Supplementary Materials 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”

Anonymous Authors[1]

Preliminaries

In this section, we first give the definition of the proximal operator.

Definition 3 (Proximal Operator)**.** *The proximal operator,* $\text{prox}_h^{\nu}(\cdot)$*, is defined as follows:*

$$\text{prox}_h^{\nu}(y) := \arg\min_x \left\{\frac{1}{2\nu}\|x-y\|^2 + h(x)\right\}. \tag{16}$$

Before giving the convergence analysis of our algorithms, we first give the following properties and lemmas.

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

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

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

$$\begin{aligned}&\mathbb{E}\left[\left\|\widetilde{\nabla} f_{i_k}(x_k^s) - \nabla f(x_k^s)\right\|^2\right]\\ &\le 2L\big[f(\widetilde{x}^{s-1}) - f(x_k^s) + \langle \nabla f(x_k^s),\ x_k^s - \widetilde{x}^{s-1}\rangle\big]\,.\end{aligned} \tag{17}$$

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 3 to the mini-batch setting.

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

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

Theoretical Analysis for DAVIS

In this section, we give some detailed proofs for the convergence analysis of DAVIS (i.e., Algorithm 1), which mainly include the proofs for Lemmas 1 and 2, and Theorem 1 in the main paper.

Now we sketch the proof of Theorem 1 as follows: The proof of Theorem 1 relies on telescoping the upper bound of one-epoch in Lemma 4 below. Lemmas 1 and 2 in the main paper play a key role for obtaining the upper bound of one-epoch in Lemma 4. That is, we first give the upper bound in Lemma 1 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 2 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 2, which can be used to offset by the the residual term $\mathcal{R}$ in Lemma 1. Therefore, we obtain a tight upper bound of one-epoch in Lemma 4 by using Lemmas 1 and 2.

Before giving the detailed proof for Theorem 1, we first analyze the convergence behavior of our DAVIS algorithm in a single iteration.

Proof of Lemma 1 (Upper bound of new snapshot update)

In this subsection, we prove the upper bound for a single iteration of our deterministic gradient descent in Algorithm 1.

Lemma 1 (Upper bound of new snapshot update)**.** *Suppose that Assumption 1 holds. Let $\{\overline{x}^s\}$ be the sequence generated by our deterministic gradient descent step in Algorithm 1, for any $p \in \mathbb{R}^d$, we have*

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

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

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,

$$\nabla f(\widetilde{x}^{s-1}) + \xi + \frac{\theta_s}{2m\eta}(\overline{z}^{s-1} - \widetilde{x}^{s-1}) = 0, \tag{18}$$

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 using the update rule $\overline{x}^{s-1} = \theta_s\overline{z}^{s-1} + (1-\theta_s)\widetilde{x}^{s-1}$, the following inequality holds

$$\begin{aligned}F(\overline{x}^{s-1}) &= h(\overline{x}^{s-1}) + f(\overline{x}^{s-1})\\ &\leq h(\overline{x}^{s-1}) + f(\widetilde{x}^{s-1}) + \left\langle \nabla f(\widetilde{x}^{s-1}),\, \overline{x}^{s-1} - \widetilde{x}^{s-1}\right\rangle + \frac{L}{2}\|\overline{x}^{s-1} - \widetilde{x}^{s-1}\|^2\\ &\leq h(\overline{x}^{s-1}) + f(\widetilde{x}^{s-1}) + \theta_s\left\langle \nabla f(\widetilde{x}^{s-1}),\, x^* - \widetilde{x}^{s-1}\right\rangle + \frac{\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2\\ &- \theta_s\left\langle \nabla f(\widetilde{x}^{s-1}),\, x^* - \overline{z}^{s-1}\right\rangle\\ &= h(\overline{x}^{s-1}) + f(\widetilde{x}^{s-1}) + \theta_s\left\langle \nabla f(\widetilde{x}^{s-1}),\, x^* - \widetilde{x}^{s-1}\right\rangle + \frac{\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2\\ &+ \theta_s\left\langle \xi + \frac{\theta_s}{m\eta}(\overline{z}^{s-1} - \widetilde{x}^{s-1}),\, x^* - \overline{z}^{s-1}\right\rangle\\ &\leq \theta_s F(x^*) + (1-\theta_s)F(\widetilde{x}^{s-1}) + \frac{\theta_s^2}{m\eta}\left(\|x^* - \widetilde{x}^{s-1}\|^2 - \|x^* - \overline{z}^{s-1}\|^2\right)\\ &+ \frac{\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2\end{aligned} \tag{19}$$

where the first inequality holds due to the smoothness of $f(\cdot)$, the third equality holds due to the optimal condition in (18), and Property 1, that is,and the last inequality holds due to the convexities of $h(\cdot)$ and $f(\cdot)$, and the following fact

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

where the first equality follows Property 1. That is,

$$\begin{aligned}&F(\overline{x}^{s-1})-F(x^*)\\ \leq&(1-\theta_s)(F(\widetilde{x}^{s-1})-F(x^*))+\frac{\theta_s^2}{2m\eta}\left(\|x^*-\widetilde{x}^{s-1}\|^2-\|x^*-\overline{z}^{s-1}\|^2\right)+\mathcal{R}^s.\end{aligned}\tag{20}$$

This completes the proof. □

Proof of Lemma 2 (Upper Bound of One-iteration Inner Loop)

In this subsection, we give and prove the following upper bound for our stochastic updates in one iteration (e.g., for a fixed k) of Algorithm 1.

Lemma 2 (Upper Bound of One-iteration)**.** *Suppose that Assumption 1 holds. Let $\{x_k^s, z_k^s\}$ be the sequence generated by our momentum accelerated update rules of Algorithm 1. Then we have*

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

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

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 $g_k^s=\nabla f_{i_k}(y_k^s)-\nabla f_{i_k}(\widetilde{x}^{s-1})+\nabla f(\widetilde{x}^{s-1})$, 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

$$z_k^s\triangleq\arg\min_z\left\{h(z)+\left\langle\widetilde{\nabla}_{i_k}(y_k^s),\ z\right\rangle+\frac{3m\theta_s}{2\eta}\|z-p_k^s+2(\overline{z}^{s-1}-\widetilde{x}^{s-1})\|^2\right\},$$

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

$$\widetilde{\nabla}_{i_k}(y_k^s)+\zeta_k^s+\frac{3m\theta_s}{2\eta}\left[z_k^s-p_k^s+2(\overline{z}^{s-1}-\widetilde{x}^{s-1})\right]=0.\tag{21}$$

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:

$$\begin{aligned}y_k^s&=\frac{\theta_s}{m}p_k^s+\left(1-\frac{\theta_s}{m}\right)\overline{x}^{s-1},\\ x_k^s&=\frac{\theta_s}{m}(z_k^s-p_k^s)+y_k^s\\ &=\frac{\theta_s}{m}z_k^s+\left(1-\frac{\theta_s}{m}\right)\overline{x}^{s-1}.\end{aligned}\tag{22}$$

Below we give the detailed proof of Lemma 2.

Proof of Lemma 2:

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

$$
\begin{aligned}
&\mathbb{E}[F(x_k^s)]\\
&\leq \mathbb{E}\left[h(x_k^s) + f(y_k^s) + \langle \nabla f(y_k^s),\, x_k^s - y_k^s\rangle + \frac{L}{2}\|x_k^s - y_k^s\|^2\right]\\
&\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]\\
&\quad + \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]\\
&\quad + \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]\\
&\quad + \frac{3\theta_s^2}{4\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)\\
&\quad + \mathbb{E}\left[\frac{\theta_s}{m}\left\langle \zeta_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}(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{3\theta_s^2}{4\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]\\
&\quad + \underbrace{\mathbb{E}\left[h(x_k^s) + \frac{\theta_s}{m}\left\langle \zeta_k^s,\, x^* - z_k^s\right\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}\\
&\quad + \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 + \frac{L\theta_s^2}{2m^2}\|z_k^s - p_k^s\|^2\right]}_{A_3},
\end{aligned}
\tag{23}
$$

where the first inequality follows from the smoothness of $f(\cdot)$ (i.e., $f(x_k^s) \leq f(y_k^s) + \langle \nabla f(y_k^s),\, x_k^s - y_k^s\rangle + \frac{L}{2}\|x_k^s - y_k^s\|^2$); the equality $\overset{\text{a}}{=}$ 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 (21) and Property 1, that is,

$$
\begin{aligned}
&\frac{\theta_s}{m}\left\langle -\widetilde{\nabla}_{i_k}(y_k^s),\, x^* - z_k^s\right\rangle\\
&= \frac{\theta_s}{m}\left\langle \frac{3m\theta_s}{2\eta}(z_k^s - p_k^s + 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})),\, x^* - z_k^s\right\rangle\\
&= \frac{3\theta_s^2}{2\eta}\left\langle z_k^s - p_k^s + 2(\overline{z}^{s-1} - \widetilde{x}^{s-1}),\, x^* - z_k^s\right\rangle\\
&= \frac{3\theta_s^2}{4\eta}\left(\|x^* - p_k^s + 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 - \|x^* - z_k^s\|^2 - \|z_k^s - p_k^s + 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2\right),
\end{aligned}
$$

Next we need to bound the terms A_1, A_2, and A_3 in the inequality (23). And we first bound the term A_1. Using the update rule of $x_k^s = \frac{\theta_s}{m} z_k^s + \left(1 - \frac{\theta_s}{m}\right) \overline{x}^{s-1}$, we have

$$
\begin{aligned}
A_1 \; =& \mathbb{E}\left[h(x_k^s) + \frac{\theta_s}{m} \langle \zeta_k^s,\, x^* - z_k^s \rangle\right] \\
=& \, \mathbb{E}\left[h\left(\frac{\theta_s}{m} z_k^s + \left(1 - \frac{\theta_s}{m}\right) \overline{x}^{s-1}\right) + \frac{\theta_s}{m} \langle \zeta_k^s,\, x^* - z_k^s \rangle\right] \\
\leq& \, \mathbb{E}\left[\frac{\theta_s}{m} h(z_k^s) + \left(1 - \frac{\theta_s}{m}\right) h(\overline{x}^{s-1})\right] \\
& \; + \; \mathbb{E}\left[\left\langle \zeta_k^s,\, \frac{\theta_s}{m}(x^* - z_k^s) \right\rangle\right] \\
\leq& \, \mathbb{E}\left[\frac{\theta_s}{m} h(z_k^s) + \left(1 - \frac{\theta_s}{m}\right) h(\overline{x}^{s-1})\right] \\
& \; + \; \mathbb{E}\left[\frac{\theta_s}{m} \left(h(x^*) - h(z_k^s)\right)\right] \\
=& \, \mathbb{E}\left[\left(1 - \frac{\theta_s}{m}\right) h(\overline{x}^{s-1}) + \frac{\theta_s}{m} h(x^*)\right],
\end{aligned} \tag{24}
$$

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 $\widetilde{\nabla}_{i_k}(y_k^s)$ (i.e., $\widetilde{\nabla}_{i_k}(y_k^s) = g_k^s + \frac{m\theta_s}{\eta}\left(\overline{z}^{s-1} - \widetilde{x}^{s-1}\right)$), the term A_2 in the inequality (23) is rewritten as follows:

$$
\begin{aligned}
A_2 =& \, \mathbb{E}\left[\frac{\theta_s}{m} \left\langle \widetilde{\nabla}_{i_k}(y_k^s),\; x^* - p_k^s \right\rangle\right] \\
=& \mathbb{E}\left[\frac{\theta_s}{m} \left\langle \widetilde{\nabla}_{i_k}(y_k^s),\; x^* - p_k^s \right\rangle\right] \\
=& \, \frac{\theta_s}{m} \langle \nabla f(y_k^s),\; x^* - p_k^s \rangle + \frac{\theta_s^2}{\eta} \left\langle \overline{z}^{s-1} - \widetilde{x}^{s-1},\, x^* - p_k^s \right\rangle \\
=& \, \frac{\theta_s}{m} \langle \nabla f(y_k^s),\; x^* - p_k^s \rangle + \frac{\theta_s^2}{2\eta} \left\langle 2(\overline{z}^{s-1} - \widetilde{x}^{s-1}),\, x^* - p_k^s \right\rangle \\
=& \, \frac{\theta_s}{m} \langle \nabla f(y_k^s),\; x^* - p_k^s \rangle + \frac{\theta_s^2}{4\eta} \left(\|x^* - p_k^s + 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 - \|x^* - p_k^s\|^2 - 4\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2\right),
\end{aligned} \tag{25}
$$

where the last equality follows Property 1.

