Supplementary Materials for "Guaranteed Sufficient Decrease for Stochastic Variance Reduced Gradient Optimization"

Fanhua Shang, Yuanyuan Liu*, Kaiwen Zhou, James Cheng, Kelvin K.W. Ng
Department of Computer Science and Engineering, The Chinese University of Hong Kong

Yuichi Yoshida
National Institute of Informatics, Tokyo, Japan

In this supplementary material, we give the detailed proofs for some lemmas, theorems and corollaries stated in the main paper. Moreover, we also report more experimental results for both of our algorithms on several dense and sparse data sets.

Notations

Throughout this paper, $\|\cdot\|$ denotes the standard Euclidean norm, and $\|\cdot\|_1$ is the ℓ_1-norm, i.e., $\|x\|_1 = \sum_{i=1}^{d}|x_i|$. We denote by $\nabla f(x)$ the full gradient of $f(x)$ if it is differentiable, or $\partial f(x)$ the subdifferential of $f(\cdot)$ at x if it is only Lipschitz continuous. Note that Assumption 2 is the general form for the two cases when $F(x)$ is smooth or non-smooth[1]. That is, if $F(x)$ is smooth, the inequality in (12) in Assumption 2 becomes the following form:

$$F(y) \geq F(x) + \nabla F(x)(y-x) + \frac{\mu}{2}\|y-x\|^2.$$

In the main paper, we assume that all component functions have the same smoothness parameter, L. In fact, we can extend the theoretical result for the case, when the gradients of all component functions have the same Lipschitz constant L, to the more general case, when some component functions $f_i(\cdot)$ have different degrees of smoothness.

Definition 1. *The SVRG estimator in the mini-batch setting is defined as follows:*

$$\widetilde{\nabla} f_{I_k^s}(x_k^s) = \frac{1}{b}\sum_{i\in I_k^s}\big[\nabla f_i(x_k^s) - \nabla f_i(\widetilde{x}^{s-1})\big] + \nabla f(\widetilde{x}^{s-1})$$

where $I_k^s \subset [n]$ is a mini-batch of size b.

[1]Strictly speaking, when the function $F(\cdot)$ is non-smooth, $\vartheta \in \partial F(x)$; while $F(\cdot)$ is smooth, $\vartheta = \nabla F(x)$.

(*) Corresponding author

Using the above definition, our algorithms naturally generalize to the mini-batch setting.

Appendix A: Proof of Theorem 1

Although the proposed SVRG-SD algorithm is a variant of SVRG, it is non-trivial to analyze its convergence property. Before proving Theorem 1, we first give the following lemma.

Lemma 1. *Let x^* be the optimal solution of Problem (1), then the following inequality holds*

$$\mathbb{E}\Big[\big\|\nabla f_{i_k^s}(x_{k-1}^s) - \nabla f(x_{k-1}^s) - \nabla f_{i_k^s}(\widetilde{x}^{s-1}) + \nabla f(\widetilde{x}^{s-1})\big\|^2\Big]$$
$$\leq 4L\big[F(x_{k-1}^s) - F(x^*) + F(\widetilde{x}^{s-1}) - F(x^*)\big].$$

Lemma 1 provides the upper bound on the expected variance of the variance reduced gradient estimator in (9) (i.e., the SVRG estimator independently introduced in [5, 10]), which satisfies $\mathbb{E}[\widetilde{\nabla} f_{i_k^s}(x_{k-1}^s)] = \nabla f(x_{k-1}^s)$. This lemma is essentially identical to Corollary 3.5 in [9] and Lemma A.2 in [2]. In addition, the upper bound on the variance of $\widetilde{\nabla} f_{i_k^s}(x_k^s)$ can be extended to the mini-batch setting as in [6].

Using Lemma 1, we immediately get the following result, which is useful in our convergence analysis.

Corollary 2. *For any $\alpha \geq \beta > 0$, the following inequality holds*

$$\alpha\mathbb{E}\Big[\big\|\nabla f_{i_k^s}(x_{k-1}^s) - \nabla f(x_{k-1}^s) - \nabla f_{i_k^s}(\widetilde{x}^{s-1}) + \nabla f(\widetilde{x}^{s-1})\big\|^2\Big]$$
$$- \beta\mathbb{E}\Big[\big\|\nabla f_{i_k^s}(x_{k-1}^s) - \nabla f_{i_k^s}(\widetilde{x}^{s-1})\big\|^2\Big]$$
$$\leq 4L(\alpha-\beta)\big[F(x_{k-1}^s) - F(x^*) + F(\widetilde{x}^{s-1}) - F(x^*)\big].$$

Proof.

$$\begin{aligned}
&\alpha\mathbb{E}\Big[\big\|\nabla f_{i_k^s}(x_{k-1}^s)-\nabla f(x_{k-1}^s)-\nabla f_{i_k^s}(\widetilde{x}^{s-1})+\nabla f(\widetilde{x}^{s-1})\big\|^2\Big]-\beta\mathbb{E}\Big[\big\|\nabla f_{i_k^s}(x_{k-1}^s)-\nabla f_{i_k^s}(\widetilde{x}^{s-1})\big\|^2\Big]\\
=&\,\alpha\mathbb{E}\Big[\big\|[\nabla f_{i_k^s}(x_{k-1}^s)-\nabla f_{i_k^s}(\widetilde{x}^{s-1})]-[\nabla f(x_{k-1}^s)-\nabla f(\widetilde{x}^{s-1})]\big\|^2\Big]-\beta\mathbb{E}\Big[\big\|\nabla f_{i_k^s}(x_{k-1}^s)-\nabla f_{i_k^s}(\widetilde{x}^{s-1})\big\|^2\Big]\\
=&\,\alpha\mathbb{E}\Big[\big\|\nabla f_{i_k^s}(x_{k-1}^s)-\nabla f_{i_k^s}(\widetilde{x}^{s-1})\big\|^2\Big]-\alpha\big\|\nabla f(x_{k-1}^s)-\nabla f(\widetilde{x}^{s-1})\big\|^2-\beta\mathbb{E}\Big[\big\|\nabla f_{i_k^s}(x_{k-1}^s)-\nabla f_{i_k^s}(\widetilde{x}^{s-1})\big\|^2\Big]\\
\le&\,\alpha\mathbb{E}\Big[\big\|\nabla f_{i_k^s}(x_{k-1}^s)-\nabla f_{i_k^s}(\widetilde{x}^{s-1})\big\|^2\Big]-\beta\mathbb{E}\Big[\big\|\nabla f_{i_k^s}(x_{k-1}^s)-\nabla f_{i_k^s}(\widetilde{x}^{s-1})\big\|^2\Big]\\
=&\,(\alpha-\beta)\mathbb{E}\Big[\big\|\big[\nabla f_{i_k^s}(x_{k-1}^s)-\nabla f_{i_k^s}(x^*)\big]-\big[\nabla f_{i_k^s}(\widetilde{x}^{s-1})-\nabla f_{i_k^s}(x^*)\big]\big\|^2\Big]\\
\le&\,2(\alpha-\beta)\Big\{\mathbb{E}\Big[\big\|\nabla f_{i_k^s}(x_{k-1}^s)-\nabla f_{i_k^s}(x^*)\big\|^2\Big]+\mathbb{E}\Big[\big\|\nabla f_{i_k^s}(\widetilde{x}^{s-1})-\nabla f_{i_k^s}(x^*)\big\|^2\Big]\Big\}\\
\le&\,4L(\alpha-\beta)\big[F(x_{k-1}^s)-F(x^*)+F(\widetilde{x}^{s-1})-F(x^*)\big],
\end{aligned}$$

where the second equality holds due to the fact that $\mathbb{E}[\|x-\mathbb{E}x\|^2]=\mathbb{E}[\|x\|^2]-\|\mathbb{E}x\|^2$; the second inequality holds due to the fact that $\|a-b\|^2\le 2(\|a\|^2+\|b\|^2)$; and the last inequality follows from Lemma 3.4 in [9] (i.e., $\mathbb{E}[\|\nabla f_i(x)-\nabla f_i(x^*)\|^2]\le 2L[F(x)-F(x^*)]$). □

Moreover, we also introduce the following lemmas [3, 7], which are useful in our convergence analysis.

