Supplementary Materials for "ASVRG: Accelerated Proximal SVRG"

Fanhua Shang FHSHANG@XIDIAN.EDU.CN
Licheng Jiao LCHJIAO@MAIL.XIDIAN.EDU.CN
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, China

Kaiwen Zhou KWZHOU@CSE.CUHK.EDU.HK
James Cheng JCHENG@CSE.CUHK.EDU.HK
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong

Yan Ren CRANE_ROCK@OUTLOOK.COM
Yufei Jin JESTY@JESTYF.CN
School of Computer Science and Technology, Xidian University, China

Editors: Jun Zhu and Ichiro Takeuchi

In this supplementary material, we give the detailed proofs for some lemmas, theorems and properties.

Appendix A.

Appendix A1: Proof of Proposition 1

Proof Using Theorem 1, we have

$$\rho(\omega) = 1 - \omega + \frac{\omega^2}{\mu m \eta}.$$

Obviously, it is desirable to have a small convergence factor $\rho(\omega)$. So, we minimize $\rho(\omega)$ with given η. Then we have

$$\omega_\star = m\mu\eta/2 \leq 1 - \frac{\widetilde{L}\eta}{1 - \widetilde{L}\eta},$$

and

$$\rho(\omega_\star) = 1 - \frac{m\mu\eta}{4} > 0.$$

The above two inequalities imply that

$$\eta \leq \frac{1 + 4c_1 - \sqrt{1 + 16c_1^2}}{2\widetilde{L}} = \frac{1 + 4c_1 - \sqrt{1 + 16c_1^2}}{2c_1 m\mu} \quad \text{and} \quad \eta < \frac{4}{m\mu},$$

where $c_1 = \widetilde{L}/(m\mu) > 0$. This completes the proof. ■

Appendix A2: ASVRG Pseudo-Codes

We first give the details on Algorithm 1 with $\omega = 1$ for optimizing smooth objective functions such as ℓ_2-norm regularized logistic regression, as shown in Algorithm 3, which is almost identical to the regularized SVRG in (Babanezhad et al., 2015) and the original SVRG in (Johnson and Zhang, 2013). The main differences between Algorithm 3 and the latter two are the initialization of x_0^s and the choice of the snapshot point $\widetilde{x}^s$. Moreover, we can use the doubling-epoch technique in (Mahdavi et al., 2013; Allen-Zhu and Yuan, 2016) to further speed up our ASVRG method for both SC and non-SC cases. Besides, all the proposed algorithms can be extended to the mini-batch setting as in (Nitanda, 2014; Konečný et al., 2016). In particular, our ASVRG method can be extended to an accelerated incremental aggregated gradient method with the SAGA estimator in (Defazio et al., 2014).

Algorithm 3 ASVRG with $\omega = 1$

Input: The number of epochs S, the number of iterations m per epoch, and the step size η.
Initialize: $x_0^1 = \widetilde{x}^0$, $m_1 = n/4$, $\rho > 1$, and the probability $P = [p_1, \ldots, p_n]$.

```
for s = 1, 2, ..., S do
    ∇̃ = (1/n) Σ_{i=1}^{n} ∇f_i(x̃^{s-1});
    for t = 1, 2, ..., m_s do
        Pick i_t from {1, ..., n} randomly based on P;
        ∇̃f_{i_t}(x_{t-1}^s) = [∇f_{i_t}(x_{t-1}^s) − ∇f_{i_t}(x̃^{s-1})] / (n p_{i_t}) + ∇̃;
        x_t^s = x_{t-1}^s − η [∇̃f_{i_t}(x_{t-1}^s) + ∇g(x_{t-1}^s)];
    end for
    x̃^s = (1/m_s) Σ_{t=1}^{m_s} x_t^s,  x_0^{s+1} = x_{m_s}^s,  m_{s+1} = min(⌊ρ m_s⌋, m);
end for
```

Output: $\widetilde{x}^S$.

Appendix A3: Elastic-Net Regularized Logistic Regression

In this paper, we mainly focus on the following elastic-net regularized logistic regression problem for binary classification,

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-b_i a_i^T x)) + \frac{\lambda_1}{2} \|x\|^2 + \lambda_2 \|x\|_1,$$

where $\{(a_i, b_i)\}$ is a set of training examples, and $\lambda_1, \lambda_2 \geq 0$ are the regularization parameters. Note that $f_i(x) = \log(1 + \exp(-b_i a_i^T x)) + (\lambda_1/2)\|x\|^2$.

In this paper, we used the two publicly available data sets in the experiments: Covtype and RCV1, as listed in Table 1. For fair comparison, we implemented the state-of-the-art stochastic methods such as SAGA (Defazio et al., 2014), SVRG (Johnson and Zhang, 2013), Acc-Prox-SVRG (Nitanda, 2014), Catalyst (Lin et al., 2015), and Katyusha (Allen-Zhu, 2018), and our ASVRG method in C++ with a Matlab interface, and conducted all the experiments on a PC with an Intel i5-4570 CPU and 16GB RAM.

Table 1: Summary of data sets used for our experiments.

Data sets	Covtype	RCV1
Number of training samples, n	581,012	20,242
Number of dimensions, d	54	47,236
Sparsity	22.12%	0.16%
Size	50M	13M

Appendix B. Proof of Lemma 2

Before proving the key Lemma 2, we first give the following lemma and properties, which are useful for the convergence analysis of our ASVRG method.