Furthermore, we give the upper bound of the term A_3 in the inequality (23) as follows:

$$\begin{aligned}
A_3 =& \mathbb{E}\Big[\frac{\theta_s}{m}\left\langle \nabla f(y_k^s) - \widetilde{\nabla}_{i_k}(y_k^s), z_k^s - p_k^s \right\rangle\Big] \\
=& \frac{\theta_s}{m}\mathbb{E}\Big[\Big\langle \nabla f(y_k^s) - \widetilde{\nabla}_{i_k}(y_k^s), z_k^s - p_k^s + 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\Big\rangle\Big] \\
& - \frac{\theta_s}{m}\mathbb{E}\Big[\Big\langle \nabla f(y_k^s) - \widetilde{\nabla}_{i_k}(y_k^s), 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\Big\rangle\Big] \\
\overset{\text{a}}{=}& \frac{\theta_s}{m}\mathbb{E}\Big[\Big\langle \nabla f(y_k^s) - \widetilde{\nabla}_{i_k}(y_k^s), z_k^s - p_k^s + 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\Big\rangle\Big] - \frac{2\theta_s^2}{\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2 \qquad (26) \\
\overset{\text{b}}{\leq}& \frac{\eta}{2m^2}\left\|\nabla f(y_k^s) - \widetilde{\nabla}_{i_k}(y_k^s)\right\|^2 + \frac{\theta_s^2}{2\eta}\|z_k^s - p_k^s + 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 - \frac{3\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2 \\
\overset{\text{c}}{\leq}& \frac{1}{m}\left[f(\overline{x}^{s-1}) - f(y_k^s) + \langle \nabla f(y_k^s),\ y_k^s - \overline{x}^{s-1}\rangle\right] \\
& + \frac{\theta_s^2}{2\eta}\|z_k^s - p_k^s + 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 - \frac{3\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2
\end{aligned}$$

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

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

And the equality $\overset{\text{b}}{\leq}$ in (27) follow from the Young's inequality; the $\overset{\text{c}}{\leq}$ in (27) due to the following fact

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

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

Combing the equality (25) and the inequality (26), we have

$$\begin{aligned}
&A_2 + A_3\\
\leq &\left\langle \nabla f(y_k^s), \frac{\theta_s}{m}\left(x^* - p_k^s\right) + \frac{1}{m}(y_k^s - \widetilde{x}^{s-1}) \right\rangle\\
&+ \frac{\theta_s^2}{2\eta}\|z_k^s - p_k^s + 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 - \frac{3\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2\\
&+ \frac{\theta_s^2}{4\eta}\left(\|x^* - p_k^s + 2(\overline{z}^s - \widetilde{x}^s)\|^2 - \|x^* - p_k^s\|^2 - 4\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2\right)\\
\leq &\frac{\theta_s}{m} f(x^*) + \left(1 - \frac{\theta_s}{m}\right) f(\overline{x}^{s-1}) - f(y_k^s)\\
&+ \frac{\theta_s^2}{2\eta}\|z_k^s - p_k^s + 2(\overline{z}^{s-1} - x^{s-1})\|^2 - \frac{5\theta_s^2}{2\eta}\|\overline{z}^{s-1} - x^{s-1}\|^2\\
&+ \frac{\theta_s^2}{4\eta}\left(\|x^* - p_k^s + 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|^2 - \|x^* - p_k^s\|^2\right)
\end{aligned} \tag{27}$$

where the first equality holds due to 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:

$$\begin{aligned}
&\left\langle \nabla f(y_k^s), \frac{\theta_s}{m}\left(x^* - p_k^s\right) + \frac{1}{m}(y_k^s - \overline{x}^{s-1}) \right\rangle\\
= &\left\langle \nabla f(y_k^s), \frac{\theta_s}{m} x^* + \left(1 - \frac{\theta_s}{m} - \frac{1}{m}\right)\overline{x}^{s-1} + \frac{1}{m} y_k^s - y_k^s \right\rangle + \frac{1}{m}\left[f(\overline{x}^{s-1}) - f(y_k^s)\right]\\
\leq &f\left(\frac{\theta_s}{m} x^* + \left(1 - \frac{\theta_s}{m} - \frac{1}{m}\right)\overline{x}^{s-1} + \frac{1}{m} y_k^s\right) - f(y_k^s) + \frac{1}{m}\left[f(\overline{x}^{s-1}) - f(y_k^s)\right]\\
\leq &\frac{\theta_s}{m} f(x^*) + \left(1 - \frac{\theta_s}{m}\right) f(\overline{x}^{s-1}) - f(y_k^s),
\end{aligned} \tag{28}$$

where the first inequality follows the property of f (i.e., $\langle \nabla f(x), y - x \rangle \leq f(y) - f(x)$); and the last inequality holds due to the convexity of $f(\cdot)$.

By the above analysis and combining the inequalities (23), (24) and (28), we have

$$\begin{aligned}
&\mathbb{E}[F(x_k^s) - F(x^*)]\\
\leq &\left(1 - \frac{\theta_s}{m}\right)\left[F(\overline{x}^{s-1}) - F(x^*)\right] - \frac{3\theta_s^2}{2\eta}\|\overline{z}^{s-1} - x^{s-1}\|^2\\
&+ \frac{\theta_s^2}{4\eta}\left(\|x^* - p_k^s + 2(\overline{z}^s - \widetilde{x}^s)\|^2 - \|x^* - p_k^s\|^2 - 4\|\overline{z}^{s-1} - x^{s-1}\|^2\right)\\
&+ \frac{3\theta_s^2}{4\eta}\left(\|x^* - p_k^s + 2(\overline{z}^s - \widetilde{x}^s)\|^2 - \|x^* - z_k^s\|^2\right)
\end{aligned} \tag{29}$$

where the second inequality hold due to jensen inequality (i.e., $\frac{\theta_s^2}{4\eta}\|x^* - p_k^s\|^2 + \frac{3\theta_s^2}{4\eta}\|x^* - z_k^s\|^2 \geq \frac{\theta_s^2}{\eta}\|x^* - p_k^s/4 - 3z_k^s/4\|^2$), $\mathcal{C}^s = -\frac{\theta_s^2}{\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2$; the last inequality holds due to $p_k^s = p_{k-1}^s/4 + 3z_{k-1}^s/4 + 2(\overline{z}^s - \widetilde{x}^s)$ and $r_k^s = p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})$, then $r_{k+1}^s = p_{k+1}^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1}) = p_k^s/4 + 3z_k^s/4$.

This completes the proof. □

Proof of Lemma 4 (Upper Bound of One-epoch):

Before giving the proof of Theorem 1, we first give and prove the following lemma, which provides the upper bound for one epoch of Algorithm 1.

Lemma 4. *(Upper Bound of One-epoch) Suppose that Assumption 1 holds. Let $\{\widetilde{x}^s\}$ be the sequence generated by Algorithm 1. Then we have*

$$\begin{aligned}&\mathbb{E}[F(\widetilde{x}^s)-F(x^*)]\\ &\leq (1-\theta_s)(F(\widetilde{x}^{s-1})-F(x^*))+\frac{\theta_s^2}{m\eta}\left(\|x^*-\widetilde{x}^{s-1}\|^2-\|x^*-\widetilde{x}^s\|^2\right).\end{aligned} \tag{30}$$

Proof. By the one-iteration upper bound in Lemma 2, we have

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

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

$$\begin{aligned}&\mathbb{E}[F(\widetilde{x}^s)-F(x^*)]\\ \leq &\left(1-\frac{\theta_s}{m}\right)\left[F(\overline{x}^{s-1})-F(x^*)\right]+\mathcal{C}^s\\ &+\frac{\theta_s^2}{m\eta}\left(\|x^*-r_1^s\|^2-\left\|x^*-r_{m+1}^s\right\|^2\right)\\ = &\left(1-\frac{\theta_s}{m}\right)\left[F(\overline{x}^{s-1})-F(x^*)\right]+\mathcal{C}^s\\ &+\frac{\theta_s^2}{m\eta}\left(\|x^*-\overline{z}^{s-1}\|^2-\|x^*-\widetilde{x}^s\|^2\right),\end{aligned} \tag{31}$$

where the equality holds due to the update rules $p_0^s=4\overline{z}^{s-1}-3z_0^s$ (i.e, $\overline{z}^{s-1}=r_1^s=p_0^s/4+3z_0^s/3$) and $\widetilde{x}^s=r_{k+1}^s=p_k^s/4+3z_k^s/4$.

Furthermore, by using Lemma 1, we have

$$\begin{aligned}&F(\overline{x}^{s-1})-F(x^*)\\ \leq&(1-\theta_s)(F(\widetilde{x}^{s-1})-F(x^*))+\frac{\theta_s^2}{m\eta}\left(\|x^*-\widetilde{x}^{s-1}\|^2-\|x^*-\overline{z}^{s-1}\|^2\right)+\mathcal{R}^s,\end{aligned} \tag{32}$$

By the above analysis, the upper bound of one-epoch

$$\begin{aligned}&\mathbb{E}[F(\widetilde{x}^s)-F(x^*)]\\ &\leq (1-\theta_s)(F(\widetilde{x}^{s-1})-F(x^*))+\frac{\theta_s^2}{m\eta}\left(\|x^*-\widetilde{x}^{s-1}\|^2-\|x^*-\widetilde{x}^s\|^2\right).\end{aligned} \tag{33}$$

This completes the proof. □

Proof of Theorem 1

In this subsection, we prove the convergence property of DAVIS (i.e., **Algorithm 1**). Theorem 1 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, $\mathcal{O}(1/(nS^2))$.

Theorem 1. *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[1]), then the following result holds*

$$\mathbb{E}[F(\widetilde{x}^s) - F(x^*)] \leq \mathcal{O}\Big(\frac{1}{mS^2}\Big[F(\widetilde{x}^0) - F(x^*) + L\|x^* - \widetilde{x}^0\|^2\Big]\Big).$$

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

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 (34) in Lemma 4 for all $s = 1, 2, \ldots, S$, we have

$$\begin{aligned}&\frac{1}{\theta_S^2}\mathbb{E}\left[F(\widetilde{x}^S) - F(x^*)\right]\\ \leq\ &\frac{(1-\theta_1)}{\theta_1^2}\left[F(\widetilde{x}^0) - F(x^*)\right] + \frac{3\theta_1^2}{4m\eta}\left\|x^* - \widetilde{x}^0\right\|^2\end{aligned}$$

Since $\theta_1 = 1$,

$$\begin{aligned}&\frac{1}{\theta_S^2}\mathbb{E}\big[F(\widetilde{x}^S) - F(x^*)\big]\\ \leq\ &\frac{3}{4m\eta}\left\|x^* - \widetilde{x}^0\right\|^2.\end{aligned} \tag{34}$$

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

$$\begin{aligned}&\mathbb{E}\big[F(\widetilde{x}^S) - F(x^*)\big]\\ \leq\ &\mathcal{O}\left(\frac{\left\|x^* - \widetilde{x}^0\right\|^2}{mS^2\eta}\right).\end{aligned} \tag{35}$$

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. □

Theoretical Analysis for DAVIS-ADMM

In this section, we analyze the convergence property of the proposed algorithm. Similar to Theorem 1, the proofs of Theorems 2 and 3 rely 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 7 and 8 below. Thus, we can obtain the upper bound of one-epoch in Lemma 10 by using Lemmas 6, 7 and 8.