Lemma 2. *Let $\widetilde{F}(x,y)$ be the linear approximation of $F(\cdot)$ at y with respect to f, i.e.,*

$$\widetilde{F}(x,y)=f(y)+\langle\nabla f(y),\,x-y\rangle+r(x).$$

Then

$$F(x)\le\widetilde{F}(x,y)+\frac{L}{2}\|x-y\|^2\le F(x)+\frac{L}{2}\|x-y\|^2.$$

Lemma 3. *Assume that $\hat{x}$ is an optimal solution of the following problem,*

$$\min_{x\in\mathbb{R}^d}\frac{\tau}{2}\|x-y\|^2+g(x),$$

where $g(x)$ is a convex function (but possibly non-differentiable). Then the following inequality holds for all $x\in\mathbb{R}^d$:

$$g(\hat{x})+\frac{\tau}{2}\|\hat{x}-y\|^2+\frac{\tau}{2}\|x-\hat{x}\|^2\le g(x)+\frac{\tau}{2}\|x-y\|^2.$$

Proof of Theorem 1:

Proof. Let $\eta=\frac{1}{L\alpha}$ and $p_{i_k^s}=\widetilde{\nabla}f_{i_k^s}(x_{k-1}^s)=\nabla f_{i_k^s}(x_{k-1}^s)-\nabla f_{i_k^s}(\widetilde{x}^{s-1})+\nabla f(\widetilde{x}^{s-1})$. Using Lemma 2, we have

$$\begin{aligned}
F(y_k^s)&\le f(x_{k-1}^s)+\big\langle\nabla f(x_{k-1}^s),\,y_k^s-x_{k-1}^s\big\rangle+\frac{L\alpha}{2}\big\|y_k^s-x_{k-1}^s\big\|^2-\frac{L(\alpha-1)}{2}\big\|y_k^s-x_{k-1}^s\big\|^2+r(y_k^s)\\
&=f_{i_k^s}(x_{k-1}^s)+\big\langle p_{i_k^s},\,y_k^s-x_{k-1}^s\big\rangle+r(y_k^s)+\frac{L\alpha}{2}\|y_k^s-x_{k-1}^s\|^2\\
&\quad+\big\langle\nabla f(x_{k-1}^s)-p_{i_k^s},\,y_k^s-x_{k-1}^s\big\rangle-\frac{L(\alpha-1)}{2}\|y_k^s-x_{k-1}^s\|^2+f(x_{k-1}^s)-f_{i_k^s}(x_{k-1}^s).
\end{aligned}\tag{13}$$

Then

$$\begin{aligned}
&\big\langle\nabla f(x_{k-1}^s)-p_{i_k^s},\,y_k^s-x_{k-1}^s\big\rangle-\frac{L(\alpha-1)}{2}\|y_k^s-x_{k-1}^s\|^2\\
\le&\,\frac{1}{2L(\alpha-1)}\|\nabla f(x_{k-1}^s)-p_{i_k^s}\|^2+\frac{L(\alpha-1)}{2}\|y_k^s-x_{k-1}^s\|^2-\frac{L(\alpha-1)}{2}\|y_k^s-x_{k-1}^s\|^2\\
=&\,\frac{1}{2L(\alpha-1)}\|\nabla f(x_{k-1}^s)-p_{i_k^s}\|^2,
\end{aligned}\tag{14}$$

where the inequality follows from the Young's inequality, i.e., $a^Tb \leq \|a\|^2/(2\rho) + \rho\|b\|^2/2$ for any $\rho>0$. Substituting the inequality (14) into the inequality (13), we have

$$\begin{aligned}
F(y_k^s) &\leq f_{i_k^s}(x_{k-1}^s) + \langle p_{i_k^s}, y_k^s - x_{k-1}^s\rangle + r(y_k^s) + \frac{L\alpha}{2}\|y_k^s - x_{k-1}^s\|^2 \\
&\quad + \frac{1}{2L(\alpha-1)}\|\nabla f(x_{k-1}^s) - p_{i_k^s}\|^2 + f(x_{k-1}^s) - f_{i_k^s}(x_{k-1}^s) \\
&\leq f_{i_k^s}(x_{k-1}^s) + r(\widehat{w}_{k-1}^s) + \frac{L\alpha}{2}\big(\|\widehat{w}_{k-1}^s - x_{k-1}^s\|^2 - \|\widehat{w}_{k-1}^s - y_k^s\|^2\big) + \langle p_{i_k^s}, \widehat{w}_{k-1}^s - x_{k-1}^s\rangle \\
&\quad + \frac{1}{2L(\alpha-1)}\|\nabla f(x_{k-1}^s) - p_{i_k^s}\|^2 + f(x_{k-1}^s) - f_{i_k^s}(x_{k-1}^s) \\
&\leq F_{i_k^s}(\widehat{w}_{k-1}^s) + \frac{L\alpha}{2}\left(\|\widehat{w}_{k-1}^s - x_{k-1}^s\|^2 - \|\widehat{w}_{k-1}^s - y_k^s\|^2\right) + f(x_{k-1}^s) - f_{i_k^s}(x_{k-1}^s) \\
&\quad + \frac{1}{2L(\alpha-1)}\|\nabla f(x_{k-1}^s) - p_{i_k^s}\|^2 + \big\langle -\nabla f_{i_k^s}(\widetilde{x}^{s-1}) + \nabla f(\widetilde{x}^{s-1}), \widehat{w}_{k-1}^s - x_{k-1}^s\big\rangle \\
&\leq \sigma F_{i_k^s}(x^*) + (1-\sigma)F_{i_k^s}(\widehat{x}_{k-1}^s) + \frac{L\alpha\sigma^2}{2}\left(\|x^* - z_{k-1}^s\|^2 - \|x^* - z_k^s\|^2\right) \\
&\quad + \frac{1}{2L(\alpha-1)}\|\nabla f(x_{k-1}^s) - p_{i_k^s}\|^2 + f(x_{k-1}^s) - f_{i_k^s}(x_{k-1}^s) \\
&\quad + \big\langle \nabla f(\widetilde{x}^{s-1}) - \nabla f_{i_k^s}(\widetilde{x}^{s-1}), \widehat{w}_{k-1}^s - x_{k-1}^s\big\rangle,
\end{aligned} \tag{15}$$