Lemma 1 *Suppose Assumption 1 holds. Then the following inequality holds*

$$\mathbb{E}\left[\left\|\widetilde{\nabla} f_{i_t}(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2\right] \leq 2\widetilde{L}\big(f(\widetilde{x}^{s-1}) - f(x_{t-1}^s) + \big\langle \nabla f(x_{t-1}^s),\ x_{t-1}^s - \widetilde{x}^{s-1}\big\rangle\big), \quad (10)$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product (i.e., $\langle x, y\rangle = x^T y$ for all $x, y \in \mathbb{R}^d$), and $\widetilde{L} = \max_j L_j/(p_j n)$. When $p_i = 1/n$ (i.e., uniform random sampling), $\widetilde{L} = L_{\max} := \max_j L_j$, while $\widetilde{L} = L_{\text{avg}} := \frac{1}{n}\sum_{j=1}^n L_j$ when $p_i = L_i/\sum_{j=1}^n L_j$ (i.e., the sampling probabilities p_i for $i \in \{1, \ldots, n\}$ are proportional to their Lipschitz constants L_i of $\nabla f_i(\cdot)$).

The proof of Lemma 1 is similar to that of Lemma 3.4 in (Allen-Zhu, 2018). For the sake of completeness, we give the detailed proof of Lemma 1 as follows. Their main difference is that Lemma 1 provides the upper bound on the expected variance of the modified stochastic gradient estimator, i.e.,

$$\widetilde{\nabla} f_{i_t}(x_{t-1}^s) = \big[\nabla f_{i_t}(x_{t-1}^s) - \nabla f_{i_t}(\widetilde{x}^{s-1})\big]/(np_{i_t}) + \widetilde{\nabla},$$

while the upper bound in Lemma 3.4 in (Allen-Zhu, 2018) is for the standard stochastic gradient estimator in (Johnson and Zhang, 2013; Zhang et al., 2013). Obviously, the upper bound in Lemma 1 is much tighter than that in (Johnson and Zhang, 2013; Xiao and Zhang, 2014; Allen-Zhu and Yuan, 2016), e.g., Corollary 3.5 in (Xiao and Zhang, 2014) and Lemma A.2 in (Allen-Zhu and Yuan, 2016).

Proof Now we take expectations with respect to the random choice of i_t, to obtain

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{np_{i_t}}\big[\nabla f_{i_t}(x_{t-1}^s) - \nabla f_{i_t}(\widetilde{x}^{s-1})\big]\right] \\
&= \sum_{i=1}^n \frac{p_i}{np_i}\big[\nabla f_i(x_{t-1}^s) - \nabla f_i(\widetilde{x}^{s-1})\big] \\
&= \sum_{i=1}^n \frac{1}{n}\big[\nabla f_i(x_{t-1}^s) - \nabla f_i(\widetilde{x}^{s-1})\big] \\
&= \nabla f(x_{t-1}^s) - \nabla f(\widetilde{x}^{s-1}).
\end{aligned} \quad (11)$$

Theorem 2.1.5 in (Nesterov, 2004) immediately implies the following result.

$$\left\|\nabla f_i(x_{t-1}^s) - \nabla f_i(\widetilde{x}^{s-1})\right\|^2 \le 2L_i\left[f_i(\widetilde{x}^{s-1}) - f_i(x_{t-1}^s) + \langle \nabla f_i(x_{t-1}^s),\ x_{t-1}^s - \widetilde{x}^{s-1}\rangle\right].$$

Dividing both sides of the above inequality by $1/(n^2 p_i)$, and summing it over $i = 1, \ldots, n$, we obtain

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\frac{1}{np_i}\left\|\nabla f_i(x_{t-1}^s) - \nabla f_i(\widetilde{x}^{s-1})\right\|^2 \\ &\le 2\widetilde{L}\left[f(\widetilde{x}^{s-1}) - f(x_{t-1}^s) + \langle \nabla f(x_{t-1}^s),\ x_{t-1}^s - \widetilde{x}^{s-1}\rangle\right]. \end{aligned} \tag{12}$$

Using the definition of $\widetilde{\nabla} f_{i_t}(x_{t-1}^s) = [\nabla f_{i_t}(x_{t-1}^s) - \nabla f_{i_t}(\widetilde{x}^{s-1})]/(np_{i_t}) + \nabla f(\widetilde{x}^{s-1})$, (11), and (12), we have

$$\begin{aligned} &\mathbb{E}\left[\left\|\widetilde{\nabla} f_{i_t}(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2\right] \\ &= \mathbb{E}\left[\left\|\nabla f(\widetilde{x}^{s-1}) - \nabla f(x_{t-1}^s) - \frac{\nabla f_{i_t}(\widetilde{x}^{s-1}) - \nabla f_{i_t}(x_{t-1}^s)}{np_{i_t}}\right\|^2\right] \\ &\le \mathbb{E}\left[\frac{1}{n^2 p_{i_t}^2}\left\|\nabla f_{i_t}(x_{t-1}^s) - \nabla f_{i_t}(\widetilde{x}^{s-1})\right\|^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{np_i}\left\|\nabla f_i(x_{t-1}^s) - \nabla f_i(\widetilde{x}^{s-1})\right\|^2 \\ &\le 2\widetilde{L}\left[f(\widetilde{x}^{s-1}) - f(x_{t-1}^s) + \langle \nabla f(x_{t-1}^s),\ x_{t-1}^s - \widetilde{x}^{s-1}\rangle\right], \end{aligned}$$

where the first inequality follows from the fact that $\mathbb{E}[\|\mathbb{E}[x] - x\|^2] = \mathbb{E}[\|x\|^2] - \|\mathbb{E}[x]\|^2 \le \mathbb{E}[\|x\|^2]$, and the second inequality holds due to (12). ■

Property 2 (Lan (2012)) *Assume that z^* is an optimal solution of the following problem,*

$$\min_z \frac{\nu}{2}\|z - z_0\|^2 + h(z),$$

where $h(z)$ is a convex function (but possibly non-differentiable). Then for any $z \in \mathbb{R}^d$,

$$h(z^*) + \frac{\nu}{2}\|z^* - z_0\|^2 + \frac{\nu}{2}\|z - z^*\|^2 \le h(z) + \frac{\nu}{2}\|z - z_0\|^2.$$