For the more general case, we use the mini-batch version of the proposed compensated stochastic variance reduction gradient estimator in our ASADMM algorithm, which is defined as follows:

Definition 4 (Mini-batch compensated stochastic gradient estimator). *We define a new compensated stochastic variance reduction gradient estimator for our ASADMM algorithm as follows:*

$$\widehat{\nabla}_{I_k}(x) = \underbrace{\nabla f_{I_k}(x) - \nabla f_{I_k}(\widetilde{x}^{s-1}) + \nabla f(\widetilde{x}^{s-1})}_{\text{SVRG estimator}} + \underbrace{\frac{m\theta_s}{\eta}Q_s(\overline{z}^{s-1} - \widetilde{x}^{s-1})}_{\text{Compensated estimator}}, \tag{36}$$

[1] Note that we choose $\widetilde{x}^s$ to be the average point of the previous m stochastic iterates rather than the last iterate because it has been reported to work better in practice (Xiao & Zhang, 2014; Allen-Zhu, 2018; Allen-Zhu & Yuan, 2016).

where $I_k \subset \{1, 2, \ldots, n\}$ is a randomly chosen mini-batch of size b.

Compared with Definition 1, the additional matrix Q_s is need to introduced into the gradient estimator in the ADMM version.

Our ASADMM Algorithm

In this subsection, we present an efficient accelerated stochastic variance reduced ADMM algorithm for solving the structured regularization problem (2), as shown in Algorithm 2. By introducing the dual variable λ and the variable w, the augmented Lagrangian function of Problem (2) is

$$\mathcal{L}(x, w, \lambda) = f(x) + h(w) + \langle \lambda, Ax - w \rangle + \frac{\beta}{2}\|Ax - w\|^2,$$

where $\beta > 0$ is a penalty parameter.

Proofs for ASADMM

Before giving the proof of Theorem 2, we first present the following lemma (Zheng & Kwok, 2016).

Lemma 5. *Let $\varphi_k = \beta(\lambda_k - \lambda^*)$ and 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 new snapshot update)**.** *Suppose that Assumption 1 holds. Let $G_s = \nabla f(\widetilde{x}^{s-1}) + \beta A^T \overline{\lambda}^{s-1}$ for Algorithm 2, and $\{\overline{x}^s, \overline{w}^s, \overline{\lambda}^s\}$ be the sequence generated by our deterministic gradient descent step in Algorithm 2, then we have*

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

where $\mathcal{R}^s = -\frac{\eta}{2m}\|G_s\|^2_{Q_s^{-1}} + \theta_s\langle G_s,\, \widetilde{x}^{s-1} - x^*\rangle$

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

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

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

$$\big(\nabla f(\widetilde{x}^{s-1}) + \xi\big) + \frac{2\theta_s}{m\eta}Q_s(\overline{z}^{s-1} - \widetilde{x}^{s-1}) + \frac{2\beta}{m}A^T(A\overline{z}^{s-1} - \overline{w}^{s-1} + \overline{\lambda}^{s-2}) = 0, \tag{37}$$

where $\xi \in \partial h(\overline{z}^{s-1})$ is a sub-gradient of $h(\cdot)$ at $\overline{z}^{s-1}$. With $\overline{\lambda}^{s-1} = A\overline{z}^{s-1} - \overline{w}^{s-1} + \overline{\lambda}^{s-2}$, we have

$$\big(\nabla f(\widetilde{x}^{s-1}) + \xi\big) + \frac{2\theta_s}{m\eta}Q_s(\overline{z}^{s-1} - \widetilde{x}^{s-1}) + \frac{2\beta}{m}A^T\overline{\lambda}^{s-1} = 0, \tag{38}$$

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

$$\begin{aligned}
&f(\overline{x}^{s-1})\\
&\le f(\widetilde{x}^{s-1}) + \left\langle \nabla f(\widetilde{x}^{s-1}),\, \overline{x}^{s-1} - \widetilde{x}^{s-1} \right\rangle + \frac{L}{2}\|\overline{x}^{s-1} - \widetilde{x}^{s-1}\|^2\\
&\le f(\widetilde{x}^{s-1}) + \theta_s \left\langle \nabla f(\widetilde{x}^{s-1}),\, \overline{z}^{s-1} - \widetilde{x}^{s-1} \right\rangle + \frac{\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2_{Q_s},
\end{aligned} \tag{39}$$

where the first inequality holds due to the smoothness of $f(\cdot)$, and the second inequality uses our choice of $\eta \le \frac{1}{L}$ and the fact that $Q_s \succ I$. Furthermore, for any $p \in \mathbb{R}^d$ and using the optimal condition in (18), we have

$$\begin{aligned}
&f(\overline{x}^{s-1})\\
\le& f(\widetilde{x}^{s-1}) + \frac{\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2_{Q_s}\\
&+\theta_s \left\langle \nabla f(\widetilde{x}^{s-1}),\, \overline{z}^{s-1} - x^* \right\rangle + \theta_s \left\langle \nabla f(\widetilde{x}^{s-1}),\, x^* - \widetilde{x}^{s-1} \right\rangle\\
\le& f(\widetilde{x}^{s-1}) + \frac{\theta_s^2}{2\eta}\|\overline{z}^{s-1} - \widetilde{x}^{s-1}\|^2_{Q_s}\\
&+\left\langle \frac{\theta_s^2}{m\eta} Q_s(\overline{z}^{s-1} - \widetilde{x}^{s-1}) + \frac{\beta\theta_s}{m} A^T \overline{\lambda}^{s-1},\, x^* - \overline{z}^{s-1} \right\rangle + \theta_s \left\langle \nabla f(\widetilde{x}^{s-1}),\, x^* - \widetilde{x}^{s-1} \right\rangle\\
\le& \theta_s f(x^*) + (1-\theta_s) f(\widetilde{x}^{s-1}) + \frac{\theta_s^2}{2m\eta}\left(\|x^* - \widetilde{x}^{s-1}\|^2_{Q_s} - \|x^* - \overline{z}^{s-1}\|^2_{Q_s}\right),
\end{aligned} \tag{40}$$

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

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

$$\begin{aligned}
&\frac{2\beta}{m}\left\langle \overline{w}^{s-2} - \overline{w}^{s-1},\, A(x^* - \overline{z}^{s-1}) \right\rangle\\
=&\frac{\beta}{m}(\|Ax^* - \overline{w}^{s-2}\|^2 - \|Ax^* - \overline{w}^{s-1}\|^2 + \|\overline{\lambda}^{s-1} - \overline{\lambda}^{s-2}\|^2)
\end{aligned}$$

Using $\beta A^T \overline{\lambda}^{s-1} + \nabla f(\widetilde{x}^{s-1}) = \frac{\theta_s}{\eta} Q_s(\widetilde{x}^{s-1} - \overline{z}^{s-1})$, we have

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 result for $\overline{x}^{s-1}$ is given by

$$\begin{aligned}
&\left\langle \beta A^T \overline{\lambda}^{s-1},\, x^* - \overline{x}^{s-1} \right\rangle\\
=& \left\langle \nabla f(x^*),\, \overline{x}^{s-1} - x^* \right\rangle + \left\langle \beta A^T \lambda^*,\, \overline{x}^{s-1} - x^* \right\rangle\\
&+ \left\langle \beta A^T \overline{\lambda}^{s-1},\, x^* - \overline{x}^{s-1} \right\rangle\\
=& \left\langle \nabla f(x^*),\, \overline{x}^{s-1} - x^* \right\rangle + \left\langle A^T \varphi^{s-1},\, x^* - \overline{x}^{s-1} \right\rangle,
\end{aligned} \tag{41}$$

and $\overline{x}^{s-1} = \theta_s \overline{z}^{s-1} + (1 - \overline{z}^{s-1})\widetilde{x}^{s-1}$, we have

$$\begin{aligned}
&f(\overline{x}^{s-1}) - f(x^*) + \left\langle \nabla f(x^*),\, x^* - \overline{x}^{s-1} \right\rangle - \left\langle A^T \varphi^{s-1},\, x^* - \overline{x}^{s-1} \right\rangle\\
\le&(1-\theta_s)(F(\widetilde{x}^{s-1}) - F(x^*) + \left\langle \nabla f(x^*),\, x^* - \widetilde{x}^{s-1} \right\rangle - \left\langle A^T \varphi^{s-1},\, x^* - \widetilde{x}^{s-1} \right\rangle)\\
&+\frac{\theta_s^2}{2\eta}\left(\|p - \widetilde{x}^{s-1}\|^2_{Q_s} - \|p - \overline{z}^{s-1}\|^2_{Q_s}\right)\\
&+\frac{\theta_s^2}{\eta}\left\langle p - x^*, Q_s(\widetilde{x}^{s-1} - \overline{z}^{s-1}) \right\rangle + \theta_s \left\langle \beta A^T \overline{\lambda}^{s-1},\, x^* - \overline{z}^{s-1} \right\rangle.
\end{aligned} \tag{42}$$

Let $p = x^* - v_k^s$, we have

$$\begin{aligned}
&f(\overline{x}^{s-1}) - f(x^*) + \left\langle \nabla f(x^*),\ x^* - \overline{x}^{s-1} \right\rangle - \left\langle A^T \varphi^{s-1},\ x^* - \overline{x}^{s-1} \right\rangle \\
\leq &(1-\theta_s)(F(\widetilde{x}^{s-1}) - F(x^*) + \left\langle \nabla f(x^*),\ x^* - \widetilde{x}^{s-1} \right\rangle - \left\langle A^T \varphi^{s-1},\ x^* - \widetilde{x}^{s-1} \right\rangle) \\
&+ \frac{\theta_s^2}{2\eta}\left(\|x^* - v_k^s - \widetilde{x}^{s-1}\|_{Q_s}^2 - \|x^* - v_k^s - \overline{z}^{s-1}\|_{Q_s}^2\right) \\
&+ \frac{\theta_s^2}{\eta}\left\langle v_k^s, Q_s(\overline{z}^{s-1} - \widetilde{x}^{s-1})\right\rangle + \theta_s \left\langle \beta A^T \overline{\lambda}^{s-1},\ x^* - \overline{z}^{s-1} \right\rangle .
\end{aligned} \tag{43}$$

This completes the proof. □

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

$$\overline{x}^{s-1} = \arg\min_x \left\{ \left\langle m\nabla f(\widetilde{x}^{s-1}), x\right\rangle + \frac{m}{2\eta}\|x - \widetilde{x}^{s-1}\|_{Q_s}^2 + \frac{\beta}{2\theta_s}\|Ax - \overline{w}^{s-2} - q^s + \frac{\theta_s}{m}\overline{\lambda}^{s-2}\|^2 \right\},$$

