Closing the Convergence Gaps in Non-Strongly Convex Optimization by Directly Accelerated SVRG with Double Compensation and Snapshots

Yuanyuan Liu[1], Fanhua Shang[2], Weixin An[1],

Hongying Liu[1] and Zhouchen Lin[3]

[1]Xidian University

[2]Tianjin University

[3]Peking University

ICML | 2022

Thirty-ninth International Conference on Machine Learning

Optimization Problems

Finite-Sum Composite Minimization Problem

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

where $f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ is the finite average of n convex component functions $f_i(x) : \mathbb{R}^d \to \mathbb{R}$, and $h(x) : \mathbb{R}^d \to \mathbb{R}$ is a "simple" possibly non-smooth convex function.

- Linear regression: $f_i(x) = \frac{1}{2}(a_i^T x - b_i)^2,\ \ h(x) = 0;$
- Logistic regression:
  $$f_i(x) = \log(1 + \exp(-b_i a_i^T x)),\ \ h(x) = \frac{\lambda}{2}\|x\|^2;$$
- Lasso: $f_i(x) = \frac{1}{2}(a_i^T x - b_i)^2,\ \ h(x) = \lambda\|x\|_1;$
- SVM: $f_i(x) = \max\{0, 1 - b_i\langle a_i, x\rangle\},\ \ h(x) = \frac{\lambda}{2}\|x\|^2;$
- ...

Optimization Problems

- **Equality-Constrained Finite-Sum Problem**

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

where $A\in\mathbb{R}^{d_1\times d}$, $f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, each $f_i(\cdot)$ is convex and $h(\cdot)$ is convex but possibly non-smooth.

Graph-Guided Fused Lasso

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

where $f_i(\cdot)$ is the logistic loss function, $\lambda\geq 0$ is the regularization parameter, $A = [G; I]$, and G is the sparsity pattern of the graph obtained by sparse inverse covariance selection.

Motivations and Contributions

Motivations

We attempt to answer the following questions, which are not fully addressed in the existing literature yet:

- For Problem (1), there is still a gap between the best-known oracle complexity [1] and its lower bound [2]. **Can we design a simple algorithm to close the gap in theory?**
- For structure-regularized problem (2), there is a big gap between the convergence rates of prior works and the lower bound in [3]. **Can we obtain the optimal convergence rate in both theory and practice?**

[1] Song, C., Jiang, Y., and Ma, Y. Variance reduction via accelerated dual averaging for finite-sum optimization. NeurIPS, 2020.
[2] Woodworth, B. and Srebro, N. Tight complexity bounds for optimizing composite objectives. NIPS, 2016.
[3] Xie, G., Luo, L., Lian, Y., and Zhang, Z. Lower complexity bounds for finite-sum convex-concave minimax optimization problems. ICML, 2020.

Motivations and Contributions

Contributions

- For Problem (1), we propose a novel directly accelerated stochastic variance reduced gradient (DAVIS) method, which has two snapshots and new momentum accelerated rules with a new compensated stochastic gradient operator.
- we prove that DAVIS obtains an **optimal** convergence rate $O(1/(nS^2))$, and the oracle complexity of DAVIS is $O(n+\sqrt{nL/\epsilon})$, which is identical to the lower bound in [1].