Property 3 *Assume that the stochastic momentum weight ω_s in Algorithm 2 satisfies the following conditions:*

$$\omega_0 \le 1 - \frac{1}{\alpha - 1} \quad \text{and} \quad \frac{1 - \omega_s}{\omega_s^2} = \frac{1}{\omega_{s-1}^2}, \tag{13}$$

where $\alpha = 1/(\widetilde{L}\eta)$. Then the following properties hold:

$$\omega_s = \frac{\sqrt{\omega_{s-1}^4 + 4\omega_{s-1}^2} - \omega_{s-1}^2}{2}, \quad \omega_s \le \frac{2}{s+2}.$$

Proof Using the equality in (13), it is easy to show that

$$\omega_s = \frac{\sqrt{\omega_{s-1}^4 + 4\omega_{s-1}^2} - \omega_{s-1}^2}{2} \geq 0.$$

In the following, we will prove by induction that $\omega_s \leq \frac{2}{s+2}$. Firstly, we have

$$\omega_0 \leq 1 - \frac{1}{\alpha - 1} \leq 1 = \frac{2}{0+2}.$$

Assume that $\omega_{s-1} \leq \frac{2}{s+1}$, then we have

$$\begin{aligned}
\omega_s &= \frac{\sqrt{\omega_{s-1}^4 + 4\omega_{s-1}^2} - \omega_{s-1}^2}{2} = \frac{2}{1 + \sqrt{1 + \frac{4}{\omega_{s-1}^2}}} \\
&\leq \frac{2}{1 + \sqrt{1 + (s+1)^2}} \\
&\leq \frac{2}{s+2}.
\end{aligned}$$

This completes the proof. ■

Proof of Lemma 2:

Proof Let $\widetilde{\nabla}_t := \left[\nabla f_{i_t}(x_{t-1}^s) - \nabla f_{i_t}(\widetilde{x}^{s-1})\right]/(np_{i_t}) + \nabla f(\widetilde{x}^{s-1})$. Suppose each component function $f_i(\cdot)$ is L_i-smooth, which implies that the gradient of the average function $f(x)$ is convex and also Lipschitz-continuous, i.e., there exists a Lipschitz constant $L_f > 0$ such that for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\|,$$

whose equivalent form is

$$f(y) \leq f(x) + \langle \nabla f(x),\ y - x \rangle + \frac{L_f}{2}\|y - x\|^2.$$

Moreover, it is easy to verify that $L_f \leq L_{\text{avg}} = \frac{1}{n}\sum_{j=1}^n L_j \leq \widetilde{L}$. Let $\eta = 1/(\widetilde{L}\alpha)$ and $\alpha > 2$ be a suitable constant, then we have

$$\begin{aligned}
f(x_t^s) &\leq f(x_{t-1}^s) + \left\langle \nabla f(x_{t-1}^s),\ x_t^s - x_{t-1}^s \right\rangle + \frac{L_f}{2}\left\|x_t^s - x_{t-1}^s\right\|^2 \\
&\leq f(x_{t-1}^s) + \left\langle \nabla f(x_{t-1}^s),\ x_t^s - x_{t-1}^s \right\rangle + \frac{\widetilde{L}}{2}\left\|x_t^s - x_{t-1}^s\right\|^2 \\
&= f(x_{t-1}^s) + \left\langle \nabla f(x_{t-1}^s),\ x_t^s - x_{t-1}^s \right\rangle + \frac{\widetilde{L}\alpha}{2}\left\|x_t^s - x_{t-1}^s\right\|^2 - \frac{\widetilde{L}(\alpha-1)}{2}\left\|x_t^s - x_{t-1}^s\right\|^2 \quad (14) \\
&= f(x_{t-1}^s) + \left\langle \widetilde{\nabla}_t,\ x_t^s - x_{t-1}^s \right\rangle + \frac{\widetilde{L}\alpha}{2}\|x_t^s - x_{t-1}^s\|^2 - \frac{\widetilde{L}(\alpha-1)}{2}\|x_t^s - x_{t-1}^s\|^2 \\
&\quad + \left\langle \nabla f(x_{t-1}^s) - \widetilde{\nabla}_t,\ x_t^s - x_{t-1}^s \right\rangle.
\end{aligned}$$

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\langle \nabla f(x_{t-1}^s) - \widetilde{\nabla}_t,\, x_t^s - x_{t-1}^s \Big\rangle\Big] \\
\le& \mathbb{E}\left[\frac{1}{2\widetilde{L}(\alpha-1)}\left\|\nabla f(x_{t-1}^s) - \widetilde{\nabla}_t\right\|^2 + \frac{\widetilde{L}(\alpha-1)}{2}\left\|x_t^s - x_{t-1}^s\right\|^2\right] \\
\le& \frac{1}{\alpha-1}\big[f(\widetilde{x}^{s-1}) - f(x_{t-1}^s) + \big\langle \nabla f(x_{t-1}^s),\, x_{t-1}^s - \widetilde{x}^{s-1}\big\rangle\big] + \frac{\widetilde{L}(\alpha-1)}{2}\mathbb{E}\Big[\left\|x_t^s - x_{t-1}^s\right\|^2\Big],
\end{aligned} \tag{15}
$$

where the first inequality holds due to the Young's inequality, i.e., $a^T b \le \|a\|^2/(2\theta) + \theta\|b\|^2/2$ for all $\theta > 0$, and the second inequality follows from Lemma 1.

