}^{s-1}\big\rangle\right].
\end{aligned}$$

Lemma 8 (Upper bound of one-iteration of f)**.** *Suppose that Assumption 1 holds. Let $\{\widetilde{x}^{s},\widetilde{w}^{s},\widetilde{\lambda}^{s}\}$ be sequence generated by Algorithm 2, then we have*

$$\begin{aligned}
&\mathbb{E}\left[f(\widetilde{x}^{s})-f(x^{*})+\langle\nabla f(x^{*}),\ x^{*}-\widetilde{x}^{s}\rangle-\frac{\theta_s}{m^{2}}\sum_{k=1}^{m}\langle A^{T}\varphi_k^s,\ x^{*}-x^{s-1}+v_k^s-z_k^s\rangle\right]\\
\leq\ &\mathbb{E}\left[\left(1-\frac{\theta_s}{m}\right)\left[f(\overline{x}^{s-1})-f(x^{*})+\langle\nabla f(x^{*}),\ x^{*}-\overline{x}^{s-1}\rangle\right]\right]\\
&+\mathbb{E}\left[\frac{\theta_s^{2}}{2m\eta}\left(\|x^{*}-z_0^s\|_{Q_s}^{2}-\|x^{*}-z_m^s\|_{Q_s}^{2}\right)\right]+\mathcal{C}^{s},
\end{aligned}\tag{36}$$

where $\mathcal{C}^{s}=\left(\frac{\theta_s^{2}}{2m\eta}-\frac{\theta_s^{2}}{2\eta}\right)\|\overline{z}^{s-1}-\widetilde{x}^{s-1}\|_{Q_s}^{2}$ and $\varphi_k^s=\beta(\lambda_k^s-\lambda^{*})$.

Proof. 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 $g_{I_k}(y_k^s)=\nabla f_{I_k}(y_k^s)-\nabla f_{I_k}(\widetilde{x}^{s-1})+\nabla f(\widetilde{x}^{s-1})$, we have $\widehat{\nabla}_{I_k}(y_k^s)=g_{I_k}(y_k^s)+\frac{m\theta_s}{\eta}Q_s(\overline{z}^{s-1}-\widetilde{x}^{s-1})$. Since the function $f(\cdot)$ is L-smooth, and by using the update rule of x_k^s and the similar derivation as in Lemma 3.4, we have

$$\begin{aligned}
f(x_k^s) \leq\, & f(y_k^s) + \left\langle \nabla f(y_k^s),\, \frac{\theta_s}{m}(z_k^s - p_k^s)\right\rangle + \frac{L\theta_s^2}{2m^2}\|z_k^s - p_k^s\|^2 \\
\overset{\text{a}}{=}\, & f(y_k^s) + \frac{\theta_s}{m}\left\langle \widehat{\nabla}_{I_k}(y_k^s),\, x^* - p_k^s\right\rangle + \frac{\theta_s}{m}\left\langle \frac{m\theta_s}{\eta}Q_s(z_k^s - p_k^s) + \beta A^T\lambda_k^s,\, x^* - z_k^s\right\rangle \\
& + \left\langle \nabla f(y_k^s) - \widehat{\nabla}_{I_k}(y_k^s),\, \frac{\theta_s}{m}(z_k^s - p_k^s)\right\rangle + \frac{L\theta_s^2}{2m^2}\|z_k^s - p_k^s\|^2 \\
=\, & f(y_k^s) + \frac{\theta_s}{m}\left\langle \widehat{\nabla}_{I_k}(y_k^s),\, x^* - p_k^s\right\rangle + \frac{\theta_s}{m}\left\langle \beta A^T\lambda_k^s,\, x^* - z_k^s\right\rangle \\
& + \left\langle \nabla f(y_k^s) - \widehat{\nabla}_{I_k}(y_k^s),\, \frac{\theta_s}{m}(z_k^s - p_k^s)\right\rangle + \frac{L\theta_s^2}{2m^2}\|z_k^s - p_k^s\|^2 \\
& + \frac{\theta_s^2}{2\eta}(\|x^* - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 - \|x^* - z_k^s\|^2 - \|z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2),
\end{aligned} \tag{37}$$

where the equality $\overset{\text{a}}{=}$ holds due to the update rule of λ_k^s and the optimality condition with respect to z, i.e.,

$$\begin{aligned}
& \widehat{\nabla}_{I_k}(y_k^s) + \beta A^T\left(Az_k^s - w_k^s + \lambda_{k-1}^s\right) + \frac{m\theta_s}{\eta}Q_s(z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})) \\
=\, & \widehat{\nabla}_{I_k}(y_k^s) + \beta A^T\lambda_k^s + \frac{m\theta_s}{\eta}Q_s(z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})) \\
=\, & 0.
\end{aligned} \tag{38}$$