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

$$\nabla f(\widetilde{x}^{s-1}) + \frac{1}{\eta}Q_s(\overline{x}^{s-1} - \widetilde{x}^{s-1}) + \frac{\beta}{m}A^T[(A\overline{x}^{s-1} - \overline{w}^{s-2} - q^s)/\theta_s + \overline{\lambda}^{s-2}/m] = 0. \tag{44}$$

Let $G_s = \nabla f(\widetilde{x}^{s-1}) + \frac{\beta}{m}A^T[(A\overline{x}^{s-1} - \overline{w}^{s-2} - q^s)/\theta_s + \overline{\lambda}^{s-2}]$ and $\overline{\lambda}^{s-1} = m(A\overline{x}^{s-1} - \overline{w}^{s-1} - q^s)/\theta_s + \overline{\lambda}^{s-2}$, we have

$$\begin{aligned}
G_s &= \nabla f(\widetilde{x}^{s-1}) + \frac{\beta}{m}A^T[(A\overline{x}^{s-1} - \overline{w}^{s-2} - q^s)/\theta_s + \overline{\lambda}^{s-2}/m] \\
&= \nabla f(\widetilde{x}^{s-1}) + \frac{\beta}{m}A^T[(A\overline{x}^{s-1} - \overline{w}^{s-1} - q^s)/\theta_s + \overline{\lambda}^{s-2}/m + \overline{w}^{s-1} - \overline{w}^{s-2}] \\
&= \nabla f(\widetilde{x}^{s-1}) + \frac{\beta}{m^2}A^T\overline{\lambda}^{s-1} + \frac{\beta}{m}A^T(\overline{w}^{s-1} - \overline{w}^{s-2})
\end{aligned} \tag{45}$$

Thus

$$G_s + \frac{1}{\eta}Q_s(\overline{x}^{s-1} - \widetilde{x}^{s-1}) = 0. \tag{46}$$

Using the convexity of $f(\cdot)$, we have

$$f(\widetilde{x}^{s-1}) + \langle \nabla f(\widetilde{x}^{s-1}),\ x - \widetilde{x}^{s-1} \rangle \leq f(x). \tag{47}$$

Since $f(\cdot)$ is L-smooth, the following result holds

$$\begin{aligned}
f(\overline{x}^{s-1}) &\leq f(\widetilde{x}^{s-1}) + \left\langle \nabla f(\widetilde{x}^{s-1}), \overline{x}^{s-1} - \widetilde{x}^{s-1} \right\rangle + \frac{L}{2}\|\overline{x}^{s-1} - \widetilde{x}^{s-1}\|^2 \\
&\leq f(\widetilde{x}^{s-1}) + \left\langle \nabla f(\widetilde{x}^{s-1}), \overline{x}^{s-1} - \widetilde{x}^{s-1} \right\rangle + \frac{1}{2\eta}\|\overline{x}^{s-1} - \widetilde{x}^{s-1}\|_{Q_s}^2,
\end{aligned} \tag{48}$$

where the first inequality holds due to the smoothness of $f(\cdot)$, and the second inequality holds due to the choice $\eta \leq \frac{1}{L}$ and the fact that $Q_s \succ I$.

Using the results in the inequalities (47) and (48), and given any x, the following result holds

$$\begin{aligned}
f(\overline{x}^{s-1}) &\leq f(x) + \langle \nabla f(\widetilde{x}^{s-1}),\ \overline{x}^{s-1} - x \rangle + \frac{1}{2\eta}\|\overline{x}^{s-1} - \widetilde{x}^{s-1}\|_{Q_s}^2 \\
&= f(x) + \langle G_s, \widetilde{x}^{s-1} - x \rangle + \langle G_s,\ \overline{x}^{s-1} - \widetilde{x}^{s-1} \rangle + \frac{1}{2\eta}\|\overline{x}^{s-1} - \widetilde{x}^{s-1}\|_{Q_s}^2 \\
&\quad - \left\langle \frac{\beta}{m^2}A^T\overline{\lambda}^{s-1} + \frac{\beta}{m}(\overline{w}^{s-1} - \overline{w}^{s-2}),\ \overline{x}^{s-1} - x \right\rangle,
\end{aligned} \tag{49}$$

where the equality holds due to the result of (45).

Due to the fact that $\overline{x}^{s-1}-\widetilde{x}^{s-1}=-\eta Q_s^{-1}G_s$ and let $u^{s-1}=\frac{\beta}{m^2}A^T\overline{\lambda}^{s-1}+\frac{\beta}{m}A^T(\overline{w}^{s-1}-\overline{w}^{s-2})$, the result in the inequality (49) can be rewritten as follows:

$$f(\overline{x}^{s-1})\leq f(x)+\langle G_s,\widetilde{x}^{s-1}-x\rangle-\frac{\eta}{2}\|G_s\|^2_{Q_s^{-1}}-\langle u^{s-1},\ \overline{x}^{s-1}-\widetilde{x}^{s-1}\rangle. \tag{50}$$

Let $x=\widetilde{x}^{s-1}$ or $x=x^*$, the above inequality can be reformulated as follows:

$$f(\overline{x}^{s-1})\leq f(\widetilde{x}^{s-1})-\frac{\eta}{2}\|G_s\|^2_{Q_s^{-1}}-\langle u^{s-1},\ \overline{x}^{s-1}-\widetilde{x}^{s-1}\rangle, \tag{51}$$

and

$$f(\overline{x}^{s-1})\leq f(x^*)+\langle G_s,\ \widetilde{x}^{s-1}-x^*\rangle-\frac{\eta}{2}\|G_s\|^2_{Q_s^{-1}}-\langle u^{s-1},\ \overline{x}^{s-1}-x^*\rangle. \tag{52}$$

Multiplying each side of the inequality (51) by $(1-\theta_s)$ and the inequality (52) by θ_s, respectively, and then combining the two resulting inequalities, we obtain

$$\begin{aligned}
f(\overline{x}^{s-1})\leq\ &\theta_s f(x^*)+(1-\theta_s)f(\widetilde{x}^{s-1})-\frac{\eta}{2}\|G_s\|^2_{Q_s^{-1}}+\theta_s\langle G_s,\ \widetilde{x}^{s-1}-x^*\rangle\\
&-\langle u^{s-1},\ \overline{x}^{s-1}-\theta_s x^*-(1-\theta_s)\widetilde{x}^{s-1}\rangle\\
=\ &\theta_s f(x^*)+(1-\theta_s)f(\widetilde{x}^{s-1})-\frac{\eta}{2}\|G_s\|^2_{Q_s^{-1}}+\theta_s\langle G_s,\ \widetilde{x}^{s-1}-x^*\rangle\\
&-\beta\langle A^T\overline{\lambda}^{s-1}/m^2,\ \overline{x}^{s-1}-x^*+(1-\theta_s)(x^*-\widetilde{x}^{s-1})\rangle\\
&-\frac{\beta}{m}\left\langle\overline{w}^{s-1}-\overline{w}^{s-2},\ A\left(\overline{x}^{s-1}-x^*+(1-\theta_s)(x^*-\widetilde{x}^{s-1})\right)\right\rangle\\
=\ &\theta_s f(x^*)+(1-\theta_s)f(\widetilde{x}^{s-1})-\frac{\eta}{2}\|G_s\|^2_{Q_s^{-1}}+\theta_s\langle G_s,\ \widetilde{x}^{s-1}-x^*\rangle\\
&-\beta\langle A^T\overline{\lambda}^{s-1}/m^2,\ \overline{x}^{s-1}-x^*+(1-\theta_s)(x^*-\widetilde{x}^{s-1})\rangle\\
&+\frac{\beta\theta_s}{m}\left(\|Ax^*-\overline{w}^{s-2}\|^2-\|Ax^*-\overline{w}^{s-1}\|^2+\frac{1}{m}\|\overline{\lambda}^{s-1}-\overline{\lambda}^{s-2}\|^2\right),
\end{aligned} \tag{53}$$

where the last equality follows from the similar derivation as in Lemma 3 in (Zheng & Kwok, 2016) with the update rule $\overline{\lambda}^{s-1}=m(A\overline{x}^{s-1}-\overline{w}^{s-1}-q^s)/\theta_s+\overline{\lambda}^{s-2}$.

Subtracting $f(x^*)+\beta\langle A^T\overline{\lambda}^{s-1}/m^2,\ x^*-\overline{x}^{s-1}\rangle$ from both sides of the inequality (53), we have

$$\begin{aligned}
&f(\overline{x}^{s-1})-f(x^*)-\beta\langle A^T\overline{\lambda}^{s-1}/m^2,\ x^*-\overline{x}^{s-1}\rangle\\
\leq\ &(1-\theta_s)\left[f(\widetilde{x}^{s-1})-f(x^*)-\beta\langle A^T\overline{\lambda}^{s-1}/m^2,\ x^*-\widetilde{x}^{s-1}\rangle\right]-\frac{\eta}{2}\|G_s\|^2_{Q_s^{-1}}+\theta_s\langle G_s,\ \widetilde{x}^{s-1}-x^*\rangle\\
&+\frac{\beta\theta_s}{m}\left(\|Ax^*-\overline{w}^{s-2}\|^2-\|Ax^*-\overline{w}^{s-1}\|^2+\frac{1}{m}\|\overline{\lambda}^{s-1}-\overline{\lambda}^{s-2}\|^2\right).
\end{aligned} \tag{54}$$

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

$$\begin{aligned}
&\left\langle\beta A^T\overline{\lambda}^{s-1}/m^2,\ x^*-\overline{x}^{s-1}\right\rangle\\
=\ &\left\langle\nabla f(x^*),\ \overline{x}^{s-1}-x^*\right\rangle+\left\langle\beta A^T\lambda^*,\ \overline{x}^{s-1}-x^*\right\rangle\\
&+\left\langle\beta A^T\overline{\lambda}^{s-1}/m^2,\ x^*-\overline{x}^{s-1}\right\rangle\\
=\ &\left\langle\nabla f(x^*),\ \overline{x}^{s-1}-x^*\right\rangle+\left\langle A^T\varphi^{s-1},\ x^*-\overline{x}^{s-1}\right\rangle,
\end{aligned} \tag{55}$$

and the result for $\widetilde{x}^{s-1}$ is

$$\begin{aligned}
&\left\langle \beta A^T \overline{\lambda}^{s-1}/m^2,\ x^* - \widetilde{x}^{s-1} \right\rangle \\
&= \left\langle \nabla f(x^*),\ \widetilde{x}^{s-1} - x^* \right\rangle + \left\langle \beta A^T \lambda^*,\ \widetilde{x}^{s-1} - x^* \right\rangle \\
&\quad + \left\langle \beta A^T \overline{\lambda}^{s-1}/m^2,\ x^* - \widetilde{x}^{s-1} \right\rangle \\
&= \left\langle \nabla f(x^*),\ \widetilde{x}^{s-1} - x^* \right\rangle + \left\langle A^T \varphi^{s-1},\ x^* - \widetilde{x}^{s-1} \right\rangle.
\end{aligned} \tag{56}$$