where $\widehat{w}_{k-1}^s = \sigma x^* + (1-\sigma)\widehat{x}_{k-1}^s$, and $\widehat{x}_{k-1}^s = \theta_{k-1}x_{k-2}^s$. The second inequality follows from Lemma 3 with $g(x) := \langle p_{i_k^s}, x - x_{k-1}^s\rangle + r(x)$, $\tau = L\alpha$, $\hat{x} = y_k^s$, $x = \widehat{w}_{k-1}^s$ and $y = x_{k-1}^s$; the third inequality holds due to the convexity of the component function $f_{i_k^s}(x)$ (i.e., $f_{i_k^s}(x_{k-1}^s) + \langle \nabla f_{i_k^s}(x_{k-1}^s), \widehat{w}_{k-1}^s - x_{k-1}^s\rangle \leq f_{i_k^s}(\widehat{w}_{k-1}^s)$); and the last inequality holds due to the convexity of the function $F_{i_k^s}(x) := f_{i_k^s}(x) + r(x)$, and

$$z_{k-1}^s = [x_{k-1}^s - (1-\sigma)\widehat{x}_{k-1}^s]/\sigma, \;\; z_k^s = [y_k^s - (1-\sigma)\widehat{x}_{k-1}^s]/\sigma,$$

which mean that $\widehat{w}_{k-1}^s - x_{k-1}^s = \sigma(x^* - z_{k-1}^s)$ and $\widehat{w}_{k-1}^s - y_k^s = \sigma(x^* - z_k^s)$.

Using Property 1 with $\zeta = \frac{\delta\eta}{1-L\eta}$ and $\eta = 1/L\alpha$,[2] we obtain

$$\begin{aligned}
F(\theta_k x_{k-1}^s) = F(\widehat{x}_k^s) &\leq F(x_{k-1}^s) - \frac{(\theta_k - 1)^2}{2L(\alpha-1)}\|\nabla f_{i_k^s}(x_{k-1}^s) - \nabla f_{i_k^s}(\widetilde{x}^{s-1})\|^2 \\
&\leq F(x_{k-1}^s) - \frac{\beta_k}{2L(\alpha-1)}\|\nabla f_{i_k^s}(x_{k-1}^s) - \nabla f_{i_k^s}(\widetilde{x}^{s-1})\|^2,
\end{aligned} \tag{16}$$

where $\beta_k = \min\big[1/\alpha_k, (\theta_k - 1)^2\big]$, and α_k is defined below. Then there exists $\overline{\beta}_k$ such that

$$\mathbb{E}\left[\frac{\beta_k}{2L(\alpha-1)}\|\nabla f_{i_k^s}(x_{k-1}^s) - \nabla f_{i_k^s}(\widetilde{x}^{s-1})\|^2\right] = \frac{\overline{\beta}_k}{2L(\alpha-1)}\mathbb{E}\big[\|\nabla f_{i_k^s}(x_{k-1}^s) - \nabla f_{i_k^s}(\widetilde{x}^{s-1})\|^2\big], \tag{17}$$

where $\overline{\beta}_k = \mathbb{E}[\beta_k\|\nabla f_{i_k^s}(x_{k-1}^s) - \nabla f_{i_k^s}(\widetilde{x}^{s-1})\|^2]/\mathbb{E}[\|\nabla f_{i_k^s}(x_{k-1}^s) - \nabla f_{i_k^s}(\widetilde{x}^{s-1})\|^2]$, and $\overline{\beta}_k < (\alpha-1)/2$. Using the inequality (16), then we have

$$\begin{aligned}
\mathbb{E}[F(\widehat{x}_k^s) - F(x^*)] &\leq \mathbb{E}\left[F(x_{k-1}^s) - F(x^*) - \frac{\beta_k}{2L(\alpha-1)}\|\nabla f_{i_k^s}(x_{k-1}^s) - \nabla f_{i_k^s}(\widetilde{x}^{s-1})\|^2\right] \\
&= \mathbb{E}\big[F(x_{k-1}^s) - F(x^*)\big] - \frac{\overline{\beta}_k}{2L(\alpha-1)}\mathbb{E}\big[\|\nabla f_{i_k^s}(x_{k-1}^s) - \nabla f_{i_k^s}(\widetilde{x}^{s-1})\|^2\big].
\end{aligned} \tag{18}$$

[2]Note that our fast versions of SVRG-SD (i.e., SVRG-SD with randomly partial sufficient decrease) have the similar convergence properties as SVRG-SD because Property 1 still holds in the case when $\theta_k = 1$. That is, the main difference between their convergence properties is the different values of β_k, as shown below.

There must exist a constant $\alpha_k > 0$ such that $\mathbb{E}[F(y_k^s)-F(x^*)] = \alpha_k\mathbb{E}\big[F(x_{k-1}^s)-F(x^*)\big]$. Since $\mathbb{E}\big[f(x_{k-1}^s)-f_{i_k^s}(x_{k-1}^s)\big]=0$, $\mathbb{E}\big[\nabla f_{i_k^s}(\widetilde{x}^{s-1})\big]=\nabla f(\widetilde{x}^{s-1})$, $\mathbb{E}\big[F_{i_k^s}(x^*)\big]=F(x^*)$, and $\mathbb{E}\big[F_{i_k^s}(x_{k-1}^s)\big]=F(x_{k-1}^s)$, and taking the expectation of both sides of (15), we have

$$\begin{aligned}
&\alpha_k\mathbb{E}\big[F(x_{k-1}^s)-F(x^*)\big]-\frac{c_k\overline{\beta}_k}{2L(\alpha-1)}\mathbb{E}\big[\|\nabla f_{i_k^s}(x_{k-1}^s)-\nabla f_{i_k^s}(\widetilde{x}^{s-1})\|^2\big]\\
\leq&(1-\sigma)\mathbb{E}\big[F(\widehat{x}_{k-1}^s)-F(x^*)\big]+\frac{L\alpha\sigma^2}{2}\mathbb{E}\big[\|x^*-z_{k-1}^s\|^2-\|x^*-z_k^s\|^2\big]\\
&+\frac{1}{2L(\alpha-1)}\mathbb{E}\|\nabla f(x_{k-1}^s)-p_{i_k^s}\|^2-\frac{c_k\overline{\beta}_k}{2L(\alpha-1)}\mathbb{E}\big[\|\nabla f_{i_k^s}(x_{k-1}^s)-\nabla f_{i_k^s}(\widetilde{x}^{s-1})\|^2\big]\\
\leq&(1-\sigma)\mathbb{E}\big[F(\widehat{x}_{k-1}^s)-F(x^*)\big]+\frac{L\alpha\sigma^2}{2}\mathbb{E}\big[\|x^*-z_{k-1}^s\|^2-\|x^*-z_k^s\|^2\big]\\
&+\frac{2(1-c_k\overline{\beta}_k)}{\alpha-1}\left[F(x_{k-1}^s)-F(x^*)+F(\widetilde{x}^{s-1})-F(x^*)\right],
\end{aligned}\tag{19}$$

where the second inequality follows from Lemma 1 and Corollary 2. Here, $c_k=\alpha_k-[2(1-c_k\overline{\beta}_k)]/(\alpha-1)$, i.e.,

$$c_k=\frac{\alpha_k(\alpha-1)-2}{\alpha-1-2\overline{\beta}_k}.$$

Since $\frac{2}{\alpha-1}<\sigma$ with the suitable choices of α and σ, we have $c_k>\alpha_k-\frac{2}{\alpha-1}>1-\sigma$. Thus, (19) is rewritten as follows:

$$\begin{aligned}
&c_k\mathbb{E}\big[F(x_{k-1}^s)-F(x^*)\big]-\frac{c_k\overline{\beta}_k}{2L(\alpha-1)}\mathbb{E}\big[\|p_{i_k^s}-\nabla f_{i_k^s}(\widetilde{x}^{s-1})\|^2\big]\\
\leq&(1-\sigma)\mathbb{E}[F(\widehat{x}_{k-1}^s)-F(x^*)]+\frac{L\alpha\sigma^2}{2}\mathbb{E}\big[\|x^*-z_{k-1}^s\|^2-\|x^*-z_k^s\|^2\big]\\
&+\frac{2(1-c_k\overline{\beta}_k)}{\alpha-1}\mathbb{E}\big[F(\widetilde{x}^{s-1})-F(x^*)\big].
\end{aligned}\tag{20}$$