Moreover, the last equality in (37) follows from Property 1. Taking expectation over the random choice of I_k, the inequality (37) can be rewritten as follows:

$$\begin{aligned}
& \mathbb{E}[f(x_k^s)] \\
\leq\, & \mathbb{E}\Bigg[f(y_k^s) + \frac{\theta_s}{m}\left\langle \widehat{\nabla}_{I_k}(y_k^s),\, x^* - p_k^s\right\rangle + \frac{\theta_s}{m}\left\langle \beta A^T\lambda_k^s,\, x^* - z_k^s\right\rangle\Bigg] \\
& + \mathbb{E}\Bigg[\left\langle \nabla f(y_k^s) - \widehat{\nabla}_{I_k}(y_k^s),\, \frac{\theta_s}{m}(z_k^s - p_k^s)\right\rangle + \frac{L\theta_s^2}{2m^2}\|z_k^s - p_k^s\|^2\Bigg] \\
& + \mathbb{E}\left[\frac{\theta_s^2}{2\eta}(\|x^* - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 - \|x^* - z_k^s\|^2 - \|z_k^s - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2)\right].
\end{aligned} \tag{39}$$

Using the inequality (39) and the similar derivation as in Lemmas 3.4 and 6, the following result holds

$$\begin{aligned}
& \mathbb{E}[f(x_k^s) - f(x^*)] \\
\leq\, & \mathbb{E}\Bigg[\left(1 - \frac{\theta_s}{m}\right)\left[f(\overline{x}^{s-1}) - f(x^*)\right]\Bigg] + \mathbb{E}\Bigg[\frac{\theta_s}{m}\left\langle \beta A^T\lambda_k^s,\, x^* - z_k^s\right\rangle\Bigg] \\
& + \mathbb{E}\Bigg[\frac{\theta_s^2}{2\eta}\left(\left\|x^* - z_{k-1}^s\right\|_{Q_s}^2 - \|x^* - p_k^s\|_{Q_s}^2\right)\Bigg] + \mathcal{C}^s.
\end{aligned} \tag{40}$$

Let $\varphi_k^s = \beta\left(\lambda_k^s - \lambda^*\right)$. Using the optimality condition of Problem (2) and the update rule of x_k^s in Algorithm 2, i.e., $\nabla f(x^*) + \beta A^T\lambda^* = 0$, and the similar derivation as in Lemmas 3.4 and 6, the inequality (40) can be rewritten as follows:

$$\begin{aligned}
& \mathbb{E}\Bigg[f(x_k^s) - f(x^*) + \langle \nabla f(x^*),\, x^* - x_k^s)\rangle - \langle A^T\varphi_k^s,\, \frac{\theta_s}{m}(x^* - z_k^s)\rangle\Bigg] \\
\leq\, & \mathbb{E}\Bigg[\left(1 - \frac{\theta_s}{m}\right)\left[f(\overline{x}^{s-1}) - f(x^*) + \langle \nabla f(x^*),\, x^* - \overline{x}^{s-1}\right]\Bigg] + \mathcal{C}^s \\
& + \frac{\theta_s^2}{2\eta}\mathbb{E}\Big[\left\|x^* - z_{k-1}^s\right\|_{Q_s}^2 - \|x^* - z_k^s\|_{Q_s}^2\Big].
\end{aligned} \tag{41}$$

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 $k=1,2,\cdots,m$, and dividing both side of the resulting inequality by m, and using the update rules of $\widetilde{x}^s=\frac{1}{m}\sum_{k=1}^{m}x_k^s$, $f(\widetilde{x}^s)\leq\frac{1}{m}\sum_{k=1}^{m}f(x_k^s)$, and $x_0^s=\widetilde{x}^{s-1}$, we have

$$\begin{aligned}
&\mathbb{E}\left[f(\widetilde{x}^s)-f(x^*)+\langle\nabla f(x^*),\ x^*-\widetilde{x}^s\rangle-\frac{\theta_s}{m^2}\sum_{k=1}^{m}\langle A^T\varphi_k^s,\ x^*-z_k^s\rangle\right]\\
\leq\ &\mathbb{E}\left[\left(1-\frac{\theta_s}{m}\right)\left[f(\overline{x}^{s-1})-f(x^*)+\langle\nabla f(x^*),\ x^*-\overline{x}^{s-1}\right]\right]\\
&+\mathbb{E}\left[\frac{\theta_s^2}{2m\eta}\left(\|x^*-z_0^s\|_{Q_s}^2-\|x^*-z_m^s\|_{Q_s}^2\right)\right]+\mathcal{C}^s.
\end{aligned}$$