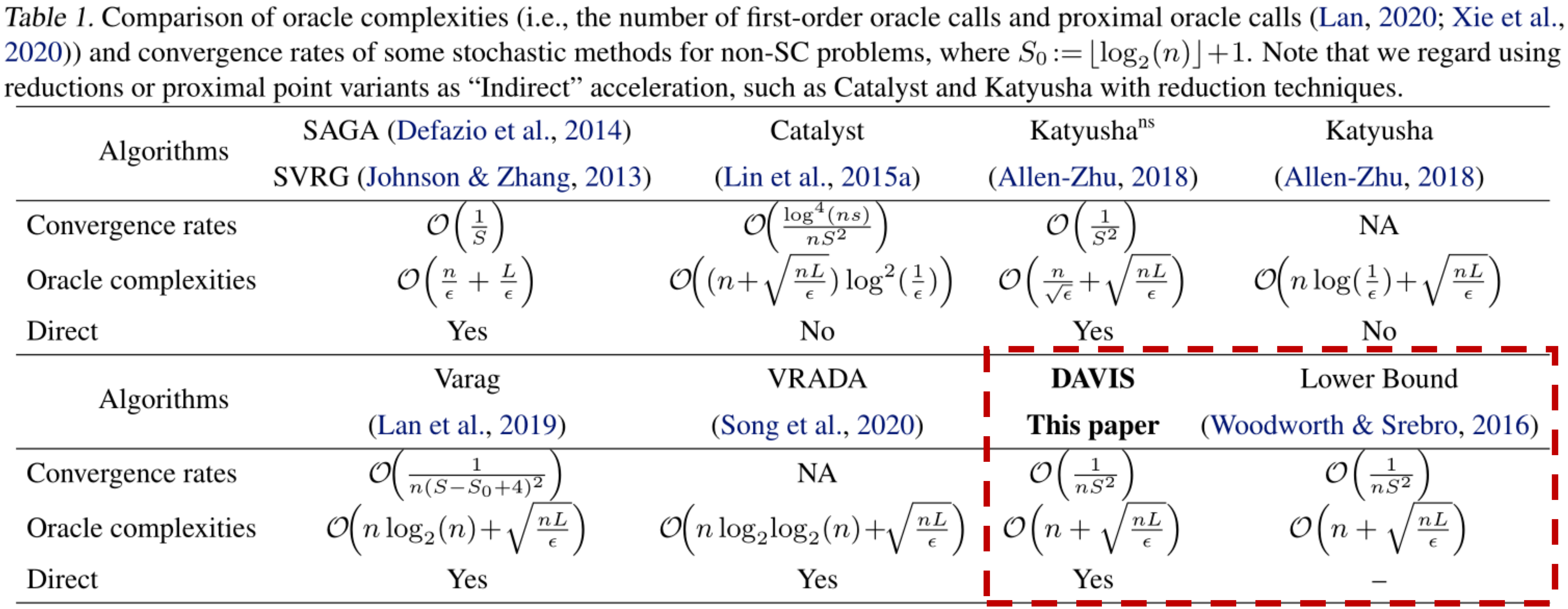


Table 1. Comparison of oracle complexities (i.e., the number of first-order oracle calls and proximal oracle calls (Lan, 2020; Xie et al., 2020)) and convergence rates of some stochastic methods for non-SC problems, where $S_0 := \lfloor \log_2(n) \rfloor + 1$. Note that we regard using reductions or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

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

Motivations and Contributions

Contributions

- We also propose a directly accelerated stochastic ADMM (DAVIS-ADMM) algorithm to solve Problem (2), and prove that DAVIS-ADMM attains the **optimal** rate $O\left(\frac{1}{nS}\right)$, and the **optimal** oracle complexity $O\left(n+\frac{L}{\epsilon}\right)$.

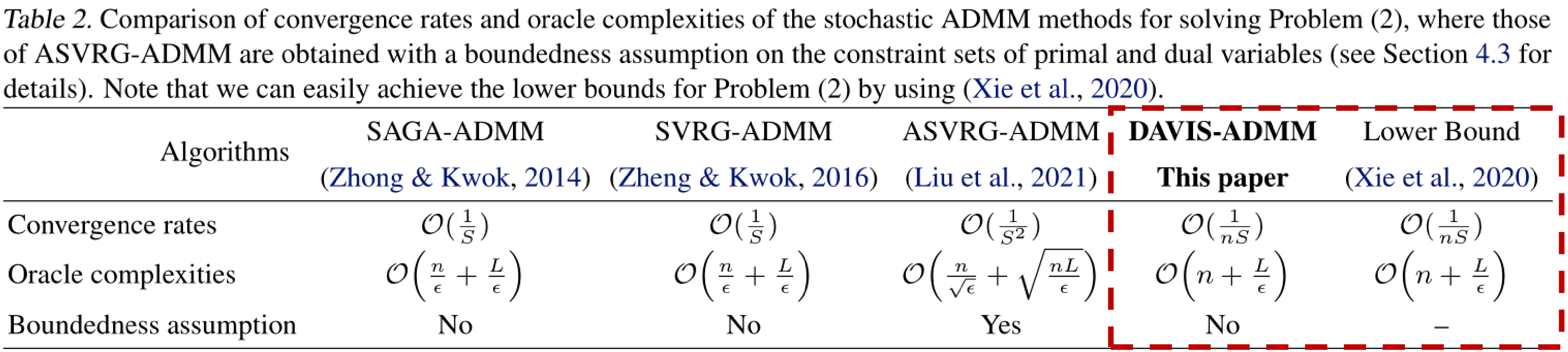


Table 2. Comparison of convergence rates and oracle complexities of the stochastic ADMM methods for solving Problem (2), where those of ASVRG-ADMM are obtained with a boundedness assumption on the constraint sets of primal and dual variables (see Section 4.3 for details). Note that we can easily achieve the lower bounds for Problem (2) by using (Xie et al., 2020).