Taking the expectation over the random choice of i_t, and substituting the inequality (15) into the inequality (14), then we have

$$
\begin{aligned}
\mathbb{E}[F(x_t^s)] \le& f(x_{t-1}^s) + \mathbb{E}\left[\Big\langle \widetilde{\nabla}_t,\, x_t^s - x_{t-1}^s \Big\rangle + \frac{\widetilde{L}\alpha}{2}\|x_t^s - x_{t-1}^s\|^2 + g(x_t^s)\right] \\
&+ \frac{1}{\alpha-1}\big[f(\widetilde{x}^{s-1}) - f(x_{t-1}^s) + \big\langle \nabla f(x_{t-1}^s),\, x_{t-1}^s - \widetilde{x}^{s-1}\big\rangle\big] \\
\le& f(x_{t-1}^s) + \mathbb{E}\left[\omega_{s-1}\Big\langle \widetilde{\nabla}_t,\, y_t^s - y_{t-1}^s \Big\rangle + \frac{\widetilde{L}\alpha\omega_{s-1}^2}{2}\|y_t^s - y_{t-1}^s\|^2 + \omega_{s-1} g(y_t^s)\right] \\
&+ (1-\omega_{s-1})g(\widetilde{x}^{s-1}) + \frac{1}{\alpha-1}\big[f(\widetilde{x}^{s-1}) - f(x_{t-1}^s) + \big\langle \nabla f(x_{t-1}^s),\, x_{t-1}^s - \widetilde{x}^{s-1}\big\rangle\big] \\
\le& f(x_{t-1}^s) + \mathbb{E}\left[\omega_{s-1}\Big\langle \widetilde{\nabla}_t,\, x^\star - y_{t-1}^s \Big\rangle + \frac{\widetilde{L}\alpha\omega_{s-1}^2}{2}(\|x^\star - y_{t-1}^s\|^2 - \|x^\star - y_t^s\|^2) + \omega_{s-1} g(x^\star)\right] \\
&+ (1-\omega_{s-1})g(\widetilde{x}^{s-1}) + \frac{1}{\alpha-1}\big[f(\widetilde{x}^{s-1}) - f(x_{t-1}^s) + \big\langle \nabla f(x_{t-1}^s),\, x_{t-1}^s - \widetilde{x}^{s-1}\big\rangle\big] \\
=& f(x_{t-1}^s) + \mathbb{E}\left[\frac{\widetilde{L}\alpha\omega_{s-1}^2}{2}\big(\|x^\star - y_{t-1}^s\|^2 - \|x^\star - y_t^s\|^2\big) + \omega_{s-1} g(x^\star)\right] + (1-\omega_{s-1})g(\widetilde{x}^{s-1}) \\
&+ \left\langle \nabla f(x_{t-1}^s),\, \omega_{s-1}x^\star + (1-\omega_{s-1})\widetilde{x}^{s-1} - x_{t-1}^s + \frac{1}{\alpha-1}(x_{t-1}^s - \widetilde{x}^{s-1}) \right\rangle \\
&+ \mathbb{E}\big[\big\langle -\nabla f_{i_t}(\widetilde{x}^{s-1}) + \nabla f(\widetilde{x}^{s-1}),\, \omega_{s-1}x^\star + (1-\omega_{s-1})\widetilde{x}^{s-1} - x_{t-1}^s \big\rangle\big] + \frac{f(\widetilde{x}^{s-1}) - f(x_{t-1}^s)}{\alpha-1} \\
=& f(x_{t-1}^s) + \mathbb{E}\left[\frac{\widetilde{L}\alpha\omega_{s-1}^2}{2}\big(\|x^\star - y_{t-1}^s\|^2 - \|x^\star - y_t^s\|^2\big) + \omega_{s-1} g(x^\star)\right] + (1-\omega_{s-1})g(\widetilde{x}^{s-1}) \\
&+ \left\langle \nabla f(x_{t-1}^s),\, \omega_{s-1}x^\star + (1-\omega_{s-1})\widetilde{x}^{s-1} - x_{t-1}^s + \frac{1}{\alpha-1}(x_{t-1}^s - \widetilde{x}^{s-1}) \right\rangle \\
&+ \frac{1}{\alpha-1}\left(f(\widetilde{x}^{s-1}) - f(x_{t-1}^s)\right),
\end{aligned} \tag{16}
$$

where the first inequality holds due to the inequalities (14) and (15); the second inequality follows from the facts that $x_t^s = \widetilde{x}^{s-1} + \omega_{s-1}(y_t^s - \widetilde{x}^{s-1}) = \omega_{s-1}y_t^s + (1-\omega_{s-1})\widetilde{x}^{s-1}$, $x_t^s - x_{t-1}^s = \omega_{s-1}(y_t^s - y_{t-1}^s)$, and

$$
g(\omega_{s-1}y_t^s + (1-\omega_{s-1})\widetilde{x}^{s-1}) \le \omega_{s-1}g(y_t^s) + (1-\omega_{s-1})g(\widetilde{x}^{s-1}).
$$

Since y_t^s is the optimal solution of the problem (5), the third inequality follows from Property 2 with $z^* = y_t^s$, $z = x^\star$, $z_0 = y_{t-1}^s$, $\nu = \widetilde{L}\alpha\omega_{s-1} = \omega_{s-1}/\eta$ and $h(y) := \langle \widetilde{\nabla}_t,\, y - y_{t-1}^s\rangle + g(y)$. The first equality holds due to the facts that

$$\begin{aligned}
&\omega_{s-1}\left\langle \widetilde{\nabla}_t,\ x^\star - y_{t-1}^s\right\rangle \\
&= \left\langle \widetilde{\nabla}_t,\ \omega_{s-1}x^\star + (1-\omega_{s-1})\widetilde{x}^{s-1} - x_{t-1}^s\right\rangle \\
&= \left\langle \nabla f_{i_t}(x_{t-1}^s),\ \omega_{s-1}x^\star + (1-\omega_{s-1})\widetilde{x}^{s-1} - x_{t-1}^s\right\rangle \\
&\quad + \left\langle -\nabla f_{i_t}(\widetilde{x}^{s-1}) + \nabla f(\widetilde{x}^{s-1}),\ \omega_{s-1}x^\star + (1-\omega_{s-1})\widetilde{x}^{s-1} - x_{t-1}^s\right\rangle,
\end{aligned}$$

and $\mathbb{E}[\nabla f_{i_t}(x_{t-1}^s)] = \nabla f(x_{t-1}^s)$, and the last equality follows from the fact that

$$\mathbb{E}\left[\langle -\nabla f_{i_t}(\widetilde{x}^{s-1}) + \nabla f(\widetilde{x}^{s-1}),\ \omega_{s-1}x^\star + (1-\omega_{s-1})\widetilde{x}^{s-1} - x_{t-1}^s\rangle\right] = 0.$$