This completes the proof. □

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

$$\begin{aligned}
&\mathbb{E}\left[h(\widetilde{w}^s)-h(w^*)+\hat{\nabla}h(w^*)^T(w^*-\widetilde{w}^s)-\frac{\theta_s}{m^2}\sum_{k=1}^{m}\langle\varphi_k^s,w^*-w_k^s\rangle\right]\\
\leq\ &(1-\frac{\theta_s}{m})\left[h(\overline{w}^{s-1})-h(w^*)+\hat{\nabla}h(w^*)^T(w^*-\overline{w}^{s-1})\right]\\
&+\frac{\beta\theta_s}{2m}\mathbb{E}\left[\|Az_0^s-w^*\|^2-\|Az_m^s-w^*\|^2+\sum_{k=1}^{m}\|\lambda_k^s-\lambda_{k-1}^s\|^2\right].
\end{aligned}\tag{42}$$

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

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

Since $h(x)-h(w^*)+\hat{\nabla}h(x^*)^T(w^*-w)\geq 0$, using the update rules in Algorithm 2 and summing up the above inequality for all the iterations $k=1,2,\cdots,m$, 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*

$$\begin{aligned}
&\mathbb{E}[P(\widetilde{x}^s,\widetilde{w}^s)-\langle A\widetilde{x}^s-\widetilde{w}^s,\ \varphi\rangle]\\
\leq\ &(1-\theta_s)\,[P(\widetilde{x}^{s-1},\widetilde{w}^{s-1})-\langle A\widetilde{x}^{s-1}-\widetilde{w}^{s-1},\ \varphi\rangle]\\
&+\frac{\theta_s^2}{2m\eta}\mathbb{E}\Big[\|x^*-z_0^s\|_{Q_s}^2-\|x^*-z_m^s\|_{Q_s}^2\Big]\\
&+\frac{\beta\theta_s}{2m}\mathbb{E}\big[\|Az_0^s-w^*\|^2-\|Az_m^s-w^*\|^2\big]\\
&+\frac{\beta\theta_s}{2m}\mathbb{E}\big[\|\lambda_0^s-\lambda^*-\lambda\|^2-\|\lambda_m^s-\lambda^*-\lambda\|^2\big]\\
&+\frac{\beta\theta_s}{2m}\mathbb{E}\Big[\|\overline{\lambda}^{s-2}-\lambda^*-\lambda\|^2-\|\overline{\lambda}^{s-1}-\lambda^*-\lambda\|^2\Big].
\end{aligned}\tag{43}$$

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

$$\begin{aligned}
&\mathbb{E}\left[P(\widetilde{x}^s, \widetilde{w}^s) - \frac{\theta_s}{m^2}\sum_{k=1}^{m}\langle A^T\varphi_k^s,\ x^* - z_k^s\rangle - \frac{\theta_s}{m^2}\sum_{k=1}^{m}\langle \varphi_k^s,\ w^* - w_k^s\rangle\right]\\
\leq\ &\left(1 - \frac{\theta_s}{m}\right) P(\overline{x}^{s-1}, \overline{w}^{s-1})\\
&+ \frac{\theta_s^2}{2m\eta}\mathbb{E}\Big[\|x^* - z_0^s\|_{Q_s}^2 - \|x^* - z_m^s\|_{Q_s}^2\Big] + \mathcal{C}^s\\
&+ \frac{\beta\theta_s}{2m}\mathbb{E}\left[\|Az_0^s - w^*\|^2 - \|Az_m^s - w^*\|^2 + \sum_{k=1}^{m}\|\lambda_{k-1}^s - \lambda_k^s\|^2\right].
\end{aligned} \tag{44}$$