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

Main Ideas of DAVIS for Problem (1)

New Scheme of Double Snapshots in Outer Loop

The second snapshot *The first snapshot*

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

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

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

Main Ideas of DAVIS for Problem (1)

New Stochastic Update Schemes in Inner Loop

Momentum Acceleration

$$y_k^s = \frac{\theta_s}{m} p_k^s + (1 - \frac{\theta_s}{m}) \overline{x}^{s-1}$$

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

where z_k^s is obtained by solving the following problem:

$$z_k^s \triangleq \underset{z}{\operatorname{argmin}} \left\{ \widehat{h}(z) + \langle \widetilde{\nabla}_{i_k}(y_k^s), z \rangle + \frac{m\theta_s}{2\eta} \| z - \alpha_k^s \|^2 \right\},$$

where $\widehat{h}(z) = h(z) + \frac{\theta_s}{2\eta} \| z - \overline{z}^{s-1} \|^2.$

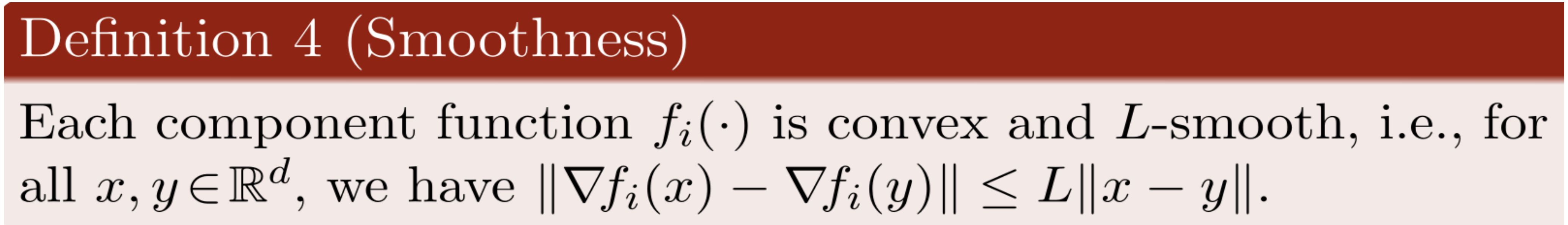
Main Ideas of DAVIS for Problem (1)

New Stochastic Update Schemes in Inner Loop

Compensated stocashtic gradient estimator：

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

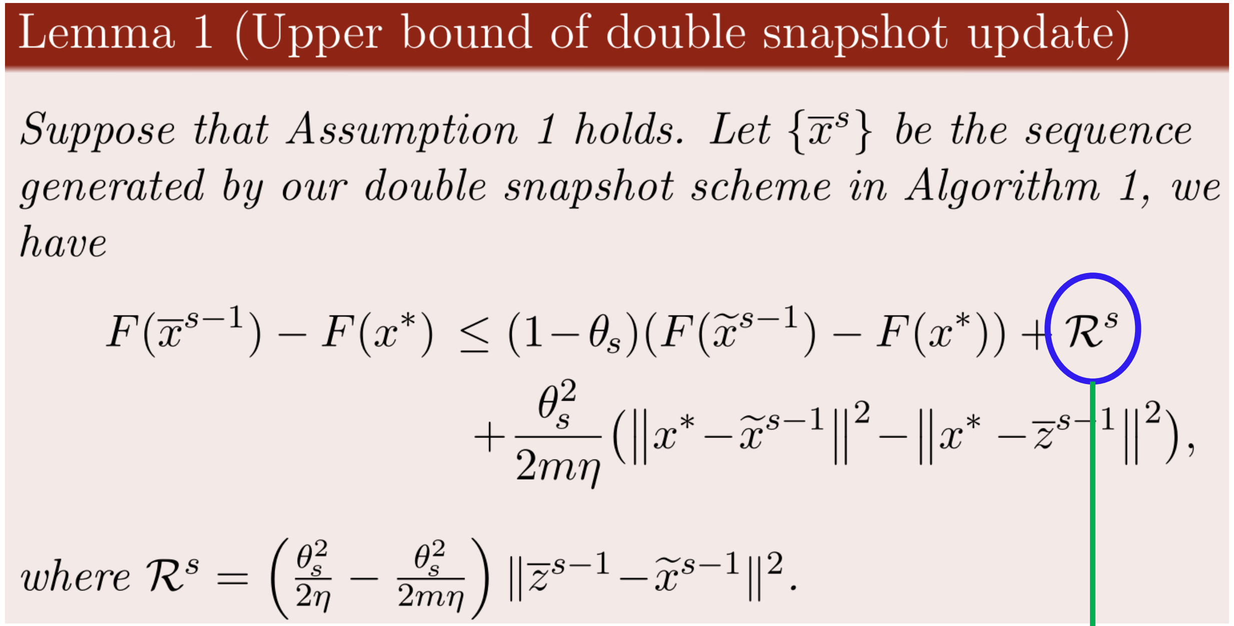
Optimal Convergence Guarantees



Definition 4 (Smoothness)

Each component function $f_i(\cdot)$ is convex and L-smooth, i.e., for all $x, y \in \mathbb{R}^d$, we have $\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x - y\|$.