Furthermore,

$$\begin{aligned}
&\left\langle \nabla f(x_{t-1}^s),\ (1-\omega_{s-1})\widetilde{x}^{s-1} + \omega_{s-1}x^\star - x_{t-1}^s + \frac{1}{\alpha-1}(x_{t-1}^s - \widetilde{x}^{s-1})\right\rangle \\
&= \left\langle \nabla f(x_{t-1}^s),\ \omega_{s-1}x^\star + (1-\omega_{s-1} - \frac{1}{\alpha-1})\widetilde{x}^{s-1} + \frac{1}{\alpha-1}x_{t-1}^s - x_{t-1}^s\right\rangle \\
&\le f\left(\omega_{s-1}x^\star + (1-\omega_{s-1} - \frac{1}{\alpha-1})\widetilde{x}^{s-1} + \frac{1}{\alpha-1}x_{t-1}^s\right) - f(x_{t-1}^s) \\
&\le \omega_{s-1}f(x^\star) + \left(1-\omega_{s-1} - \frac{1}{\alpha-1}\right)f(\widetilde{x}^{s-1}) + \frac{1}{\alpha-1}f(x_{t-1}^s) - f(x_{t-1}^s),
\end{aligned} \tag{17}$$

where the first inequality holds due to the fact that $\langle \nabla f(x),\, y - x\rangle \le f(y) - f(x)$, and the last inequality follows from the convexity of the function $f(\cdot)$ and the assumption that $1-\omega_{s-1}-\frac{1}{\alpha-1} = 1 - \omega_{s-1} - \frac{\widetilde{L}\eta}{1-\widetilde{L}\eta} \ge 0$. Substituting the inequality (17) into the inequality (16), we have

$$\begin{aligned}
\mathbb{E}[F(x_t^s)] &\le f(x_{t-1}^s) + \mathbb{E}\left[\frac{\widetilde{L}\alpha\omega_{s-1}^2}{2}\left(\|x^\star - y_{t-1}^s\|^2 - \|x^\star - y_t^s\|^2\right) + \omega_{s-1}g(x^\star) + (1-\omega_{s-1})g(\widetilde{x}^{s-1})\right] \\
&\quad + \omega_{s-1}f(x^\star) + \left(1-\omega_{s-1} - \frac{1}{\alpha-1}\right)f(\widetilde{x}^{s-1}) + \frac{1}{\alpha-1}f(x_{t-1}^s) - f(x_{t-1}^s) \\
&\quad + \frac{1}{\alpha-1}\left(f(\widetilde{x}^{s-1}) - f(x_{t-1}^s)\right) \\
&= \omega_{s-1}F(x^\star) + (1-\omega_{s-1})F(\widetilde{x}^{s-1}) + \frac{\widetilde{L}\alpha\omega_{s-1}^2}{2}\mathbb{E}\left[\|x^\star - y_{t-1}^s\|^2 - \|x^\star - y_t^s\|^2\right].
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}[F(x_t^s) - F(x^\star)] \\
&\le (1-\omega_{s-1})\mathbb{E}\left[[F(\widetilde{x}^{s-1}) - F(x^\star)\right] + \frac{\widetilde{L}\alpha\omega_{s-1}^2}{2}\mathbb{E}\left[\|x^\star - y_{t-1}^s\|^2 - \|x^\star - y_t^s\|^2\right].
\end{aligned}$$

Since

$$\widetilde{x}^s = \frac{1}{m}\sum_{t=1}^m x_t^s \quad \text{and} \quad F\left(\frac{1}{m}\sum_{t=1}^m x_t^s\right) \le \frac{1}{m}\sum_{t=1}^m F(x_t^s),$$

by taking the expectation over the random choice of the history of random variables $i_1, \cdots, i_m$ on the above inequality, and summing it over $t = 1, \cdots, m$ at the s-th stage, then we have

$$\begin{aligned}
&\mathbb{E}[F(\widetilde{x}^s) - F(x^\star)] \\
&\leq (1-\omega_{s-1})\mathbb{E}\big[F(\widetilde{x}^{s-1}) - F(x^\star)\big] + \frac{\widetilde{L}\alpha\omega_{s-1}^2}{2m}\mathbb{E}\Big[\|x^\star - y_0^s\|^2 - \|x^\star - y_m^s\|^2\Big] \\
&= (1-\omega_{s-1})\mathbb{E}\big[F(\widetilde{x}^{s-1}) - F(x^\star)\big] + \frac{\omega_{s-1}^2}{2m\eta}\mathbb{E}\Big[\|x^\star - y_0^s\|^2 - \|x^\star - y_m^s\|^2\Big].
\end{aligned}$$

This completes the proof. ■

Appendix C.

Appendix C1: Proof of Theorem 3

Proof Since the regularizer $g(x)$ is μ-strongly convex, then the objective function $F(x)$ is also strongly convex with the parameter $\widetilde{\mu} \geq \mu$, i.e. there exists a constant $\widetilde{\mu} > 0$ such that for all $x \in \mathbb{R}^d$

$$F(x) \geq F(x^\star) + \xi^T(x - x^\star) + \frac{\widetilde{\mu}}{2}\|x - x^\star\|^2, \quad \forall \xi \in \partial F(x^\star),$$

where $\partial F(x)$ is the subdifferential of $F(\cdot)$ at x.