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

$$\begin{aligned}
&-\frac{\theta_s}{m}\sum_{k=1}^{m}\langle Az_k^s - w_k^s,\ \varphi_k^s - \varphi\rangle\\
=\ &\frac{\beta\theta_s}{2m}\left(\|\lambda_0^s - \lambda^* - \lambda\|^2 - \|\lambda_m^s - \lambda^* - \lambda\|^2\right) - \frac{\beta\theta_s}{2m}\sum_{k=1}^{m}\|\lambda_{k-1}^s - \lambda_k^s\|^2,
\end{aligned} \tag{45}$$

where the second 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

$$\begin{aligned}
&\mathbb{E}[P(\widetilde{x}^s, \widetilde{w}^s) - \langle A\widetilde{x}^s - \widetilde{w}^s, \varphi\rangle]\\
\leq\ &\left(1 - \frac{\theta_s}{m}\right)[P(\overline{x}^{s-1}, \overline{w}^{s-1}) - \langle A\overline{x}^{s-1} - \overline{w}^{s-1}, \varphi\rangle]\\
&+ \frac{\theta_s^2}{2m\eta}\mathbb{E}\Big[\|x^* - r_0^s\|_{Q_s}^2 - \|x^* - r_m^s\|_{Q_s}^2\Big] + \mathcal{C}^s\\
&+ \frac{\beta\theta_s}{2m^2}\mathbb{E}\big[\|Az_0^s - w^*\|^2 - \|Az_m^s - w^*\|^2 + \|\lambda_0^s - \lambda^* - \lambda\|^2 - \|\lambda_m^s - \lambda^* - \lambda\|^2\big].
\end{aligned} \tag{46}$$

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

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

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

This completes the proof. □

Proof of Theorem 4.3 (i.e., the 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 θ_s instead of θ_s^2 in Theorem 4.3. According to the update rule of θ_s, and summing up the inequality (43) for all the stages ($s = 1, 2, \cdots, S$) with $w^* = Ax^*$ and $\overline{\lambda}^{-1} = 0$, we have

According to the update rule of θ_s, and summing up the above inequality for all the stages ($s = 1, 2, \cdots, S$) with $w^* = Ax^*$ and $\overline{\lambda}^{-1} = 0$, we have

$$\begin{aligned}
&\frac{1}{\theta_S}\mathbb{E}\left[P(\widetilde{x}^S, \widetilde{w}^S) - \sum_{s=1}^{S}\langle\varphi,\ A\widetilde{x}^s - \widetilde{w}^s\rangle\right]\\
&\leq \frac{[(1-\theta_1)]}{\theta_1}[P(\widetilde{x}^0, \widetilde{w}^0) - \langle\varphi,\ A\widetilde{x}^0 - \widetilde{w}^0\rangle]\\
&\quad + \frac{\theta_1}{2m\eta}\left\|x^* - \widetilde{x}^0\right\|_{Q_1}^2\\
&\quad + \frac{\beta}{2m}\mathbb{E}\Big[\|A\widetilde{x}^0 - w^*\|^2 + \|\widetilde{\lambda}^0 - \lambda^* - \lambda\|^2\Big]\\
&\quad + \frac{\beta}{2m}\mathbb{E}\big[\|\lambda^* - \lambda\|^2 + \|A\widetilde{x}^0 - w^*\|^2\big].
\end{aligned} \tag{47}$$

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

$$\begin{aligned}
&\mathbb{E}\left[P(\widetilde{x}^S, \widetilde{w}^S) - \langle\varphi,\ \frac{1}{S}\sum_{s=1}^{S}(A\widetilde{x}^s - \widetilde{w}^s)\rangle\right]\\
&\leq \frac{2}{m(S+1)}[P(\widetilde{x}^0, \widetilde{w}^0) - \langle\varphi,\ A\widetilde{x}^0 - \widetilde{w}^0\rangle]\\
&\quad + \frac{1}{m\eta(S+1)}\left\|x^* - \widetilde{x}^0\right\|_{Q_1}^2\\
&\quad + \frac{\beta}{m(S+1)}\left[2\|A^TA\|_2^2\|x^* - \widetilde{x}^0\|^2 + \|\widetilde{\lambda}^0 - \lambda^* - \lambda\|^2 + \|\lambda^* - \lambda\|^2\right].
\end{aligned} \tag{48}$$