■ Upper bound of double snapshot update



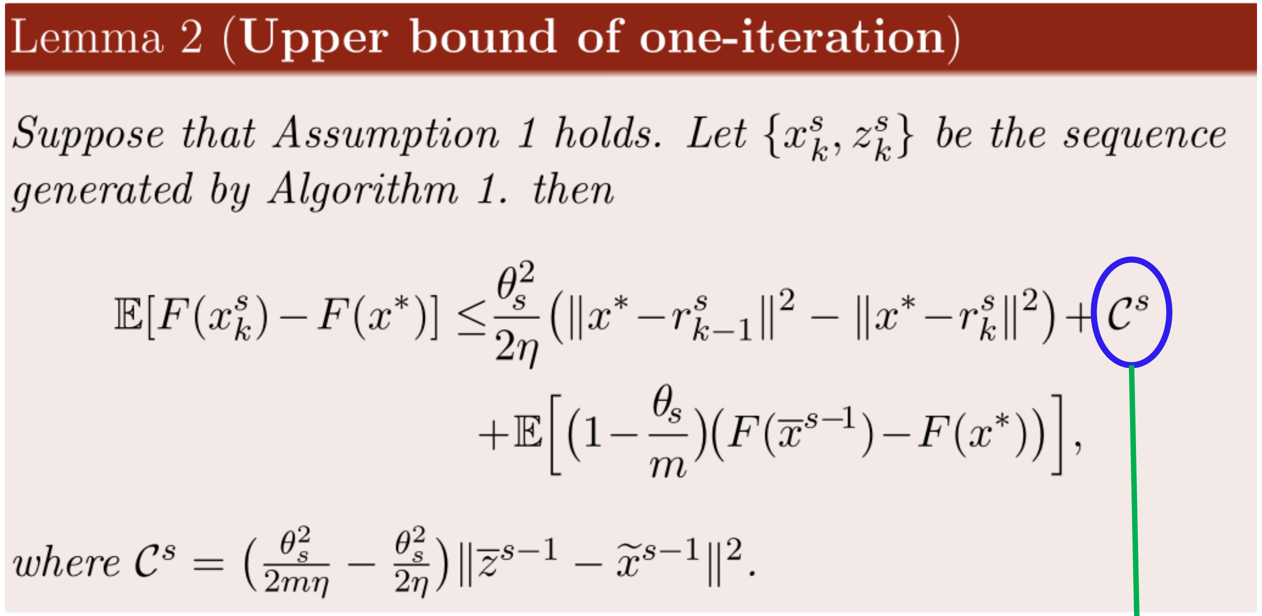
Lemma 1 (Upper bound of double snapshot update)

Suppose that Assumption 1 holds. Let $\{\overline{x}^s\}$ be the sequence generated by our double snapshot scheme in Algorithm 1, we have

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

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

■ Upper bound of one-iteration



Lemma 2 (Upper bound of one-iteration)

Suppose that Assumption 1 holds. Let $\{x_k^s, z_k^s\}$ be the sequence generated by Algorithm 1. then