Using (54), (55) and (56), we have

$$\begin{aligned}
&f(\overline{x}^{s-1}) - f(x^*) + \left\langle \nabla f(x^*),\ x^* - \overline{x}^{s-1} \right\rangle - \left\langle A^T \varphi^{s-1},\ x^* - \overline{x}^{s-1} \right\rangle \\
&\leq (1-\theta_s) \left[f(\widetilde{x}^{s-1}) - f(x^*) + \left\langle \nabla f(x^*),\ x^* - \widetilde{x}^{s-1} \right\rangle - \left\langle A^T \varphi^{s-1},\ x^* - \widetilde{x}^{s-1} \right\rangle \right] \\
&\quad - \frac{\eta}{2} \|G_s\|^2_{Q_s^{-1}} + \theta_s \langle G_s,\ \widetilde{x}^{s-1} - x^* \rangle \\
&\quad + \frac{\beta\theta_s}{m} \left(\|Ax^* - \overline{w}^{s-2}\|^2 - \|Ax^* - \overline{w}^{s-1}\|^2 + \frac{1}{m} \|\overline{\lambda}^{s-1} - \overline{\lambda}^{s-2}\|^2 \right).
\end{aligned} \tag{57}$$

Thus, we have

$$\begin{aligned}
&f(\overline{x}^{s-1}) - f(x^*) + \left\langle \nabla f(x^*),\ x^* - \overline{x}^{s-1} \right\rangle - \left\langle A^T \varphi^{s-1},\ x^* - \overline{x}^{s-1} - (1-\theta_s)(x^* - \widetilde{x}^{s-1}) \right\rangle \\
&\leq (1-\theta_s) [f(\widetilde{x}^{s-1}) - f(x^*) + \left\langle \nabla f(x^*),\ x^* - \widetilde{x}^{s-1} \right\rangle] + \mathcal{R}^s \\
&\quad + \frac{\beta\theta_s}{m} \left(\|Ax^* - \overline{w}^{s-2}\|^2 - \|Ax^* - \overline{w}^{s-1}\|^2 + \frac{1}{m} \|\overline{\lambda}^{s-1} - \overline{\lambda}^{s-2}\|^2 \right),
\end{aligned} \tag{58}$$

where $\mathcal{R}^s = -\frac{\eta}{2} \|G_s\|^2_{Q_s^{-1}} + \theta_s \langle G_s,\ \widetilde{x}^{s-1} - x^* \rangle$

Using the convexity of $h(\cdot)$ and $\zeta \in \partial h(\overline{w}^{s-1})$, we have

$$h(\overline{w}^{s-1}) - h(w) \leq \langle -\zeta,\ w - \overline{w}^{s-1} \rangle.$$

Let $w = \widetilde{w}^{s-1}$ or $w = w^*$, the above inequality can be reformulated as follows:

$$h(\overline{w}^{s-1}) - h(\widetilde{w}^{s-1}) \leq\ \langle -\zeta,\ \widetilde{w}^{s-1} - \overline{w}^{s-1} \rangle, \tag{59}$$

and

$$h(\overline{w}^{s-1}) - h(w^*) \leq\ \langle -\zeta,\ w^* - \overline{w}^{s-1} \rangle. \tag{60}$$

Multiplying each side of the inequality (59) by $(1-\theta_s)$ and the inequality (60) by θ_s, and then combining the two resulting inequalities, we obtain

$$\begin{aligned}
&h(\overline{w}^{s-1}) - h(w^*) \\
&\leq (1-\theta_s) \left(h(\widetilde{w}^{s-1}) - h(w^*) \right) + \langle -\zeta,\ \theta_s w^* + (1-\theta_s) \widetilde{w}^{s-1} - \overline{w}^{s-1} \rangle \\
&= (1-\theta_s) \left(h(\widetilde{w}^{s-1}) - h(w^*) \right) + \beta \langle \overline{\lambda}^{s-1}/m^2, \theta_s w^* + (1-\theta_s) \widetilde{w}^{s-1} - \overline{w}^{s-1} \rangle,
\end{aligned} \tag{61}$$

where the last equality holds due to the optimal condition with $\overline{\lambda}^{s-1} = m(A\overline{x}^{s-1} - \overline{w}^{s-1} - q^s)/\theta_s + \overline{\lambda}^{s-2}$

$$\zeta + \beta \left((A\overline{x}^{s-1} - \overline{w}^{s-1} - q^s)/\theta_s + \overline{\lambda}^{s-2}/m \right)/m = \zeta + \beta \overline{\lambda}^{s-1}/m^2 = 0,$$

where ζ is a subgradient of $h(\widetilde{w}^{s-1})$.

Furthermore, using the optimal condition (i.e., $\hat{\nabla} h(w^*) + \beta\lambda^* = 0$) and the result in (61), we have

$$\begin{aligned} & h(\overline{w}^{s-1}) - h(y^*) + \langle \hat{\nabla} h(w^*),\ w^* - \overline{w}^{s-1} \rangle - \left\langle \varphi^{s-1},\ w^* - \overline{w}^{s-1} \right\rangle \\ \leq\ & (1-\theta_s)[h(\widetilde{w}^{s-1}) - h(w^*) + \left\langle \hat{\nabla} h(w^*),\ w^* - \widetilde{w}^{s-1} \right\rangle - \left\langle \varphi^{s-1},\ w^* - \widetilde{w}^{s-1} \right\rangle]. \end{aligned} \tag{62}$$

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

$$\begin{aligned} & \langle A^T \varphi^{s-1},\, x^* - \overline{x}^{s-1} \rangle + \langle \varphi^{s-1},\, w^* - \overline{w}^{s-1} \rangle + \langle A\overline{x}^{s-1} - \overline{w}^{s-1},\, \varphi^{s-1} - \varphi \rangle \\ =\ & -\langle A\overline{x}^{s-1} - \overline{w}^{s-1},\, \varphi \rangle, \end{aligned} \tag{63}$$

and

$$\begin{aligned} & \langle A^T \varphi^{s-1},\, x^* - \widetilde{x}^{s-1} \rangle + \langle \varphi^{s-1},\, w^* - \widetilde{w}^{s-1} \rangle + \langle A\widetilde{x}^{s-1} - \widetilde{w}^{s-1},\, \varphi^{s-1} - \varphi \rangle \\ =\ & -\langle A\widetilde{x}^{s-1} - \widetilde{w}^{s-1},\, \varphi \rangle. \end{aligned} \tag{64}$$

Multiplying each side of the inequality (64) by $-(1-\theta_s)$ and combining the inequality (63), we have

$$\begin{aligned} & \langle A^T \varphi^{s-1},\, x^* - \overline{x}^{s-1} - (1-\theta_s)(x^* - \widetilde{x}^{s-1}) \rangle + \langle \varphi^{s-1},\, w^* - \overline{w}^{s-1} - (1-\theta_s)(w^* - \widetilde{w}^{s-1}) \rangle \\ & + \langle (A\overline{x}^{s-1} + \overline{w}^{s-1}) - (1-\theta_s)(A\widetilde{x}^{s-1} + \widetilde{w}^{s-1}) - \theta_s c,\, \varphi^{s-1} - \varphi \rangle \\ =\ & -\langle A\overline{x}^{s-1} + \overline{w}^{s-1} - (1-\theta_s)(A\widetilde{x}^{s-1} + \widetilde{w}^{s-1}) - \theta_s c,\, \varphi \rangle. \end{aligned} \tag{65}$$

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

$$\begin{aligned} & -\langle A\overline{x}^{s-1} - \overline{w}^{s-1} - (1-\theta_s)(A\widetilde{x}^{s-1} - \widetilde{w}^{s-1}),\, \varphi^{s-1} - \varphi \rangle \\ =\ & -\frac{\beta\theta_s}{m} \langle \overline{\lambda}^{s-1} - \overline{\lambda}^{s-2},\, \overline{\lambda}^{s-1} - \lambda^* - \lambda \rangle \\ =\ & \frac{\beta\theta_s}{2m} \left(\|\overline{\lambda}^{s-2} - \lambda^* - \lambda\|^2 - \|\overline{\lambda}^{s-1} - \lambda^* - \lambda\|^2 - \|\overline{\lambda}^{s-2} - \overline{\lambda}^{s-1}\|^2 \right), \end{aligned} \tag{66}$$

where the last equality holds due to Property 1.

Using the results in (64), (68) and (72), the definition of $P(x,y)$ (i.e., $P(x,y) = f(x) - f(x^*) - \nabla f(x^*)^T (x - x^*) + h(y) - h(y^*) - \hat{\nabla} h(y^*)^T (y - y^*)$) and the update rules in Algorithm 2, we have

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

This completes the proof. □

Upper bound of our stochastic gradient descent step:

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

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

Lemma 8. *Suppose that Assumption 1 holds. Let $\{\widetilde{x}^s, \widetilde{w}^s, \widetilde{\lambda}^s\}$ be sequence generated by Algorithm 2, then we have*

$$\begin{aligned}
&\mathbb{E}\Bigg[f(\widetilde{x}^s) - f(x^*) + \langle \nabla f(x^*),\ x^* - \widetilde{x}^s\rangle - \frac{\theta_s}{m^2}\sum_{k=1}^{m}\langle A^T\varphi_k^s,\ x^* - x^{s-1} + v_k^s - z_k^s\rangle\Bigg]\\
&\le \mathbb{E}\Bigg[\left(1-\frac{2}{m}\right)\left[f(\overline{x}^{s-1}) - f(x^*) + \langle \nabla f(x^*),\ x^* - x_{k-1}^s\rangle\right]\Bigg]\\
&\quad + \frac{2-\theta_s}{m}[f(\widetilde{x}^{s-1}) - f(x^*) + \langle \nabla f(x^*),\ x^* - \widetilde{x}^{s-1}\rangle]\\
&\quad + \mathbb{E}\Bigg[\frac{(m+1)\theta_s^2}{m^2\eta}\left(\|x^* - p_0^s\|_{Q_s}^2 - \|x^* - p_m^s\|_{Q_s}^2\right)\Bigg] + \mathcal{C}^s,
\end{aligned} \tag{67}$$

where $\mathcal{C}^s = (1-\frac{2}{m})\theta_s\langle G_s,\ x^* - \widetilde{x}^{s-1}\rangle + \frac{\eta(1-\frac{2}{m})^2}{2}\|G_s\|^2$.

Proof. We give the upper bound for our stochastic gradient descent step in this lemma. Using the similar derivation as in Lemma 2, we have the following result for our stochastic gradient descent step.