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

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

and

$$\|\lambda\|^2 = \|\varphi + \lambda^*\|^2 \leq 2\|\varphi\|^2 + 2\|\lambda^*\|^2 = 2\delta^2 + 2\|\lambda^*\|^2.$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\big[P(\widetilde{x}^S, \widetilde{w}^S) + \delta\|A\widetilde{x}^S - \widetilde{w}^S\|\big]\\
&\leq \frac{2}{m(S+1)}\big[P(\widetilde{x}^0, \widetilde{w}^0) + \delta\|A\widetilde{x}^0 - \widetilde{w}^0\|\big]\\
&\quad + \frac{1}{m\eta(S+1)}\left\|x^* - \widetilde{x}^0\right\|_{Q_1}^2\\
&\quad + \frac{\beta}{m(S+1)}\left[2\|A^TA\|_2^2\|x^* - \widetilde{x}^0\|^2 + \|\widetilde{\lambda}^0 - \lambda^* - \lambda\|^2 + \|\lambda^* - \lambda\|^2\right]\\
&\leq \frac{2}{m(S+1)}\big[P(\widetilde{x}^0, \widetilde{w}^0) + \delta\|A\widetilde{x}^0 - \widetilde{w}^0\|\big]\\
&\quad + \frac{1}{m\eta(S+1)}\left\|x^* - \widetilde{x}^0\right\|_{Q_1}^2\\
&\quad + \frac{\beta}{m(S+1)}\left[2\|A^TA\|_2^2\|x^* - \widetilde{x}^0\|^2 + 2\|\widetilde{\lambda}^0 - \lambda^*\|^2 + 2\|\lambda^*\|^2 + 4\|\lambda\|^2\right].
\end{aligned} \tag{49}$$

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

$$\begin{aligned}
&\mathbb{E}\big[P(\widetilde{x}^S, \widetilde{w}^S) + \delta\|A\widetilde{x}^S - \widetilde{w}^S\|\big] \\
&\leq \mathcal{O}\left(\frac{2[P(\widetilde{x}^0, \widetilde{w}^0) + \delta\|A\widetilde{x}^0 - \widetilde{w}^0\|]}{n(S+1)} + \frac{\left\|x^* - \widetilde{x}^0\right\|^2_{Q_1}}{n\eta(S+1)} + \frac{c_1\beta}{n(S+1)}\right),
\end{aligned} \tag{50}$$

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

Note that the initialization values for $\widetilde{x}^0$, $\widetilde{w}^0$ and $\widetilde{\lambda}^0$ are chosen in our algorithm (i.e., Algorithm 2).

This completes the proof. □

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 results show that the proposed DS-Katyusha algorithm consistently converges much faster than the other accelerated algorithms including Katyusha (Allen-Zhu, 2018) and Varag (Lan et al., 2019). We are optimizing the following binary non-strongly convex problems with $a_i \in \mathbb{R}^d$, $b_i \in \{-1, +1\}$, $i = 1, 2, \ldots n$:

$$\ell_1\text{-norm Regularized Logistic Regression: } \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(-b_i a_i^T x\right)\right) + \lambda\|x\|_1, \tag{51}$$

where λ is the regularization parameter.

For the ℓ_1-norm regularized logistic regression and Lasso problems, 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 η from the set $\{a \times 10^{-k} : a \in \{1, 2, \ldots, 9\}, k \in \mathbb{Z}\}$ for each plot. For Katyusha, DS-Katyusha, DS-Katyusha-ADMM, the parameter τ 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 DS-Katyusha-ADMM algorithm. Moreover, the length of each epoch in SVRG-ADMM and ASVRG-ADMM is set to $m = \lceil 2n/b \rceil$, while the initial epoch length is set as $m_1 = \lceil n/4 \rceil$ for our DS-Katyusha-ADMM algorithm as in (Allen-Zhu & Yuan, 2016). It should be noted that the increasing factor $\frac{\theta_s}{\theta_{s+1}}$ approaches 1 as the number of epochs increases, which means that the epoch length increases very slowly. For ASVRG-ADMM, the parameters η and β are set according to its original paper (Liu et al., 2017).

References

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. *J. Mach. Learn. Res.*, 18:1–51, 2018.

Allen-Zhu, Z. and Yuan, Y. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *ICML*, pp. 1080–1089, 2016.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pp. 315–323, 2013.

Lan, G., Li, Z., and Zhou, Y. A unified variance-reduced accelerated gradient method for convex optimization. In *NeurIPS*, 2019.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liu, Y., Shang, F., and Cheng, J. Accelerated variance reduced stochastic ADMM. In *AAAI*, pp. 2287–2293, 2017.

Zhang, L., Mahdavi, M., and Jin, R. Linear convergence with condition number independent access of full gradients. In *NIPS*, pp. 980–988, 2013.

Zheng, S. and Kwok, J. T. Fast-and-light stochastic ADMM. In *IJCAI*, pp. 2407–2613, 2016.