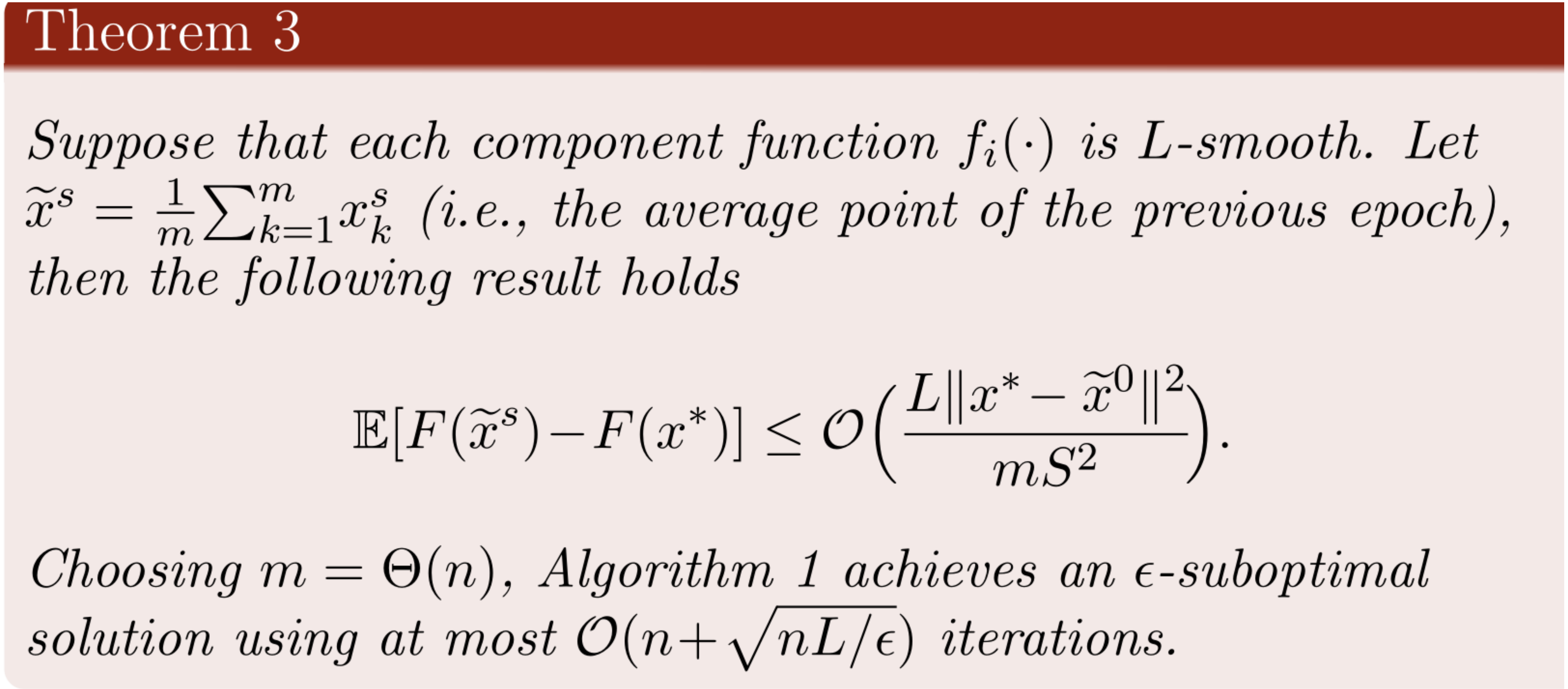
$$\mathbb{E}[F(x_k^s) - F(x^*)] \le \frac{\theta_s^2}{2\eta}\left(\|x^* - r_{k-1}^s\|^2 - \|x^* - r_k^s\|^2\right) + \boxed{\mathcal{C}^s} + \mathbb{E}\left[\left(1 - \frac{\theta_s}{m}\right)\left(F(\overline{x}^{s-1}) - F(x^*)\right)\right],$$

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

$$\mathcal{R}^s \longrightarrow \mathcal{R}^s + \mathcal{C}^s = 0 \longleftarrow \mathcal{C}^s$$