Combining the above two inequalities (18) and (20), we have

$$\begin{aligned}
&c_k\mathbb{E}[F(\widehat{x}_k^s)-F(x^*)]\\
\leq&(1-\sigma)\mathbb{E}\big[F(\widehat{x}_{k-1}^s)-F(x^*)\big]+\frac{L\alpha\sigma^2}{2}\mathbb{E}\big[\|x^*-z_{k-1}^s\|^2-\|x^*-z_k^s\|^2\big]\\
&+\frac{2(1-c_k\overline{\beta}_k)}{\alpha-1}\mathbb{E}\big[F(\widetilde{x}^{s-1})-F(x^*)\big].
\end{aligned}\tag{21}$$

Taking the expectation over the random choice of $i_1^s,i_2^s,\ldots,i_m^s$, summing up the above inequality over $k=1,\ldots,m$, and $\widehat{x}_0^s=\widetilde{x}^{s-1}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^m[c_k-(1-\sigma)]\left[F(\widehat{x}_k^s)-F(x^*)\right]\right]\\
\leq&(1-\sigma)\mathbb{E}\big[F(\widetilde{x}^{s-1})-F(x^*)\big]+\frac{L\alpha\sigma^2}{2}\mathbb{E}\big[\|x^*-z_0^s\|^2-\|x^*-z_m^s\|^2\big]\\
&+\mathbb{E}\left[\sum_{k=1}^m\frac{2(1-c_k\overline{\beta}_k)}{\alpha-1}[F(\widetilde{x}^{s-1})-F(x^*)]\right].
\end{aligned}\tag{22}$$

In addition, there exists $\widehat{\beta}^s$ for the s-th epoch such that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{m}\left[c_k-(1-\sigma)\right]\left[F(\widehat{x}_k^s)-F(x^*)\right]\right]\\
&=\mathbb{E}\left[\sum_{k=1}^{m}\left(\sigma-\frac{2}{\alpha-1}+\frac{2c_k\overline{\beta}_k}{\alpha-1}\right)\left[F(\widehat{x}_k^s)-F(x^*)\right]\right]\\
&=\left(\sigma-\frac{2}{\alpha-1}+\widehat{\beta}^s\right)\mathbb{E}\left[\sum_{k=1}^{m}[F(\widehat{x}_k^s)-F(x^*)]\right],
\end{aligned}\tag{23}$$

where

$$\widehat{\beta}^s=\frac{\mathbb{E}\Big[\sum_{k=1}^{m}\frac{2c_k\beta_k}{\alpha-1}[F(\widehat{x}_k^s)-F(x^*)]\Big]}{\mathbb{E}[\sum_{k=1}^{m}[F(\widehat{x}_k^s)-F(x^*)]]}.$$

Let $\widehat{\beta}=\min_{s=1,\ldots,S}\widehat{\beta}^s$. Using

$$\widetilde{x}^s=\frac{1}{m}\sum_{k=1}^{m}\widehat{x}_k^s,\ \ F(\widetilde{x}^s)\leq\frac{1}{m}\sum_{k=1}^{m}F(\widehat{x}_k^s),$$

(22) and (23), we have

$$\begin{aligned}
&\left(\sigma-\frac{2}{\alpha-1}+\widehat{\beta}\right)m\mathbb{E}[F(\widetilde{x}^s)-F(x^*)]\\
&\leq\left(1-\sigma+\frac{2m}{\alpha-1}\right)\mathbb{E}\big[F(\widetilde{x}^{s-1})-F(x^*)\big]\\
&\quad+\frac{L\alpha\sigma^2}{2}\mathbb{E}\big[\|x^*-z_0^s\|^2-\|x^*-z_m^s\|^2\big].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}[F(\widetilde{x}^s)-F(x^*)]\\
&\leq\left(\frac{1-\sigma}{\left(\sigma-\frac{2}{\alpha-1}+\widehat{\beta}\right)m}+\frac{2}{(\alpha-1)\left(\sigma-\frac{2}{\alpha-1}+\widehat{\beta}\right)}\right)\mathbb{E}\big[F(\widetilde{x}^{s-1})-F(x^*)\big]\\
&\quad+\frac{L\alpha\sigma^2}{2m\left(\sigma-\frac{2}{\alpha-1}+\widehat{\beta}\right)}\mathbb{E}\big[\|x^*-z_0^s\|^2-\|x^*-z_m^s\|^2\big].
\end{aligned}$$

This completes the proof. □

Appendix B: Proofs of Corollary 1

Proof. For μ-strongly convex problems, and let $x_0^s=\widehat{x}_0^s=\widetilde{x}^{s-1}$ and

$$z_0^s=\frac{x_0^s-(1-\sigma)\widehat{x}_0^s}{\sigma}=\widetilde{x}^{s-1}.$$

Due to the strong convexity of $F(\cdot)$, we have

$$\frac{\mu}{2}\|x^*-z_0^s\|^2=\frac{\mu}{2}\|x^*-\widetilde{x}^{s-1}\|^2\leq F(\widetilde{x}^{s-1})-F(x^*).\tag{24}$$

Using Theorem 1, we obtain

$$\begin{aligned}&\mathbb{E}[F(\widetilde{x}^s) - F(x^*)]\\ &\leq \left(\frac{1-\sigma}{m(\sigma-\frac{2}{\alpha-1}+\widehat{\beta})} + \frac{2}{(\alpha-1)\left(\sigma-\frac{2}{\alpha-1}+\widehat{\beta}\right)} + \frac{L\alpha\sigma^2}{m\mu\left(\sigma-\frac{2}{\alpha-1}+\widehat{\beta}\right)}\right)\mathbb{E}\big[F(\widetilde{x}^{s-1}) - F(x^*)\big].\end{aligned}$$

Replacing α and σ in the above inequality with 19 and 1/2, respectively, we have

$$\begin{aligned}&\mathbb{E}[F(\widetilde{x}^s) - F(x^*)]\\ &\leq \left(\frac{9}{(7+18\widehat{\beta})m} + \frac{2}{7+18\widehat{\beta}} + \frac{171L}{(14+36\widehat{\beta})m\mu}\right)\mathbb{E}\big[F(\widetilde{x}^{s-1}) - F(x^*)\big].\end{aligned}$$

This completes the proof. □

Appendix C: Proofs of Theorem 2

Proof. For non-strongly convex problems, and using Theorem 1 with $\alpha = 19$ and $\sigma = 1/2$, we have

$$\begin{aligned}\mathbb{E}[F(\widetilde{x}^s) - F(x^*)] \leq{}& \frac{171L}{(28+72\widehat{\beta})m}\mathbb{E}\Big[\|x^* - z_0^s\|^2 - \|x^* - z_m^s\|^2\Big]\\ &+ \left(\frac{9}{(7+18\widehat{\beta})m} + \frac{2}{7+18\widehat{\beta}}\right)\big[F(\widetilde{x}^{s-1}) - F(x^*)\big].\end{aligned} \tag{25}$$

According to the settings of Algorithm 1 for the non-strongly convex case, and let

$$x_0^s = \widehat{x}_0^s = [x_m^{s-1} - (1-\sigma)\widehat{x}_m^{s-1}]/\sigma,$$

then we have

$$z_0^s = \frac{x_0^s - (1-\sigma)\widehat{x}_0^s}{\sigma} = \frac{x_m^{s-1} - (1-\sigma)\widehat{x}_m^{s-1}}{\sigma},$$

and

$$z_m^{s-1} = \frac{x_m^{s-1} - (1-\sigma)\widehat{x}_m^{s-1}}{\sigma}.$$

Therefore, $z_0^s = z_m^{s-1}$.