Since $0 \in \partial F(x^\star)$, then we have

$$F(x) - F(x^\star) \geq \frac{\widetilde{\mu}}{2}\|x - x^\star\|^2 \geq \frac{\mu}{2}\|x - x^\star\|^2. \tag{18}$$

Using the above inequality, Lemma 2 with $\omega_s = \omega$ for all stages, and $y_0^s = \widetilde{x}^{s-1}$, we have

$$\begin{aligned}
&\mathbb{E}[F(\widetilde{x}^s) - F(x^\star)] \\
&\leq (1-\omega)\mathbb{E}\big[F(\widetilde{x}^{s-1}) - F(x^\star)\big] + \frac{\widetilde{L}\alpha\omega^2}{2m}\mathbb{E}\Big[\|x^\star - y_0^s\|^2 - \|x^\star - y_m^s\|^2\Big] \\
&\leq (1-\omega)\mathbb{E}\big[F(\widetilde{x}^{s-1}) - F(x^\star)\big] + \frac{\widetilde{L}\alpha\omega^2}{\mu m}\big[F(\widetilde{x}^{s-1}) - F(x^\star)\big] \\
&= \left(1 - \omega + \frac{\widetilde{L}\alpha\omega^2}{\mu m}\right)\mathbb{E}\big[F(\widetilde{x}^{s-1}) - F(x^\star)\big] \\
&= \left(1 - \omega + \frac{\omega^2}{\mu m\eta}\right)\mathbb{E}\big[F(\widetilde{x}^{s-1}) - F(x^\star)\big],
\end{aligned}$$

where the first inequality holds due to Lemma 2, and the second inequality follows from the inequality in (18).

This completes the proof. ■

Table 2: Theoretical suggestion for the parameters η, ω, and m.

Condition	Learning rate η	Parameter ω	Epoch Length m
$m\mu/\widetilde{L} \in [0.68623, 145.72]$	$\frac{2}{5}\sqrt{1/(\mu m\widetilde{L})}$	$\frac{2}{25}\sqrt{m\mu/\widetilde{L}}$	$\Theta(n)$
otherwise	$1/(5\widetilde{L})$	$1/5$	$2\widetilde{L}/\mu$

Appendix C2: Proof of Corollary 4

For Algorithm 1 with Option I, the theoretical suggestion of the parameter settings for the learning rate η, the momentum parameter ω and the epoch size m is shown in Table 2.

Proof Using the inequality in Theorem 3, we have

$$\mathbb{E}\big[F(\widetilde{x}^S) - F(x^\star)\big] \leq \left(1 - \omega + \frac{\omega^2}{\mu m\eta}\right)^S [F(\widetilde{x}^0) - F(x^\star)].$$

Then by setting $\eta = \sqrt{\frac{1}{a^2\mu m\widetilde{L}}}$, $\omega = \sqrt{\frac{m\mu}{b^2\widetilde{L}}}$ for some constants a and b, $m = \Theta(n)$, we have

$$\left(1 - \omega + \frac{\omega^2}{\mu m\eta}\right)^S = \left(1 - \frac{b-a}{b^2}\sqrt{\frac{m\mu}{\widetilde{L}}}\right)^S,$$

which means that our algorithm needs

$$S = O\left(\frac{b^2}{b-a}\sqrt{\frac{\widetilde{L}}{\mu n}}\right)\log\frac{F(\tilde{x}^0) - F(x^\star)}{\varepsilon},$$

epochs to an ε-suboptimal solution. Then the oracle complexity of Algorithm 1 with Option I is

$$\mathcal{O}(S(m+n)) = \mathcal{O}\left(\frac{b^2}{b-a}\sqrt{\frac{n\widetilde{L}}{\mu}}\log\frac{F(\tilde{x}^0) - F(x^\star)}{\varepsilon}\right).$$

Next we need to find the constants a, b as well as a region for $m\mu/\widetilde{L}$ that makes the above bound valid subject to some constrains,

$$0 < \omega \leq 1 - \frac{\widetilde{L}\eta}{1 - \widetilde{L}\eta}. \tag{19}$$

By substituting our parameter settings, we get

$$\frac{1}{b}\sqrt{\frac{m\mu}{\widetilde{L}}} - \left(\frac{1}{ab} + 1\right) + \frac{2}{a}\sqrt{\frac{\widetilde{L}}{\mu m}} \leq 0.$$

In order for the above inequality to has a solution, the constants a and b should satisfy the following inequalities:

$$\begin{cases} b > a > 0, \\ ab \leq 3 - 2\sqrt{2}, \text{ or } ab \geq 3 + 2\sqrt{2}. \end{cases}$$

Table 3: Theoretical suggestion for the parameters η, ω, and m.

Condition	Learning rate η	Parameter ω	Epoch Length m
$m\mu/\widetilde{L} \le 3/4$	$1/(3\widetilde{L})$	$\sqrt{(m\mu)/(3\widetilde{L})}$	$\Theta(n)$
$m\mu/\widetilde{L} > 3/4$	$1/(4m\mu)$	$1/2$	$\Theta(n)$

Suppose that the above inequalities are satisfied. Let ζ_1, ζ_2 with $\zeta_1 \le \zeta_2$ be the solutions to $x^2/b - (1/ab+1)x + 2/a = 0$, if $m\mu/\widetilde{L}$ satisfies

$$\zeta_1^2 \le \frac{m\mu}{\widetilde{L}} \le \zeta_2^2, \tag{20}$$

then the oracle complexity in this case is $\mathcal{O}\left(\sqrt{n\widetilde{L}/\mu}\log\frac{F(\tilde{x}^0)-F(x^\star)}{\varepsilon}\right)$.

For example, let $a = 2.5$, $b = 12.5$, then the range in (20) is from approximately 0.68623 to 145.72, that is, $m\mu/\widetilde{L} \in [0.68623, 145.72]$.

Now we consider the other case, i.e., out of the range in (20). Setting $\omega = 1/5$, $\eta = 1/(5\widetilde{L})$, $m = 2\widetilde{L}/\mu$ (one can easily verify that this setting satisfies the constraint in (19)), we have

$$1 - \omega + \frac{\omega^2}{\mu m \eta} = 0.9.$$

Thus, the oracle complexity for this case is $\mathcal{O}\left((n + \widetilde{L}/\mu)\log\frac{F(\tilde{x}^0)-F(x^\star)}{\varepsilon}\right)$. ■

Appendix C3: Proof of Corollary 5

For Algorithm 1 with Option II, the theoretical suggestion of the parameter settings for the learning rate η, the momentum parameter ω and the epoch size m is shown in Table 3.