Optimal Convergence Guarantees

■ OPTIMAL CONVERGENCE RESULTS



Theorem 3

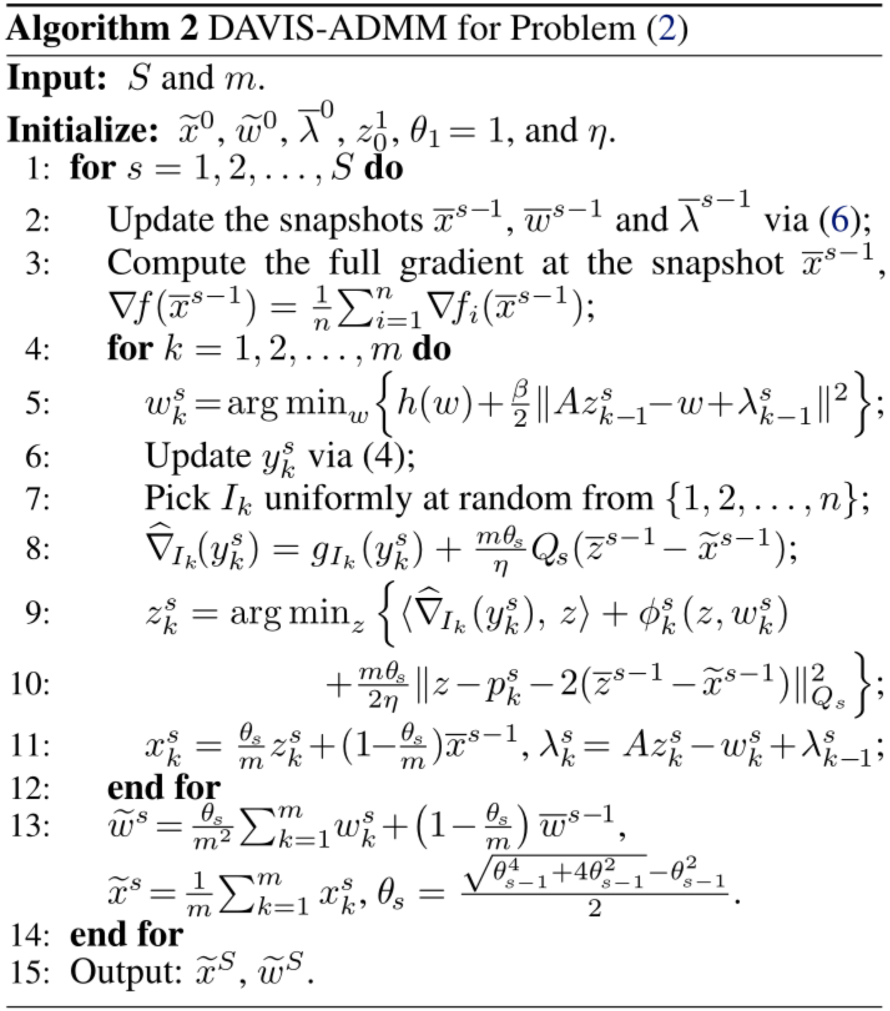
Suppose that each component function $f_i(\cdot)$ *is* L*-smooth. Let* $\widetilde{x}^s = \frac{1}{m}\sum_{k=1}^{m} x_k^s$ *(i.e., the average point of the previous epoch), then the following result holds*