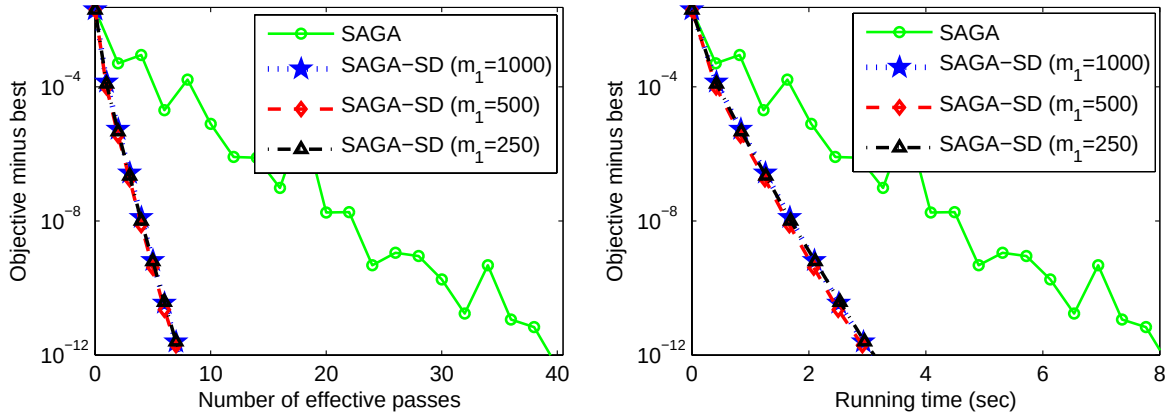
Using $z_0^0 = \widetilde{x}^0$, and summing up the inequality (25) over all $s = 1, \ldots, S$, then

$$\begin{aligned}\mathbb{E}\left[F\left(\frac{1}{S}\sum_{s=1}^{S}\widetilde{x}^s\right) - F(x^*)\right] \leq{}& \frac{171L}{(16+40\widehat{\beta})mS}\left\|x^* - \widetilde{x}^0\right\|^2\\ &+ \left(\frac{9}{(4+8\widehat{\beta})mS} + \frac{1}{(2+4\widehat{\beta})S}\right)\big[F(\widetilde{x}^0) - F(x^*)\big].\end{aligned}$$

Due to the settings of Algorithm 1 for the non-strongly convex case, we have

$$\begin{aligned}\mathbb{E}[F(\overline{x}) - F(x^*)] \leq{}& \frac{171L}{(16+40\widehat{\beta})mS}\left\|x^* - \widetilde{x}^0\right\|^2\\ &+ \left(\frac{9}{(4+8\widehat{\beta})mS} + \frac{1}{(2+4\widehat{\beta})S}\right)\big[F(\widetilde{x}^0) - F(x^*)\big].\end{aligned}$$

This completes the proof. □



(a) Objective gap vs. number of passes

(b) Objective gap vs. running time

Figure 1: Comparison of SAGA and SAGA-SD with different values of m_1 for solving ridge regression problems on the Covtype dataset.

Appendix D: Experiment Details

The C++ code of SVRG [5] was downloaded from `http://riejohnson.com/svrg_download.html`. The code of of SAGA [4] was downloaded from `http://www.aarondefazio.com/software.html`. For fair comparison, we implemented the proposed SVRG-SD and SAGA-SD algorithms, SAGA [4], Prox-SVRG [9], Catalyst [8] (which is based on SVRG and has the following three important parameters: α_k, κ, and the step size, η), and Katyusha [1] in C++ with a Matlab interface[3], and performed all the experiments on a PC with an Intel i5-2400 CPU and 16GB RAM.

Appendix E: More Experimental Results

Robustness

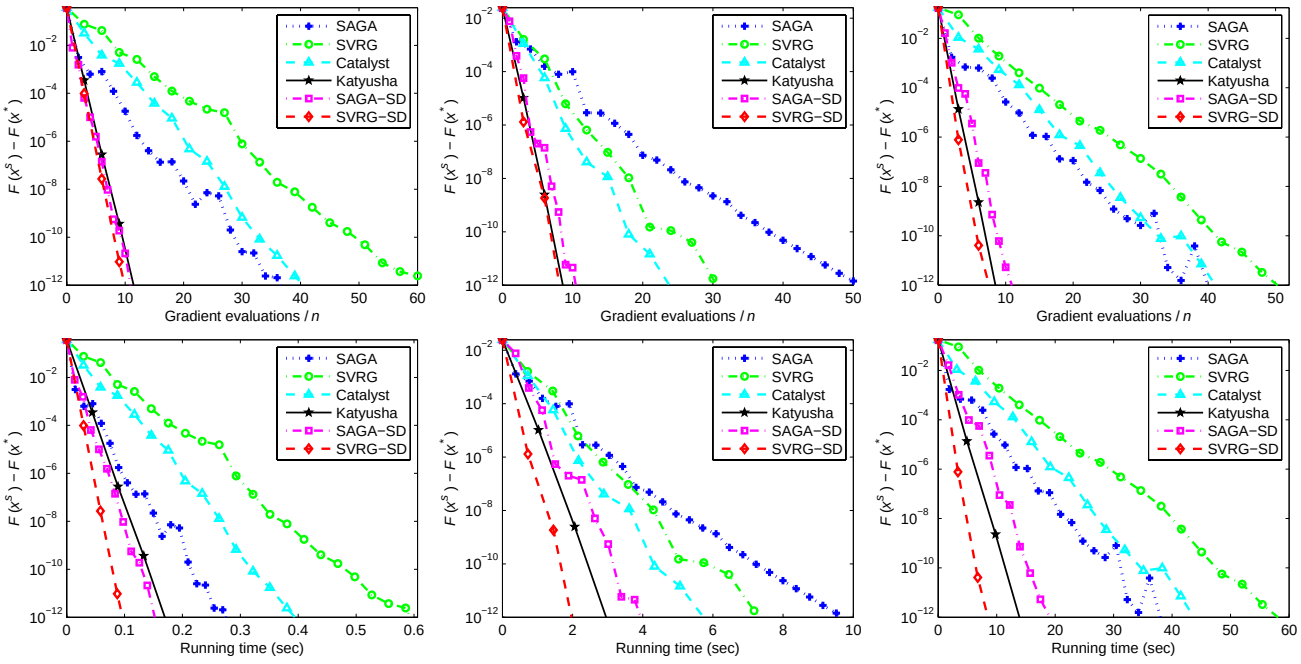
Figure 1 shows the performance of SAGA [4] and SAGA-SD with different values of m_1 for solving ridge regression problems on the Covtype data set, where the regularization parameter is $\lambda_1 = 10^{-4}$. From the result, we can observe that SAGA-SD significantly outperforms SAGA in terms of number of passes and running time. In particular, SAGA-SD, as well as SVRG-SD, has good robustness with respect to the number of iterations with sufficient decrease, which inspires us to use the partial sufficient decrease trick for both SVRG-SD and SAGA-SD.