Let $g_k^s = \nabla f_{I_k}(y_k^s) - \nabla f_{I_k}(\widetilde{x}^{s-1}) + \nabla f(\widetilde{x}^{s-1})$ and $\overline{G}_s = (1-\frac{2}{m})G_s$, we have $\widehat{\nabla}_{I_k}(y_k^s) = g_k^s + (m-2)G_s = g_k^s + m\overline{G}_s$. Since the function $f(\cdot)$ is L-smooth, and by using the update rule of $x_k^s = y_k^s + \frac{\theta_s}{m}(z_k^s - 2v_k^s)$, we have

$$\begin{aligned}
f(x_k^s) &\le f(y_k^s) + \left\langle \nabla f(y_k^s),\ \frac{\theta_s}{m}(z_k^s - 2v_k^s)\right\rangle + \frac{L\theta_s^2}{2m^2}\|z_k^s - 2v_k^s\|^2\\
&= f(y_k^s) + \frac{\theta_s}{m}\left\langle \widehat{\nabla}_{I_k}(y_k^s) + 2\theta_s v_k^s/\eta,\ x^* - \widetilde{x}^{s-1} - v_k^s\right\rangle - \frac{\theta_s}{m}\left\langle \widehat{\nabla}_{I_k}(y_k^s) + 2\theta_s v_k^s/\eta,\ x^* - \widetilde{x}^{s-1} + v_k^s - z_k^s\right\rangle\\
&\quad + \left\langle \nabla f(y_k^s) - \widehat{\nabla}_{I_k}(y_k^s) - 2\theta_s v_k^s/\eta,\ \frac{\theta_s}{m}(z_k^s - 2v_k^s)\right\rangle + \frac{L\theta_s^2}{2m^2}\|z_k^s - 2v_k^s\|^2\\
&\overset{\text{a}}{=} f(y_k^s) + \frac{\theta_s}{m}\left\langle \widehat{\nabla}_{I_k}(y_k^s) + 2\theta_s v_k^s/\eta,\ x^* - \widetilde{x}^{s-1} - v_k^s\right\rangle\\
&\quad + \frac{\theta_s}{m}\left\langle \frac{(m+2)\theta_s}{m\eta}Q_s(z_k^s - \varsigma v_k^s) + \beta A^T\lambda_k^s,\ x^* - \widetilde{x}^{s-1} + v_k^s - z_k^s\right\rangle\\
&\quad + \left\langle \nabla f(y_k^s) - \widehat{\nabla}_{I_k}(y_k^s) - 2\theta_s v_k^s/\eta,\ \frac{\theta_s}{m}(z_k^s - 2v_k^s)\right\rangle + \frac{L\theta_s^2}{2m^2}\|z_k^s - 2v_k^s\|^2\\
&\overset{\text{b}}{\le} f(y_k^s) + \frac{\theta_s}{m}\left\langle \widehat{\nabla}_{I_k}(y_k^s) + 2\theta_s v_k^s/\eta,\ x^* - \widetilde{x}^{s-1} - v_k^s\right\rangle + \frac{\theta_s}{m}\left\langle \beta A^T\lambda_k^s,\ x^* - \widetilde{x}^{s-1} + v_k^s - z_k^s\right\rangle\\
&\quad + \left\langle \nabla f(y_k^s) - \widehat{\nabla}_{I_k}(y_k^s) - 2\theta_s v_k^s/\eta,\ \frac{\theta_s}{m}(z_k^s - 2v_k^s)\right\rangle + \frac{L\theta_s^2}{2m^2}\|z_k^s - v_k^s\|_{Q_s}^2 .\\
&\quad + \frac{(m+2)\theta_s^2}{2m\eta}(\|x^* - \widetilde{x}^{s-1} - (\varsigma-1)v_k^s\|^2 - \|x^* - \widetilde{x}^{s-1} + v_k^s - z_k^s\|^2 - \|z_k^s - \varsigma v_k^s\|^2).
\end{aligned} \tag{68}$$

where the equality $\overset{\text{a}}{=}$ holds due to the update rule of $\lambda_k^s = A(z_k^s - v_k^s - \widetilde{x}^{s-1}) - w_k^s + \lambda_{k-1}^s$ and the optimality condition of (15), i.e.,

$$\begin{aligned}
&\widehat{\nabla}_{I_k}(y_k^s) + \beta A^T\left(A(z_k^s - v_k^s - \widetilde{x}^{s-1}) - w_k^s + \lambda_{k-1}^s\right) + \frac{(m+2)\theta_s}{m\eta}Q_s(z_k^s + \tau v_k^s)\\
&= \widehat{\nabla}_{I_k}(y_k^s) + \beta A^T\lambda_k^s + \frac{(m+2)\theta_s}{m\eta}Q_s(z_k^s + \tau v_k^s)\\
&= \widehat{\nabla}_{I_k}(y_k^s) + 2\theta_s v_k^s/\eta + \beta A^T\lambda_k^s + \frac{(m+2)\theta_s}{m\eta}Q_s(z_k^s - \varsigma v_k^s)\\
&= 0.
\end{aligned} \tag{69}$$

Moreover, the inequality $\overset{\text{b}}{\leq}$ in (68) follows from Property 1 and the similar derivation in (35) and (36).

Taking expectation over the random choice of I_k, the inequality (68) can be rewritten as follows:

$$\begin{aligned}
&\mathbb{E}[f(x_k^s)]\\
\leq\, &\mathbb{E}\Big[f(y_k^s) + \theta_s\left\langle\widehat{\nabla}_{I_k}(y_k^s) + 2\theta_s v_k^s/\eta,\, x^* - \widetilde{x}^{s-1} - v_k^s\right\rangle + \frac{\theta_s}{m}\left\langle\beta A^T\lambda_k^s,\, x^* - \widetilde{x}^{s-1} + v_k^s - z_k^s\right\rangle\Big]\\
&+\mathbb{E}\Big[\left\langle\nabla f(y_k^s) - \widehat{\nabla}_{I_k}(y_k^s) - 2\theta_s v_k^s/\eta,\, \frac{\theta_s}{m}(z_k^s - 2v_k^s)\right\rangle + \frac{L\theta_s^2}{2m^2}\|z_k^s - 2v_k^s\|_{Q_s}^2\Big]\\
&+\mathbb{E}\left[\frac{(m+2)\theta_s^2}{2m\eta}(\|x^* - \widetilde{x}^{s-1} - (\varsigma - 1)v_k^s\|^2 - \|x^* - \widetilde{x}^{s-1} + v_k^s - z_k^s\|^2 - \|z_k^s - \varsigma v_k^s\|^2)\right].
\end{aligned}\tag{70}$$

Using the variance upper bound in Lemma 5 and the similar derivation in (39) of Lemma 2, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\langle\nabla f(y_k^s) - \widehat{\nabla}_{I_k}(y_k^s) - 2\theta_s v_k^s/\eta,\, \frac{\theta_s}{m}(z_k^s - 2v_k^s)\right\rangle + \frac{L\theta_s^2}{2m^2}\|z_k^s - 2v_k^s\|^2\right]\\
\leq\, &\frac{(n-b)}{mb(n-1)}\left[f(\widetilde{x}^{s-1}) - f(y_k^s) + \left\langle\nabla f(y_k^s), y_k^s - \widetilde{x}^{s-1}\right\rangle\right] + \theta_s\left\langle\overline{G}_s, v_k^s\right\rangle + \frac{\eta}{2}\|\overline{G}_s\|_{Q_s^{-1}}^2 + \frac{2m\theta_s^2}{(2m-1)m\eta}\|v_k^s\|^2\\
&+\mathbb{E}\left[\frac{\theta_s^2}{m\eta}\|z_k^s - 2v_k^s\|^2 + \frac{\theta_s^2}{2\eta}\|z_k^s - v_k^s\|^2\right],,
\end{aligned}\tag{71}$$

where the first inequality holds due to the Young's inequality.

Furthermore, we also have

$$\begin{aligned}
&\frac{\theta_s}{m}\mathbb{E}\Big[\left\langle\widehat{\nabla}_{I_k}(y_k^s) + 2\theta_s v_k^s/\eta,\, x^* - \widetilde{x}^{s-1} - v_k^s\right\rangle\Big]\\
=\, &\frac{\theta_s}{m}\left\langle\nabla f(y_k^s), x^* - \widetilde{x}^{s-1} - v_k^s\right\rangle + \frac{\theta_s}{m}\left\langle m\overline{G}_s, x^* - \widetilde{x}^{s-1} - v_k^s\right\rangle + \frac{2\theta_s^2}{m\eta}\left\langle v_k^s,\, x^* - \widetilde{x}^{s-1} - v_k^s\right\rangle\\
=\, &\frac{\theta_s}{m}\left\langle\nabla f(y_k^s), x^* - \widetilde{x}^{s-1} - v_k^s\right\rangle + \theta_s\left\langle\overline{G}_s, x^* - \widetilde{x}^{s-1} - v_k^s\right\rangle\\
&+\frac{\theta_s^2}{2\eta}\left(\|x^* - \widetilde{x}^{s-1} - (\varsigma - 1)v_k^s\|^2 - \|x^* - \widetilde{x}^{s-1} - (\varsigma - 1)v_k^s - \frac{2}{m}v_k^s\|^2\right)\\
&-\left(\frac{2(2 - \varsigma - 2/m)\theta_s^2}{m} + \frac{2\theta_s^2}{m^2}\right)\|v_k^s\|^2.
\end{aligned}\tag{72}$$

Using the inequalities (76), (77), (78) and the similar derivation as in Lemma 2, the following result holds

$$\begin{aligned}
&\mathbb{E}[f(x_k^s) - f(x^*)]\\
\leq\, &\mathbb{E}\Big[\left(1 - \frac{2}{m}\right)\left[f(\overline{x}^{s-1}) - f(x^*)\right] + \frac{2 - \theta_s}{m}\left[f(\widetilde{x}^{s-1}) - f(x^*)\right]\Big]\\
&+\mathbb{E}\Big[\frac{\theta_s}{m}\left\langle\beta A^T\lambda_k^s,\, x^* - \widetilde{x}^{s-1} + v_k^s - z_k^s\right\rangle\Big]\\
&+\mathbb{E}\Big[\frac{(m+1)\theta_s^2}{m\eta}\left(\left\|x^* - p_{k-1}^s\right\|_{Q_s}^2 - \|x^* - p_k^s\|_{Q_s}^2\right)\Big] + \mathcal{C}^s,
\end{aligned}\tag{73}$$

where $\mathcal{C}^s = \theta_s\langle\overline{G}_s, x^* - \widetilde{x}^{s-1}\rangle + \frac{\eta}{2}\|\overline{G}_s\|_{Q_s^{-1}}^2 = (1 - \frac{2}{m})\theta_s\langle G_s, x^* - \widetilde{x}^{s-1}\rangle + \frac{\eta(1-\frac{2}{m})^2}{2}\|G_s\|_{Q_s^{-1}}^2$ and $\overline{G}_s = (1 - \frac{2}{m})G_s$.

Let $\varphi_k^s = \beta\left(\lambda_k^s - \lambda^*\right)$. Using the optimality condition of Problem (2), i.e., $\nabla f(x^*) + \beta A^T\lambda^* = 0$, we have

$$\begin{aligned}
&\frac{\theta_s}{m}\langle\beta A^T\lambda_k^s,\, x^* - \widetilde{x}^{s-1} + v_k^s - z_k^s\rangle\\
=\, &-\frac{\theta_s}{m}\langle\nabla f(x^*),\, x^* - \widetilde{x}^{s-1} + v_k^s - z_k^s\rangle + \frac{\theta_s}{m}\left\langle A^T\varphi_k^s,\, x^* - \widetilde{x}^{s-1} + v_k^s - z_k^s\right\rangle.
\end{aligned}$$

By the above analysis, we have

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

Using the update rule of $x_k^s=\frac{\theta_s}{m}(z_k^s-v_k^s+\widetilde{x}^{s-1})+\left(1-\frac{2}{m}\right)\overline{x}^{s-1}+\frac{2-\theta_s}{m}\widetilde{x}^{s-1}$ and adding both sides of the above inequality by $\left(1-\frac{2}{m}\right)\langle\nabla f(x^*),\ x^*-\overline{x}^{s-1}\rangle+\frac{2-\theta_s}{m}\langle\nabla f(x^*),\ x^*-\widetilde{x}^{s-1}\rangle$, we have

$$\begin{aligned}&\mathbb{E}\bigg[f(x_k^s)-f(x^*)+\langle\nabla f(x^*),\ x^*-x_k^s\rangle-\frac{\theta_s}{m}\langle A^T\varphi_k^s,\ x^*-x^{s-1}+v_k^s-z_k^s\rangle\bigg]\\ \leq\ &\mathbb{E}\bigg[\left(1-\frac{2}{m}\right)\left[f(\overline{x}^{s-1})-f(x^*)+\langle\nabla f(x^*),\ x^*-\overline{x}^{s-1}\rangle\right]\bigg]\\ &+\frac{2-\theta_s}{m}[f(\widetilde{x}^{s-1})-f(x^*)+\langle\nabla f(x^*),\ x^*-\widetilde{x}^{s-1}\rangle]\\ &+\mathbb{E}\bigg[\frac{(m+1)\theta_s^2}{m\eta}\left(\left\|x^*-p_{k-1}^s\right\|_{Q_s}^2-\|x^*-p_k^s\|_{Q_s}^2\right)\bigg]+\mathcal{C}^s.\end{aligned}\tag{74}$$