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

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

Our DAVIS-ADMM Algorithm

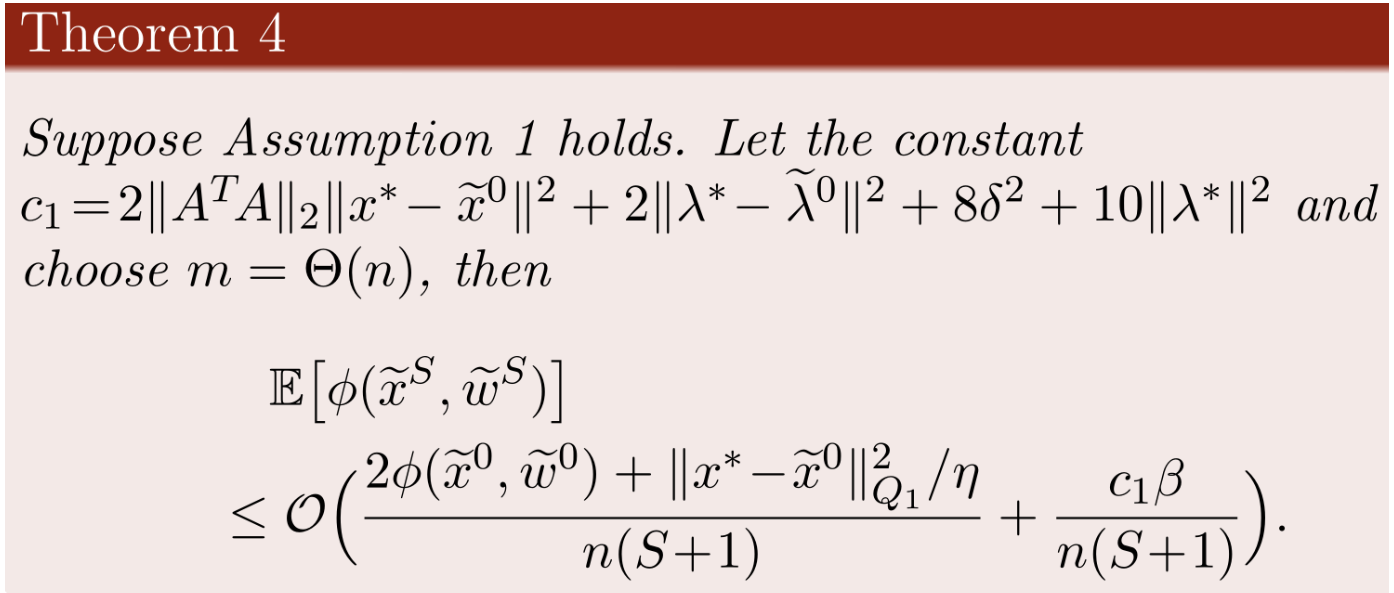
■ DAVIS-ADMM Algorithm



Algorithm 2 DAVIS-ADMM for Problem (2)

```
Input: $S$ and $m$.
Initialize: $\widetilde{x}^0$, $\widetilde{w}^0$, $\overline{\lambda}^0$, $z_0^1$, $\theta_1 = 1$, and $\eta$.
for $s = 1, 2, \ldots, S$ do
    Update the snapshots $\overline{x}^{s-1}$, $\overline{w}^{s-1}$ and $\overline{\lambda}^{s-1}$ via (6);
    Compute the full gradient at the snapshot $\overline{x}^{s-1}$, $\nabla f(\overline{x}^{s-1}) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\overline{x}^{s-1})$;
    for $k = 1, 2, \ldots, m$ do
        $w_k^s = \arg\min_w \left\{h(w) + \frac{\beta}{2}\|Az_{k-1}^s - w + \lambda_{k-1}^s\|^2\right\}$;
        Update $y_k^s$ via (4);
        Pick $I_k$ uniformly at random from $\{1, 2, \ldots, n\}$;
        $\widehat{\nabla}_{I_k}(y_k^s) = g_{I_k}(y_k^s) + \frac{m\theta_s}{\eta} Q_s(\overline{z}^{s-1} - \widetilde{x}^{s-1})$;
        $z_k^s = \arg\min_z \Big\{\langle \widehat{\nabla}_{I_k}(y_k^s),\ z\rangle + \phi_k^s(z, w_k^s)$
                    $+ \frac{m\theta_s}{2\eta}\|z - p_k^s - 2(\overline{z}^{s-1} - \widetilde{x}^{s-1})\|_{Q_s}^2 \Big\}$;
        $x_k^s = \frac{\theta_s}{m} z_k^s + (1 - \frac{\theta_s}{m})\overline{x}^{s-1}$, $\lambda_k^s = Az_k^s - w_k^s + \lambda_{k-1}^s$;
    end for
    $\widetilde{w}^s = \frac{\theta_s}{m^2}\sum_{k=1}^{m} w_k^s + \left(1 - \frac{\theta_s}{m}\right)\overline{w}^{s-1}$,
    $\widetilde{x}^s = \frac{1}{m}\sum_{k=1}^{m} x_k^s$, $\theta_s = \frac{\sqrt{\theta_{s-1}^4 + 4\theta_{s-1}^2} - \theta_{s-1}^2}{2}$.
end for
Output: $\widetilde{x}^S$, $\widetilde{w}^S$.
```

■ Optimal Convergence Guarantees



Theorem 4

Suppose Assumption 1 holds. Let the constant $c_1 = 2\|A^TA\|_2\|x^* - \widetilde{x}^0\|^2 + 2\|\lambda^* - \widetilde{\lambda}^0\|^2 + 8\delta^2 + 10\|\lambda^*\|^2$ *and choose* $m = \Theta(n)$, *then*