Comparison of Results for Ridge Regression

In this part, we first report the experimental results of SVRG [5], SAGA [4], Catalyst [8], Katyusha [1], SVRG-SD and SAGA-SD for solving strongly convex (SC) ridge regression problems with the regularization parameter $\lambda_1 = 10^{-5}$ in Figure 2, where the horizontal axis denotes the number of effective passes over the data set (evaluating n component gradients, or computing a single full gradient is considered as one effective pass) or the running time (seconds). Moreover, we report the performance of all the stochastic variance reduction methods for solving ridge regression problems with relative small regularization parameters (e.g., $\lambda_1 = 10^{-7}$) in Figure 3, which shows that SVRG-SD and SAGA-SD, as well as Katyusha, converge significantly faster than SAGA, SVRG, and Catalyst. In particular, SVRG-SD and SAGA-SD usually outperform Katyusha in terms of both number of passes and running time, which further justifies the effectiveness of our sufficient decrease technique for stochastic optimization.

[3]The codes of some algorithms can be downloaded by the following anonymous link:

`https://www.dropbox.com/s/pyjeegseht77toh/Code_SVRG_SD.zip?dl=0.`



(a) Ijcnn1, $\lambda_1 = 10^{-5}$ (b) Covtype, $\lambda_1 = 10^{-5}$ (c) SUSY, $\lambda_1 = 10^{-5}$

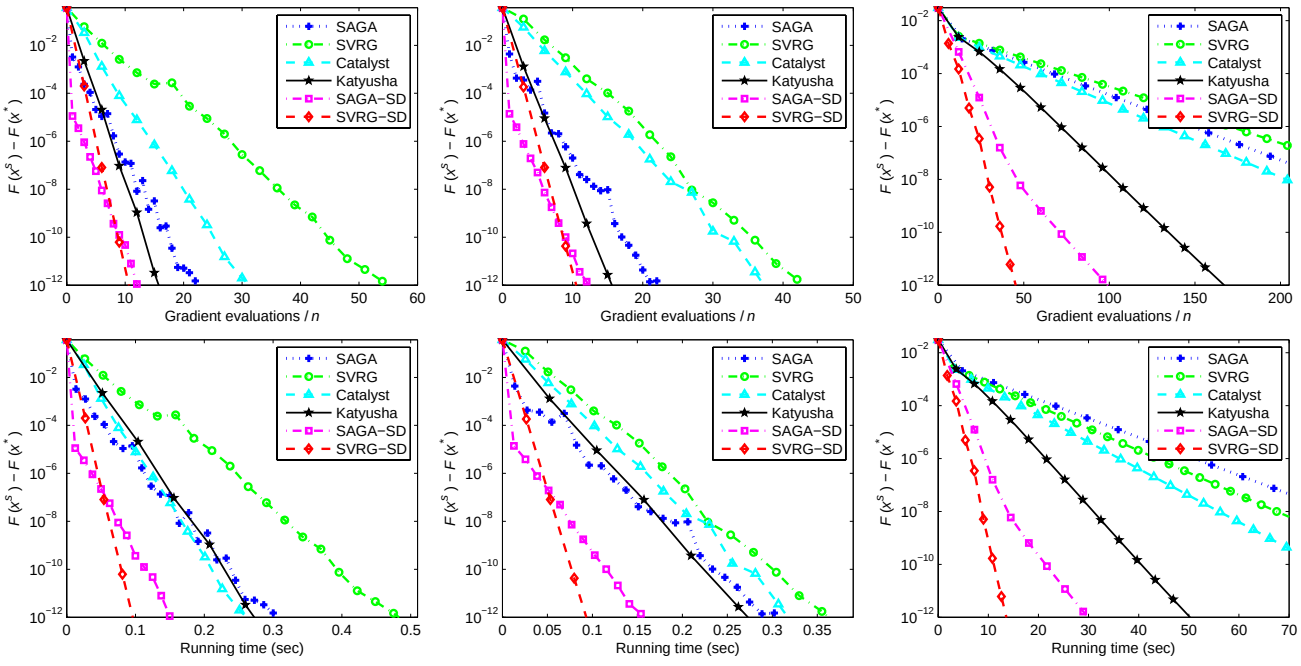
Figure 2: Comparison of all the stochastic variance reduced gradient methods for solving strongly convex ridge regression problems on the three dense data sets: Ijcnn1, Covtype and SUSY. The vertical axis is the objective value minus the minimum, and the horizontal axis denotes the number of effective passes over the data (top) or the running time (bottom).

Figure 4 shows the performance of all the methods for solving ridge regression problems with different regularization parameters on the sparse data set, Rcv1. From the results, we can observe that SVRG-SD and SAGA-SD significantly outperform their counterparts: SVRG and SAGA in terms of both number of effective passes and running time. The accelerated method, Catalyst, usually outperforms the non-accelerated methods, SVRG and SAGA. Katyusha converges much faster than SAGA, SVRG, and Catalyst for the cases when the regularization parameter is relatively small (e.g., $\lambda_1 = 10^{-5}$), whereas it sometime achieves similar or inferior performance when the regularization parameter is relatively large (e.g., $\lambda_1 = 10^{-3}$), as shown in Figures 4(a). Moreover, SVRG-SD and SAGA-SD achieve at least comparable performance with the accelerated stochastic method, Katyusha [1], in terms of number of effective passes. Since SVRG-SD and SAGA-SD have much lower per-iteration complexities than Katyusha, they have more obvious advantage over Katyusha in terms of running time.

Comparison of Results for Lasso and Elastic-Net

Finally, we report the performance of Prox-SVRG [9], SAGA [4], Catalyst [8], Katyusha [1], SVRG-SD and SAGA-SD for solving Lasso and elastic-net problems with different regularization parameters in Figures 5 and 6, respectively, from which we can observe that SVRG-SD and SAGA-SD also achieve much faster convergence speed than their counterparts: Prox-SVRG and SAGA, respectively. In particular, they also have comparable or better performance than the accelerated methods, Catalyst and Katyusha for both strongly convex and non-strongly convex problems. For the elastic-net problem, each component function $f_i(x)$ is defined as follows:

$$f_i(x) = \frac{1}{2}(a_i^T x - b_i)^2 + \frac{\lambda_1}{2}\|x\|^2.$$



(a) Ijcnn1, $\lambda_1 = 10^{-6}$ (b) Ijcnn1, $\lambda_1 = 10^{-7}$ (c) Covtype, $\lambda_1 = 10^{-7}$

Figure 3: Comparison of all the stochastic variance reduced gradient methods for solving strongly convex ridge regression problems with relatively small regularization parameters. The vertical axis is the objective value minus the minimum, and the horizontal axis denotes the number of effective passes over the data (top) or the running time (bottom).