Proof Using Lemma 2 and $\omega_s \equiv \omega$, we have

$$\mathbb{E}[F(\widetilde{x}^s) - F(x^\star)] \le (1-\omega)\mathbb{E}\left[F(\widetilde{x}^{s-1}) - F(x^\star)\right] + \frac{\omega^2}{2\eta m}\mathbb{E}\left[\|x^\star - y_0^s\|^2 - \|x^\star - y_m^s\|^2\right].$$

Let $\tilde{\Delta}_s = F(\tilde{x}^s) - F(x^\star)$, $\Lambda_t^s = \|x^\star - y_t^s\|^2$, the above inequality becomes

$$\mathbb{E}\left[\tilde{\Delta}_s\right] \le (1-\omega)\mathbb{E}\left[\tilde{\Delta}_{s-1}\right] + \frac{\omega^2}{2\eta m}\mathbb{E}\left[\Lambda_0^s - \Lambda_m^s\right].$$

Subtracting $(1-\omega)\mathbb{E}[\tilde{\Delta}_s]$ to both sides of the above inequality, we can rewrite the inequality as

$$\mathbb{E}\left[\tilde{\Delta}_s\right] \le \frac{1-\omega}{\omega}\mathbb{E}\left[\tilde{\Delta}_{s-1} - \tilde{\Delta}_s\right] + \frac{\omega}{2\eta m}\mathbb{E}\left[\Lambda_0^s - \Lambda_m^s\right].$$

Assume that our algorithm needs to restart every $\mathcal{S}$ epochs. Then in $\mathcal{S}$ epochs, by summing the above inequality overt $s = 1 \dots \mathcal{S}$, we have

$$\begin{aligned}\sum_{s=1}^{\mathcal{S}} \mathbb{E}\left[\tilde{\Delta}_s\right] &\leq \frac{1-\omega}{\omega}\mathbb{E}\left[\tilde{\Delta}_0 - \tilde{\Delta}_{\mathcal{S}}\right] + \frac{\omega}{2\eta m}\mathbb{E}\left[\Lambda_0^1 - \Lambda_m^{\mathcal{S}}\right] \\ &\leq \Big(\frac{1-\omega}{\omega} + \frac{\omega}{\eta m \mu}\Big)\tilde{\Delta}_0,\end{aligned}$$

where the last inequality holds due to the μ-strongly convex property of Problem (1). Choosing the initial vector as $x_0^{new} = \frac{1}{\mathcal{S}}\sum_{s=1}^{\mathcal{S}} \tilde{x}^s$ for the restart, we have

$$\tilde{\Delta}_0^{new} \leq \frac{\frac{1-\omega}{\omega} + \frac{\omega}{\eta m \mu}}{\mathcal{S}}\tilde{\Delta}_0.$$

By setting $\mathcal{S} = 2 \cdot \big(\frac{1-\omega}{\omega} + \frac{\omega}{\eta m \mu}\big)$, we have that $\tilde{\Delta}_0$ decreases by a factor of $1/2$ every $\mathcal{S}$ epochs. So in order to achieve an ε-suboptimal solution, the algorithm needs to perform totally $\mathcal{O}(\log \frac{\tilde{\Delta}_0}{\varepsilon})$ rounds of $\mathcal{S}$ epochs.

(I) We consider the first case, i.e., $m\mu/\widetilde{L} \leq 3/4$. Setting $m = \Theta(n)$, $\eta = 1/(3\widetilde{L})$ and $\omega = \sqrt{(m\mu)/(3\widetilde{L})} \leq 1/2$ (which satisfy the constraint in (19)), we have $\mathcal{S} = O(\sqrt{\widetilde{L}/(n\mu)})$, and then the oracle complexity of our algorithm is

$$\mathcal{O}\left(\mathcal{S} \cdot O\left(\log \frac{F(\tilde{x}^0) - F(x^\star)}{\varepsilon}\right) \cdot (m+n)\right) = \mathcal{O}\left(\sqrt{\frac{n\widetilde{L}}{\mu}} \log \frac{F(\tilde{x}^0) - F(x^\star)}{\varepsilon}\right).$$

(II) We then consider the other case, i.e., $m\mu/\widetilde{L} > 3/4$. Setting $m = \Theta(n)$, $\eta = 1/(4m\mu) < 1/(3\widetilde{L})$ and $\omega = 1/2$ (which satisfy constraint in (19)), we have $\mathcal{S} = 6 \in O(1)$. Therefore, the oracle complexity of our algorithm in this case is

$$\mathcal{O}\left(n \log \frac{F(\tilde{x}^0) - F(x^\star)}{\varepsilon}\right).$$

In short, all the results imply that the oracle complexity of Algorithm 1 is

$$\mathcal{O}\left((n + \sqrt{n\widetilde{L}/\mu}) \log \frac{F(\tilde{x}^0) - F(x^\star)}{\varepsilon}\right).$$

This completes the proof. ■

Appendix D. Proof of Theorem 7

Proof Using Lemma 2, we have

$$\frac{1}{\omega_{s-1}^2}\mathbb{E}[F(\widetilde{x}^s) - F(x^\star)] \leq \frac{1-\omega_{s-1}}{\omega_{s-1}^2}\mathbb{E}\big[F(\widetilde{x}^{s-1}) - F(x^\star)\big] + \frac{\widetilde{L}\alpha}{2m}\mathbb{E}\Big[\|x^\star - y_0^s\|^2 - \|x^\star - y_m^s\|^2\Big],$$

for all $s = 1, \ldots, S$. By the update rules $y_0^s = y_m^{s-1}$ and $(1-\omega_s)/\omega_s^2 = 1/\omega_{s-1}^2$, and summing the above inequality over $s = 1, 2, \cdots, S$, we have

$$\frac{1}{\omega_{S-1}^2}\mathbb{E}\big[F(\widetilde{x}^S) - F(x^\star)\big] \le \frac{1-\omega_0}{\omega_0^2}\left[F(\widetilde{x}^0) - F(x^\star)\right] + \frac{\widetilde{L}\alpha}{2m}\mathbb{E}\Big[\big\|x^\star - y_0^0\big\|^2 - \big\|x^\star - y_m^S\big\|^2\Big].$$