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

Experimental Results

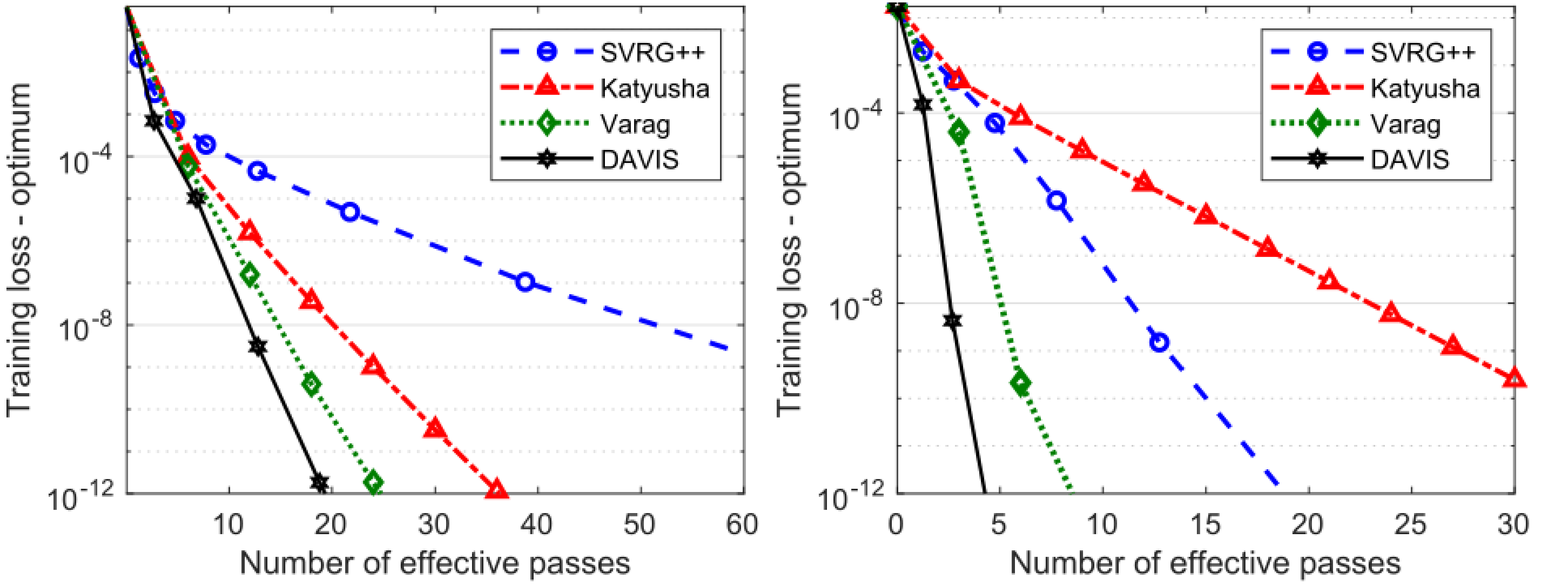


Figure 1. Comparison of all the methods for solving ℓ_1-norm regularized logistic regression problems on Adult and Covtype.

Experimental Results

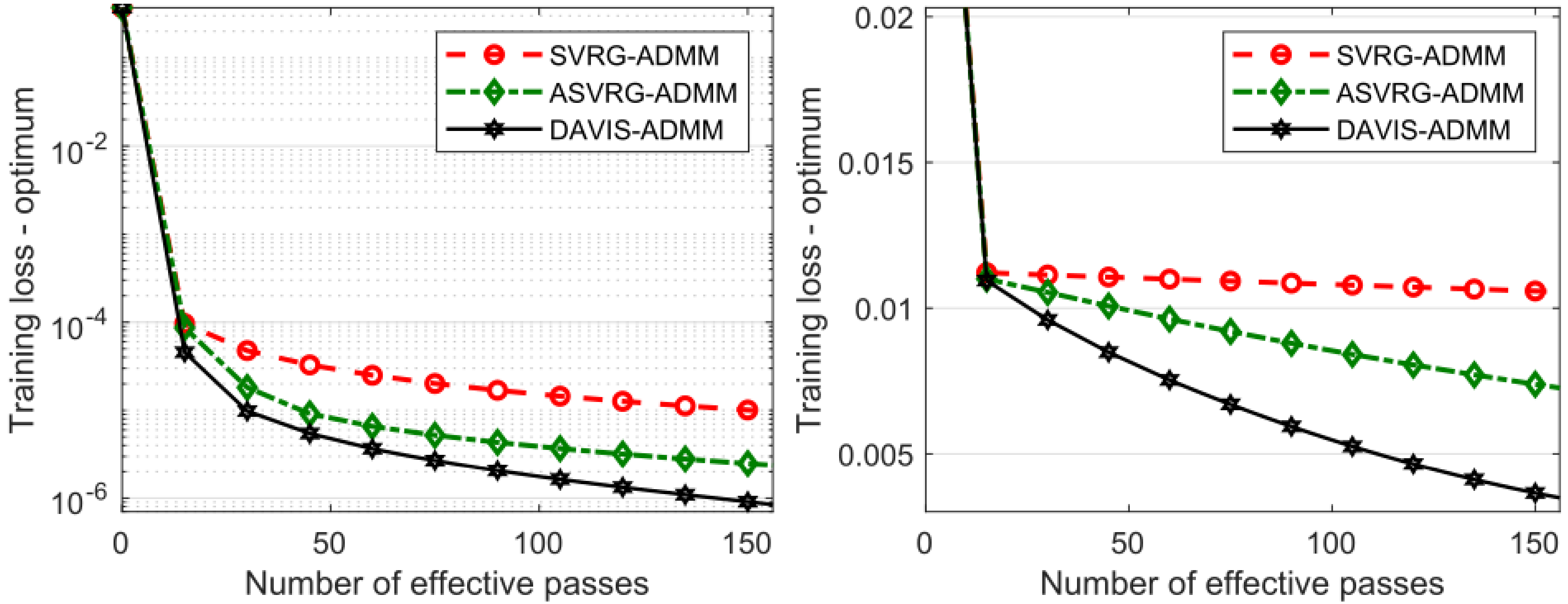


Figure 2. Comparison of all the methods for solving graph-guided fused Lasso problems on Adult and Covtype, where the regularization parameter is $\lambda = 10^{-5}$.

Thank you!