References

[1] Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *STOC*, pages 1200–1205, 2017.

[2] Z. Allen-Zhu and Y. Yuan. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. *arXiv: 1506.01972v3*, 2016.

[3] L. Baldassarre and M. Pontil. Advanced topics in machine learning part II: 5. Proximal methods. *University Lecture*, 2013.

[4] 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.

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

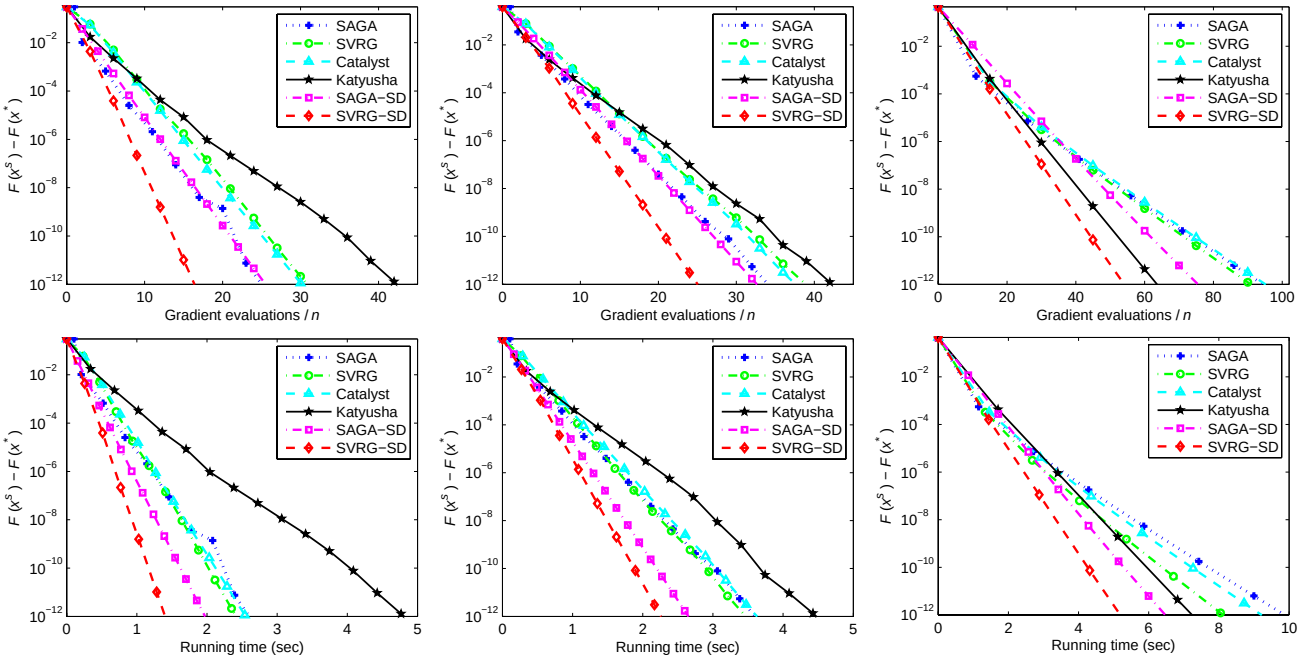
[6] Jakub Koneeny, Jie Liu, Peter Richtarik, , and Martin Takae. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE J. Sel. Top. Sign. Proces.*, 10(2):242–255, 2016.

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

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

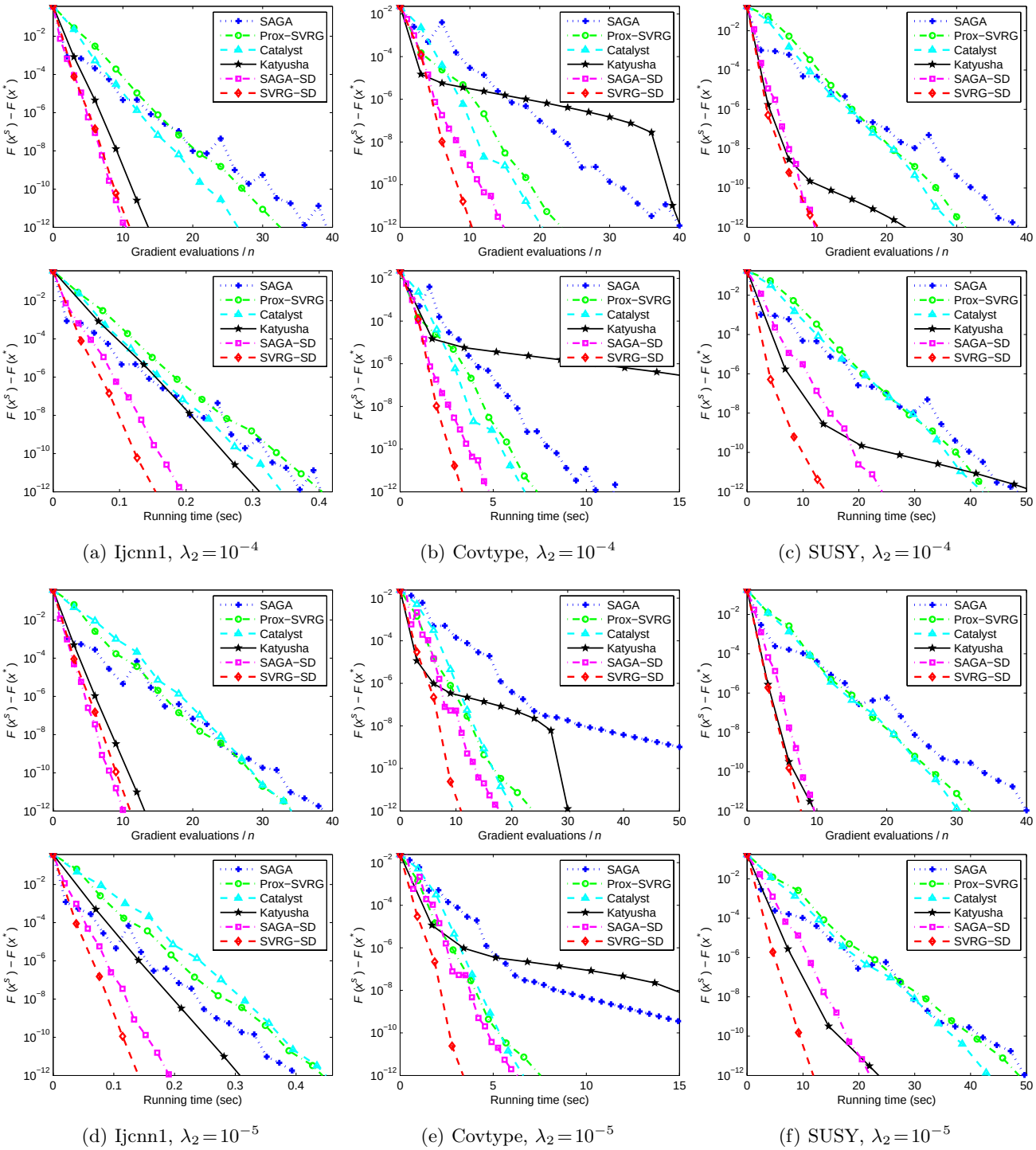
[9] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM J. Optim.*, 24(4):2057–2075, 2014.