Using Property 3, we have

$$\omega_s \le \frac{2}{s+2} \quad \text{and} \quad \omega_0 = 1 - \frac{\widetilde{L}\eta}{1-\widetilde{L}\eta} = 1 - \frac{1}{\alpha - 1},$$

where $\alpha = 1/(\widetilde{L}\eta)$. Then

$$\begin{aligned}
&\mathbb{E}\big[F(\widetilde{x}^S) - F(x^\star)\big] \\
&\le \frac{4(\alpha-1)}{(\alpha-2)^2(S+1)^2}\left[F(\widetilde{x}^0) - F(x^\star)\right] + \frac{2\widetilde{L}\alpha}{m(S+1)^2}\mathbb{E}\Big[\big\|x^\star - y_0^0\big\|^2 - \big\|x^\star - y_m^S\big\|^2\Big] \\
&\le \frac{4(\alpha-1)}{(\alpha-2)^2(S+1)^2}\left[F(\widetilde{x}^0) - F(x^\star)\right] + \frac{2}{m\eta(S+1)^2}\mathbb{E}\Big[\big\|x^\star - \widetilde{x}^0\big\|^2\Big].
\end{aligned}$$

This completes the proof. ■

Appendix E. Proof of Lemma 9

Before proving Lemma 9, we first give the following lemma (Konečný et al., 2016).

Lemma 2 *Let $\xi_i \in \mathbb{R}^d$ for all $i = 1, 2, \ldots, n$, and $\bar{\xi} := \frac{1}{n}\sum_{i=1}^n \xi_i$. b is the size of the mini-batch I_t, which is chosen independently and uniformly at random from all subsets of $[n]$. Then we have*

$$\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_t}\xi_i - \bar{\xi}\right\|^2\right] \le \frac{n-b}{nb(n-1)}\sum_{i=1}^n \|\xi_i\|^2.$$

Proof of Lemma 2:

Proof We extend the upper bound on the expected variance of the modified stochastic gradient estimator in Lemma 1 to the mini-batch setting, i.e., $b \ge 2$.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{\nabla} f_{I_t}(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_t}\left[\nabla f_i(x_{t-1}^s) - \nabla f_i(\widetilde{x}^{s-1})\right]/(np_i) + \nabla f(\widetilde{x}^{s-1}) - \nabla f(x_{t-1}^s)\right\|^2\right] \\
&\le \frac{n-b}{b(n-1)}\frac{1}{n}\sum_{i=1}^n \frac{1}{np_i}\left\|\nabla f_i(x_{t-1}^s) - \nabla f_i(\widetilde{x}^{s-1})\right\|^2 \\
&\le \frac{2\widetilde{L}(n-b)}{b(n-1)}\left[f(\widetilde{x}^{s-1}) - f(x_{t-1}^s) - \left\langle \nabla f(x_{t-1}^s),\ \widetilde{x}^{s-1} - x_{t-1}^s\right\rangle\right],
\end{aligned}$$

where the first inequality follows from Lemma 2, and the second inequality holds due to Theorem 2.1.5 in Nesterov (2004), i.e.,

$$\left\|\nabla f_i(x_{t-1}^s) - \nabla f_i(\widetilde{x}^{s-1})\right\|^2 \leq 2L_i\left[f_i(\widetilde{x}^{s-1}) - f_i(x_{t-1}^s) - \left\langle \nabla f_i(x_{t-1}^s),\ \widetilde{x}^{s-1} - x_{t-1}^s\right\rangle\right].$$

This completes the proof. ■

Appendix F. Proof of Theorem 10

The proof of Theorem 10 is similar to that of Theorem 7. Hence, we briefly sketch the proof of Theorem 10 for the sake of completeness.

Proof Let

$$\omega_0 = 1 - \frac{\tau(b)\widetilde{L}\eta}{1-\widetilde{L}\eta} = 1 - \frac{\tau(b)}{\alpha - 1},$$

where $\alpha = \frac{1}{\widetilde{L}\eta}$, and $y_0^0 = \widetilde{x}^0$, then we have

$$\begin{aligned}
&\mathbb{E}[F(\widetilde{x}^s) - F(x^\star)] \\
&\leq \frac{4(\alpha-1)\tau(b)}{(\alpha - 1 - \tau(b))^2(s+1)^2}\left[F(\widetilde{x}^0) - F(x^\star)\right] + \frac{2\widetilde{L}\alpha}{m(s+1)^2}\mathbb{E}\left[\left\|x^\star - \widetilde{x}^0\right\|^2 - \left\|x^\star - y_m^s\right\|^2\right] \\
&\leq \frac{4(\alpha-1)\tau(b)}{(\alpha - 1 - \tau(b))^2(s+1)^2}\left[F(\widetilde{x}^0) - F(x^\star)\right] + \frac{2}{\eta m(s+1)^2}\mathbb{E}\left[\left\|x^\star - \widetilde{x}^0\right\|^2\right].
\end{aligned}$$

This completes the proof. ■

References

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

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

R. Babanezhad, M. O. Ahmed, A. Virani, M. Schmidt, J. Konecny, and S. Sallinen. Stop wasting my gradients: Practical SVRG. In *NIPS*, pages 2242–2250, 2015.

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

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

J. Konečný, J. Liu, P. Richtárik, , and M. Takáč. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE J. Sel. Top. Sign. Proces.*, 10(2):242–255, 2016.

G. Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 133:365–397, 2012.

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

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

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publ., Boston, 2004.

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

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

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