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 (80) 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

$$\begin{aligned}&\mathbb{E}\bigg[f(\widetilde{x}^s)-f(x^*)+\langle\nabla f(x^*),\ x^*-\widetilde{x}^s\rangle-\frac{\theta_s}{m^2}\sum_{k=1}^m\langle A^T\varphi_k^s,\ x^*-x^{s-1}+v_k^s-z_k^s\rangle\bigg]\\ \leq\ &\mathbb{E}\bigg[\left(1-\frac{2}{m}\right)\left[f(\overline{x}^{s-1})-f(x^*)+\langle\nabla f(x^*),\ x^*-\overline{x}^{s-1}\rangle\right]\bigg]\\ &+\frac{2-\theta_s}{m}[f(\widetilde{x}^{s-1})-f(x^*)+\langle\nabla f(x^*),\ x^*-\widetilde{x}^{s-1}\rangle]\\ &+\mathbb{E}\bigg[\frac{(m+1)\theta_s^2}{m^2\eta}\left(\|x^*-p_0^s\|_{Q_s}^2-\|x^*-p_m^s\|_{Q_s}^2\right)\bigg]+\mathcal{C}^s.\end{aligned}$$

This completes the proof. □

Lemma 9. *Using the same notation as in Lemma 8, we have*

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

Proof. Using the convexity of $h(\cdot)$ and $\zeta_k^s \in \partial h(w_k^s)$, we have

$$\begin{aligned}
&h(w_k^s) - h(w^*)\\
\le\ &\langle -\zeta_k^s,\ w^* - w_k^s\rangle\\
=\ &\left\langle A(z_{k-1}^s - v_k^s + \widetilde{x}^{s-1}) - w_k^s + \lambda_{k-1}^s,\ w^* - w_k^s \right\rangle,
\end{aligned}$$

where the equality holds due to the optimal condition of Problem (14), that is, $\zeta_k^s + \beta(A(z_{k-1}^s - v_k^s + \widetilde{x}^{s-1}) - w_k^s + \lambda_{k-1}^s) = 0$. Using the similar derivation as in Lemma 3 in (Zheng & Kwok, 2016), we obtain

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

Since $h(x) - h(w^*) + \hat{\nabla} h(x^*)^T(w^* - w) \ge 0$, 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 rule of $\widetilde{w}^s = \frac{\theta_s}{m^2}\sum_{k=1}^m w_k^s + (1 - \frac{2}{m})\overline{w}^{s-1} + \frac{2-\theta_s}{m}\widetilde{w}^{s-1}$, we have

$$\begin{aligned}
&\mathbb{E}\left[h(\widetilde{w}^s) - h(w^*) + \hat{\nabla} h(w^*)^T(w^* - \widetilde{w}^s) - \frac{\theta_s}{m^2}\sum_{k=1}^m \langle \varphi_k^s, w^* - w_k^s\rangle\right]\\
\le\ &(1 - \frac{2}{m})\left[h(\overline{w}^{s-1}) - h(w^*) + \hat{\nabla} h(w^*)^T(w^* - \overline{w}^{s-1})\right]\\
&+ \frac{2-\theta_s}{m}\left[h(\widetilde{w}^{s-1}) - h(w^*) + \hat{\nabla} h(w^*)^T(w^* - \widetilde{w}^{s-1})\right]\\
&+ \frac{\beta\theta_s}{2m^2}\mathbb{E}\left[\|Az_0^s - w^*\|^2 - \|Az_m^s - w^*\|^2 + \sum_{k=1}^m \|\lambda_k^s - \lambda_{k-1}^s\|^2\right].
\end{aligned} \tag{75}$$

This completes the proof. □

Lemma 10. *Using the same notation as in Lemma 8, we have*

$$\begin{aligned}
&\mathbb{E}[P(\widetilde{x}^s, \widetilde{w}^s) - \langle A\widetilde{x}^s - \widetilde{w}^s,\ \varphi\rangle]\\
\le\ &\left(1 - \theta_s + \frac{2-\theta_s}{m}\right)[P(\widetilde{x}^{s-1}, \widetilde{w}^{s-1}) - \langle A\widetilde{x}^{s-1} - \widetilde{w}^{s-1},\ \varphi\rangle]\\
&+ \frac{(m+1)\theta_s^2}{m^2\eta}\mathbb{E}\Big[\|x^* - p_0^s\|_{Q_s}^2 - \|x^* - p_m^s\|_{Q_s}^2\Big]\\
&+ \frac{\beta\theta_s}{2m^2}\mathbb{E}\big[\|Az_0^s - w^*\|^2 - \|Az_m^s - w^*\|^2\big]\\
&+ \frac{\beta\theta_s}{2m^2}\mathbb{E}\big[\|\lambda_0^s - \lambda^* - \lambda\|^2 - \|\lambda_m^s - \lambda^* - \lambda\|^2\big]\\
&+ \frac{\beta\theta_s}{2m}\mathbb{E}\Big[\|\overline{\lambda}^{s-2} - \lambda^* - \lambda\|^2 - \|\overline{\lambda}^{s-1} - \lambda^* - \lambda\|^2\Big].
\end{aligned} \tag{76}$$

Proof. Using the definition of $P(x, y)$ and combining the inequality (67) in Lemma 8 and the inequality (81) in Lemma 9, we have

$$\begin{aligned}
&\mathbb{E}\Bigg[P(\widetilde{x}^s, \widetilde{w}^s) - \frac{\theta_s}{m^2}\sum_{k=1}^{m}\langle A^T\varphi_k^s,\ x^* - x^{s-1} + v_k^s - z_k^s\rangle - \frac{\theta_s}{m^2}\sum_{k=1}^{m}\langle \varphi_k^s,\ w^* - w_k^s\rangle\Bigg] \\
\le\ & \left(1 - \frac{2}{m}\right)P(\overline{x}^{s-1}, \overline{w}^{s-1}) + \frac{2-\theta_s}{m}P(\widetilde{x}^{s-1}, \widetilde{w}^{s-1}) \\
&+ \frac{(m+1)\theta_s^2}{m^2\eta}\mathbb{E}\Big[\|x^* - p_0^s\|_{Q_s}^2 - \|x^* - p_m^s\|_{Q_s}^2\Big] + \mathcal{C}^s \\
&+ \frac{\beta\theta_s}{2m^2}\mathbb{E}\Bigg[\|Az_0^s - w^*\|^2 - \|Az_m^s - w^*\|^2 + \sum_{k=1}^{m}\|\lambda_{k-1}^s - \lambda_k^s\|^2\Bigg].
\end{aligned} \tag{77}$$

For any $\varphi = \beta\lambda$, we have

$$\begin{aligned}
&\frac{\theta_s}{m^2}\sum_{k=1}^{m}\left(\langle A^T\varphi_k^s,\, x^* - \widetilde{x}^{s-1} + v_k^s - z_k^s\rangle + \langle\varphi_k^s,\, w^* - w_k^s\rangle\right) \\
&+ \frac{\theta_s}{m^2}\sum_{k=1}^{m}\left(\langle A(z_k^s - v_k^s + \widetilde{x}^{s-1}) - w_k^s,\, \varphi_k^s - \varphi\rangle\right) \\
=\ & -\frac{\theta_s}{m^2}\sum_{k=1}^{m}\langle A(z_k^s - v_k^s + \widetilde{x}^{s-1}) - w_k^s,\, \varphi\rangle,
\end{aligned} \tag{78}$$

where $\varphi_k^s = \beta\left(\lambda_k^s - \lambda^*\right)$, and $Ax^* - w^* = 0$.

Using Lemma 9 with $\lambda_k^s = \lambda_{k-1}^s + A(z_k^s - v_k^s + \widetilde{x}^{s-1}) - w_k^s$, $\varphi_k^s = \beta(\lambda_k^s - \lambda^*)$ and $\varphi = \beta\lambda$, we have

$$\begin{aligned}
&-\frac{\theta_s}{m^2}\sum_{k=1}^{m}\langle A(z_k^s - v_k^s + \widetilde{x}^{s-1}) - w_k^s,\, \varphi_k^s - \varphi\rangle \\
=\ & -\frac{\beta\theta_s}{m^2}\sum_{k=1}^{m}\langle\lambda_k^s - \lambda_{k-1}^s,\, \lambda_k^s - \lambda^* - \lambda\rangle \\
=\ & \frac{\beta\theta_s}{2m^2}\sum_{k=1}^{m}\left(\|\lambda_{k-1}^s - \lambda^* - \lambda\|^2 - \|\lambda_k^s - \lambda^* - \lambda\|^2 - \|\lambda_{k-1}^s - \lambda_k^s\|^2\right) \\
=\ & \frac{\beta\theta_s}{2m^2}\left(\|\lambda_0^s - \lambda^* - \lambda\|^2 - \|\lambda_m^s - \lambda^* - \lambda\|^2\right) - \frac{\beta\theta_s}{2m^2}\sum_{k=1}^{m}\|\lambda_{k-1}^s - \lambda_k^s\|^2,
\end{aligned} \tag{79}$$

where the second equality holds due to Property 1.

Adding both sides of the inequality (83) by $-\frac{\theta_s}{m^2}\sum_{k=1}^{m}\langle A(z_k^s - v_k^s + \widetilde{x}^{s-1}) - w_k^s,\, \varphi_k^s - \varphi\rangle$ and using the facts in (84) and (85), we have

$$\begin{aligned}
&\mathbb{E}\Bigg[P(\widetilde{x}^s, \widetilde{w}^s) - \frac{\theta_s}{m^2}\sum_{k=1}^{m}\langle A(z_k^s - v_k^s + \widetilde{x}^{s-1}) - w_k^s,\, \varphi\rangle\Bigg] \\
\le\ & \left(1 - \frac{2}{m}\right)P(\overline{x}^{s-1}, \overline{w}^{s-1}) + \frac{2-\theta_s}{m}P(\widetilde{x}^{s-1}, \widetilde{w}^{s-1}) \\
&+ \frac{(m+1)\theta_s^2}{m^2\eta}\mathbb{E}\Big[\|x^* - p_0^s\|_{Q_s}^2 - \|x^* - p_m^s\|_{Q_s}^2\Big] + \mathcal{C}^s \\
&+ \frac{\beta\theta_s}{2m^2}\mathbb{E}\big[\|Az_0^s - w^*\|^2 - \|Az_m^s - w^*\|^2 + \|\lambda_0^s - \lambda^* - \lambda\|^2 - \|\lambda_m^s - \lambda^* - \lambda\|^2\big].
\end{aligned} \tag{80}$$

With the update rule of $x_k^s = \frac{\theta_s}{m}(z_k^s - v_k^s - \widetilde{x}^{s-1}) + (1 - \frac{2}{m})\overline{x}^{s-1} + \frac{2-\theta_s}{m}\widetilde{x}^{s-1}$ $w_k^s = \frac{\theta_s}{m}w_k^s + (1 - \frac{2}{m})\overline{w}^{s-1} + \frac{2-\theta_s}{m}\widetilde{w}^{s-1}$, $\widetilde{x}^s = \frac{1}{m}\sum_{k=1}^{m} x_k^s$, $\widetilde{w}^s = \frac{1}{m}\sum_{k=1}^{m} w_k^s$, and adding both sides of the inequality (86) by $(1 - \frac{2}{m})\langle A\overline{x}^{s-1} - \overline{w}^{s-1},\, \varphi\rangle +$