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



(a) Rcv1, $\lambda_1 = 10^{-3}$ (b) Rcv1, $\lambda_1 = 10^{-4}$ (c) Rcv1, $\lambda_1 = 10^{-5}$

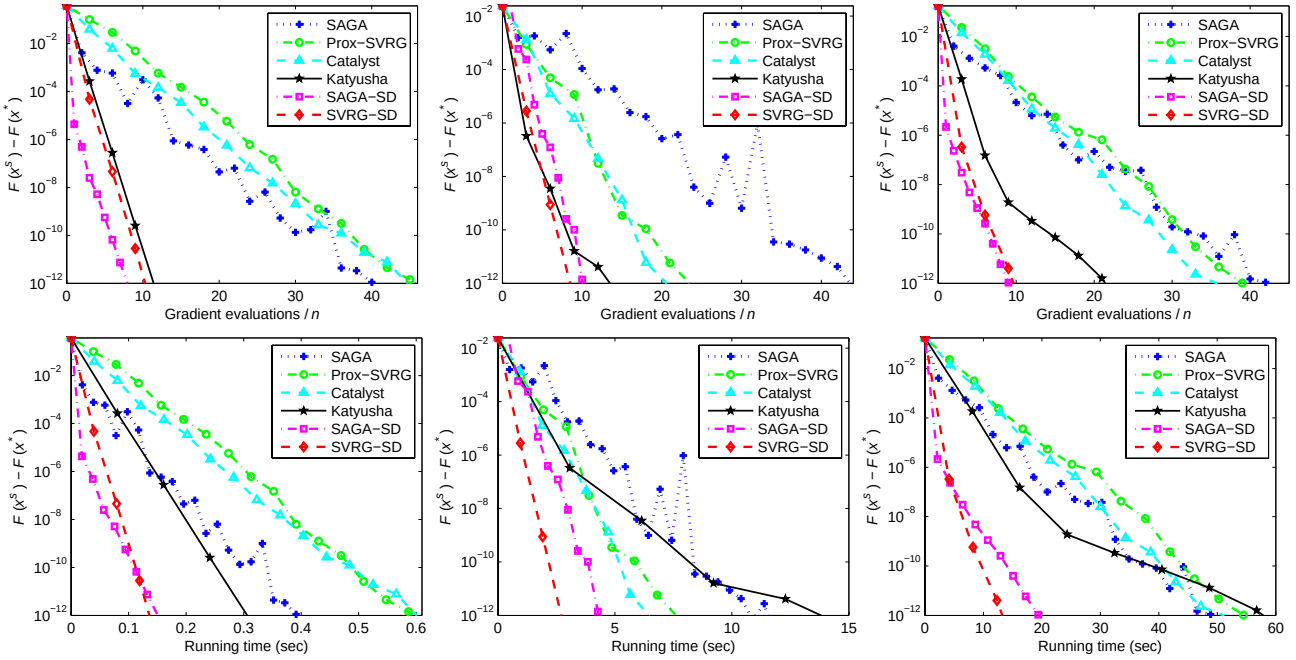
Figure 4: Comparison of all the stochastic variance reduced gradient methods for solving strongly convex ridge regression problems with different regularization parameters on the sparse data set, Rcv1. The vertical axis represents the objective value minus the minimum, and the horizontal axis denotes the number of effective passes (top) or the running time (bottom).



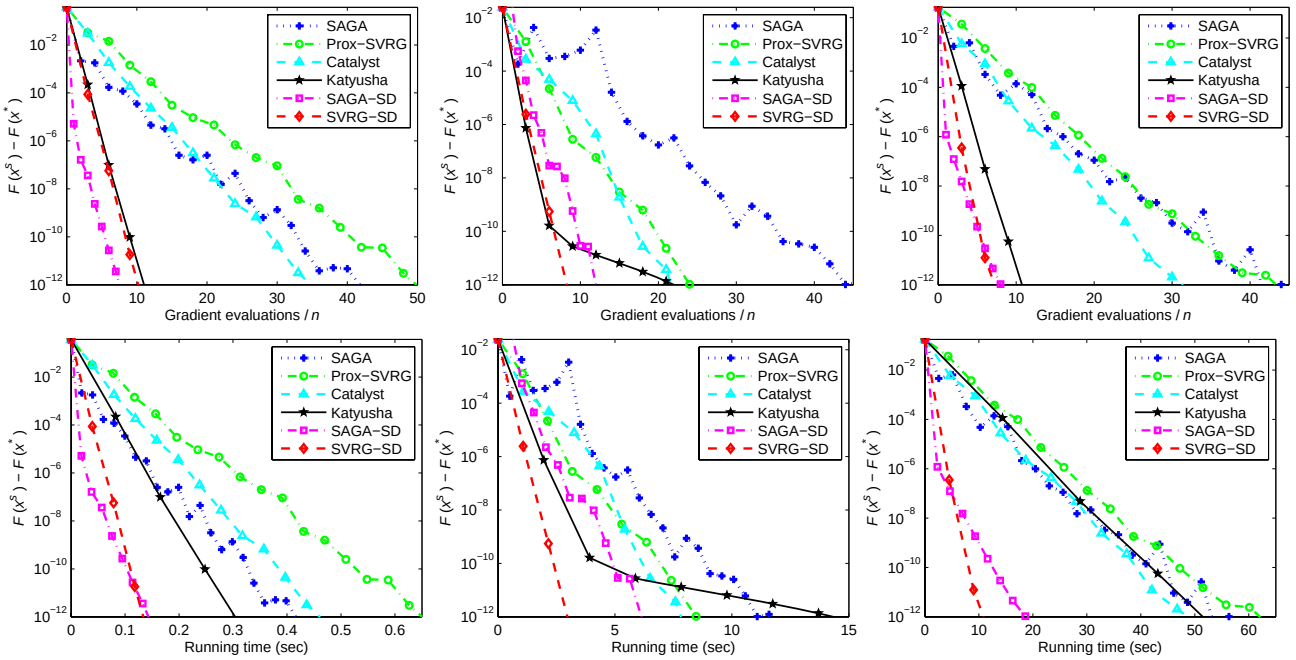
(a) Ijcnn1, $\lambda_2 = 10^{-4}$ (b) Covtype, $\lambda_2 = 10^{-4}$ (c) SUSY, $\lambda_2 = 10^{-4}$

(d) Ijcnn1, $\lambda_2 = 10^{-5}$ (e) Covtype, $\lambda_2 = 10^{-5}$ (f) SUSY, $\lambda_2 = 10^{-5}$

Figure 5: Comparison of all the stochastic variance reduced gradient methods for solving non-strongly convex Lasso problems on the three data sets. The vertical axis is the objective value minus the minimum, and the horizontal axis denotes the number of effective passes over the data (top) or the running time (seconds, bottom).



(a) $\lambda_1 = 10^{-5}$ and $\lambda_2 = 10^{-5}$



(b) $\lambda_1 = 10^{-5}$ and $\lambda_2 = 10^{-6}$

Figure 6: Comparison of all the stochastic methods for solving elastic-net (i.e., $(\lambda_1/2)\|\cdot\|^2+\lambda_2\|\cdot\|_1$) problems on Ijcnn1 (the first column), Covtype (the second column), and SUSY (the last column). The vertical axis is the objective value minus the minimum, and the horizontal axis denotes the number of effective passes over the data (top) or the running time (seconds, bottom).