$\frac{2-\theta_s}{m}\langle A\widetilde{x}^{s-1} - \widetilde{w}^{s-1},\, \varphi\rangle$, we have

$$\begin{aligned}
&\mathbb{E}[P(\widetilde{x}^s, \widetilde{w}^s) - \langle A\widetilde{x}^s - \widetilde{w}^s, \varphi\rangle] \\
\leq\; &\left(1 - \frac{2}{m}\right)[P(\overline{x}^{s-1}, \overline{w}^{s-1}) - \langle A\overline{x}^{s-1} - \overline{w}^{s-1}, \varphi\rangle] \\
&+ \frac{2-\theta_s}{m}[P(\widetilde{x}^{s-1}, \widetilde{w}^{s-1}) - \langle A\widetilde{x}^{s-1} - \widetilde{w}^{s-1}, \varphi\rangle] \\
&+ \frac{(m+2)\theta_s^2}{m^2\eta}\mathbb{E}\Big[\|x^* - p_0^s\|_{Q_s}^2 - \|x^* - p_m^s\|_{Q_s}^2\Big] + \mathcal{C}^s \\
&+ \frac{\beta\theta_s}{2m^2}\mathbb{E}\big[\|Az_0^s - w^*\|^2 - \|Az_m^s - w^*\|^2 + \|\lambda_0^s - \lambda^* - \lambda\|^2 - \|\lambda_m^s - \lambda^* - \lambda\|^2\big]\,.
\end{aligned} \tag{81}$$

Furthermore, by using Lemma 4 for the upper bound of our new snapshot point, we have

$$\begin{aligned}
&(1 - \frac{2}{m})\mathbb{E}\big[P(\overline{x}^{s-1}, \overline{w}^{s-1}) - \langle A\overline{x}^{s-1} - \overline{w}^{s-1}, \varphi\rangle\big] \\
\leq\; &(1 - \frac{2}{m})(1-\theta_s)[P(\widetilde{x}^{s-1}, \widetilde{w}^{s-1}) - \langle A\widetilde{x}^{s-1} - \widetilde{w}^{s-1}, \varphi\rangle] \\
&- (1 - \frac{2}{m})\frac{\eta}{2}\|G_s\|_{Q_s^{-1}}^2 + (1 - \frac{2}{m})\theta_s\langle G_s,\, \widetilde{x}^{s-1} - x^*\rangle \\
&+ \frac{(1-\frac{2}{m})\beta\theta_s}{2m}\mathbb{E}\big[\|A\overline{x}^{s-2} - w^*\|^2 - \|A\overline{x}^{s-1} - w^*\|^2\big] \\
&+ \frac{(1-\frac{2}{m})\beta\theta_s}{2m}\mathbb{E}\Big[\|\overline{\lambda}^{s-2} - \lambda^* - \lambda\|^2 - \|\overline{\lambda}^{s-1} - \lambda^* - \lambda\|^2\Big] \\
&+ \frac{(1-\frac{2}{m})\beta\theta_s}{2m}\mathbb{E}\big[\|Ax^* - \overline{w}^{s-2}\|^2 - \|Ax^* - \overline{w}^{s-1}\|^2\big]\,.
\end{aligned} \tag{82}$$

By adding up the above two inequalities, we have

$$\begin{aligned}
&\mathbb{E}[P(\widetilde{x}^s, \widetilde{w}^s) - \langle A\widetilde{x}^s - \widetilde{w}^s,\ \varphi\rangle] \\
\leq\; &\left[(1 - \frac{2}{m})(1-\theta_s) + \frac{2-\theta_s}{m}\right][P(\widetilde{x}^{s-1}, \widetilde{w}^{s-1}) - \langle A\widetilde{x}^{s-1} - \widetilde{w}^{s-1},\ \varphi\rangle] \\
&+ \frac{\theta_s^2}{2\eta m}\mathbb{E}\Big[\|x^* - p_0^s\|_{Q_s}^2 - \|x^* - p_m^s\|_{Q_s}^2\Big] \\
&+ \frac{\beta\theta_s}{2m^2}\mathbb{E}\big[\|Az_0^s - w^*\|^2 - \|Az_m^s - w^*\|^2\big] \\
&+ \frac{\beta\theta_s}{2m^2}\mathbb{E}\big[\|\lambda_0^s - \lambda^* - \lambda\|^2 - \|\lambda_m^s - \lambda^* - \lambda\|^2\big] \\
&+ \frac{(1-\frac{2}{m})\beta\theta_s}{2m}\mathbb{E}\Big[\|\overline{\lambda}^{s-2} - \lambda^* - \lambda\|^2 - \|\overline{\lambda}^{s-1} - \lambda^* - \lambda\|^2\Big] \\
&+ \frac{(1-\frac{2}{m})\beta\theta_s}{2m}\mathbb{E}\big[\|Ax^* - \overline{w}^{s-2}\|^2 - \|Ax^* - \overline{w}^{s-1}\|^2\big]\,.
\end{aligned} \tag{83}$$

This completes the proof. □

Proof of Theorem 2:

Proof. Using the upper bound of the s-th epoch in Lemma 9 and dividing both sides of the inequality (89) by θ_s instead of θ_s^2 in Theorem 2, we have

$$\begin{aligned}
&\frac{1}{\theta_s}\mathbb{E}[P(\widetilde{x}^s, \widetilde{w}^s) - \langle \varphi,\ A\widetilde{x}^s - \widetilde{w}^s\rangle] \\
\leq\ & \frac{\left[(1-\frac{2}{m})(1-\theta_s) + \frac{2-\theta_s}{m}\right]}{\theta_s}[P(\widetilde{x}^{s-1}, \widetilde{w}^{s-1}) - \langle A\widetilde{x}^{s-1} - \widetilde{w}^{s-1},\ \varphi\rangle] \\
&+ \frac{\theta_s}{\eta m}\mathbb{E}\Big[\|x^* - p_0^s\|_{Q_s}^2 - \|x^* - p_m^s\|_{Q_s}^2\Big] \\
&+ \frac{\beta}{2m^2}\mathbb{E}\big[\|Az_0^s - w^*\|^2 - \|Az_m^s - w^*\|^2 + \|\lambda_0^s - \lambda^* - \lambda\|^2 - \|\lambda_m^s - \lambda^* - \lambda\|^2\big] \\
&+ \frac{(1-\frac{2}{m})\beta}{2m}\mathbb{E}\Big[\|\overline{\lambda}^{s-2} - \lambda^* - \lambda\|^2 - \|\overline{\lambda}^{s-1} - \lambda^* - \lambda\|^2 + \|Ax^* - \overline{w}^{s-2}\|^2 - \|Ax^* - \overline{w}^{s-1}\|^2\Big] .
\end{aligned} \tag{84}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\theta_S}P(\widetilde{x}^S, \widetilde{w}^S) - \sum_{s=1}^{S}\sigma_s\langle \varphi,\ A\widetilde{x}^s - \widetilde{w}^s\rangle\right] \\
\leq\ & \frac{\left[(1-\frac{2}{m})(1-\theta_1) + \frac{2-\theta_1}{m}\right]}{\theta_1}[P(\widetilde{x}^0, \widetilde{w}^0) - \langle \varphi,\ A\widetilde{x}^0 - \widetilde{w}^0\rangle] \\
&+ \frac{\theta_1}{2m\eta}\big\|x^* - \widetilde{x}^0\big\|_{Q_1}^2 \\
&+ \frac{\beta}{2m^2}\mathbb{E}\Big[\|A\widetilde{x}^0 - w^*\|^2 + \|\widetilde{\lambda}^0 - \lambda^* - \lambda\|^2\Big] \\
&+ \frac{\beta}{2m}\mathbb{E}\big[\|\lambda^* - \lambda\|^2 + \|A\widetilde{x}^0 - w^*\|^2\big] ,
\end{aligned} \tag{85}$$

where $\sigma_s = \frac{1}{\theta_{s-1}} - \frac{(1-\frac{2}{m})(1-\theta_s)+\frac{2-\theta_s}{m}}{\theta_s}$, which implies that $0 < \sigma_s < 1$.

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

$$\begin{aligned}
&\mathbb{E}\left[P(\widetilde{x}^S, \widetilde{w}^S) - \langle \varphi,\ \frac{1}{S}\sum_{s=1}^{S}\sigma_s(A\widetilde{x}^s - \widetilde{w}^s)\rangle\right] \\
\leq\ & \frac{2}{m(S+1)}[P(\widetilde{x}^0, \widetilde{w}^0) - \langle \varphi,\ A\widetilde{x}^0 - \widetilde{w}^0\rangle] \\
&+ \frac{1}{m\eta(S+1)}\big\|x^* - \widetilde{x}^0\big\|_{Q_1}^2 \\
&+ \frac{\beta}{m(S+1)}\Big[2\|A^TA\|_2^2\|x^* - \widetilde{x}^0\|^2 + \|\widetilde{\lambda}^0 - \lambda^* - \lambda\|^2 + \|\lambda^* - \lambda\|^2\Big] .
\end{aligned} \tag{86}$$

Let $\widehat{x} = \frac{1}{S}\sum_{s=1}^{S}\sigma_s\widetilde{x}^s$ 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 inequality holds:

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

and

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

Therefore, we have

$$
\begin{aligned}
&\mathbb{E}\big[P(\widetilde{x}^S, \widetilde{w}^S) + \delta\|A\widetilde{x}^S - \widetilde{w}^S\|\big] \\
\le\ & \frac{2}{m(S+1)}[P(\widetilde{x}^0, \widetilde{w}^0) + \delta\|A\widetilde{x}^0 - \widetilde{w}^0\|] \\
&+ \frac{1}{m\eta(S+1)}\big\|x^* - \widetilde{x}^0\big\|_{Q_1}^2 \\
&+ \frac{\beta}{m(S+1)}\Big[2\|A^TA\|_2^2\|x^* - \widetilde{x}^0\|^2 + \|\widetilde{\lambda}^0 - \lambda^* - \lambda\|^2 + \|\lambda^* - \lambda\|^2\Big] \\
\le\ & \frac{2}{m(S+1)}[P(\widetilde{x}^0, \widetilde{w}^0) + \delta\|A\widetilde{x}^0 - \widetilde{w}^0\|] \\
&+ \frac{1}{m\eta(S+1)}\big\|x^* - \widetilde{x}^0\big\|_{Q_1}^2 \\
&+ \frac{\beta}{m(S+1)}\Big[2\|A^TA\|_2^2\|x^* - \widetilde{x}^0\|^2 + 2\|\widetilde{\lambda}^0 - \lambda^*\|^2 + 2\|\lambda^*\|^2 + 4\|\lambda\|^2\Big].
\end{aligned}
\tag{87}
$$

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

$$
\begin{aligned}
&\mathbb{E}\big[P(\widetilde{x}^S, \widetilde{w}^S) + \delta\|A\widetilde{x}^S - \widetilde{w}^S\|\big] \\
&\le \mathcal{O}\left(\frac{2[P(\widetilde{x}^0, \widetilde{w}^0) + \delta\|A\widetilde{x}^0 - \widetilde{w}^0\|]}{n(S+1)} + \frac{\big\|x^* - \widetilde{x}^0\big\|_{Q_1}^2}{n\eta(S+1)} + \frac{c_1\beta}{n(S+1)}\right),
\end{aligned}
\tag{88}
$$

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

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. □

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.

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.

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

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.