Supplementary Material of "Less than a Single Pass: Stochastically Controlled Stochastic Gradient"

A Comparison to Existing Methods

A.1 Computation Complexity

Table 1 summarizes the characteristics of 11 existing popular algorithms as well as SCSG. The table includes the computation cost of optimizing non-strongly-convex functions (column 1) and strongly convex functions (column 2). Here strong convexity is only assumed on f instead of individual f_i. In practice, the amount of tuning is of major concern. For this reason, a fixed stepsize is usually preferred to a complicated stepsize scheme and it is better that the tuning parameter does not depend on unknown quantities; e.g., D_0. These issues are documented in column 3 and column 4. Moreover, many algorithms requires $\|\nabla f_i\|$ to be bounded or at least f_i to be Lipschitz. However, this assumption is not realistic in many cases and it is better to discard it. To address this issue, we document it in column 5. For all existing method, the complexity depends on ϵ through ϵ/D_0L, which is a scale-free quantity. For convenience, we denote it by ϵ'.

	General Convex	Strongly Convex	Constant η?	Depend on D_0?	f_i Lipschitz?
SCSG	$O\left((n \wedge \frac{G_n}{L\epsilon})\frac{1}{\epsilon'}\right)$	$O\left((n \wedge \frac{1}{\epsilon'} + \kappa)\log\frac{1}{\epsilon'}\right)$	Yes	No	No
SGD[4, 12][1]	$O\left(\frac{1}{\epsilon'^2}\right)$	$O\left(\frac{\kappa}{\epsilon'}\log\frac{1}{\epsilon'}\right)$	Yes/No	No/Yes[2]	Yes
SAGA[3][3]	$O\left(\frac{n}{\epsilon'}\right)$	$O\left((n+\kappa)\log\frac{1}{\epsilon'}\right)$	Yes	No	No
SVRG[6][4]	-	$O\left((n+\kappa)\log\frac{1}{\epsilon'}\right)$	Yes	No	No
APSDCA[13][5]	$O\left(n \wedge \sqrt{\frac{n}{\epsilon'}}\right)$	$O\left((n+\kappa)\log\frac{1}{\epsilon'}\right)$	Yes	No	No
APCG[9][6]	$O\left(\frac{n}{\sqrt{\epsilon'}}\right)$	$O\left(n\sqrt{\kappa}\log\frac{1}{\epsilon'}\right)$	Yes	No	No
SPDC[16][7]	-	$O\left((n+\sqrt{n\kappa})\log\frac{1}{\epsilon'}\right)$	Yes	No	No
Catalyst[8][8]	$O\left(\frac{n}{\sqrt{\epsilon'}}\right)$	$O\left((n+\sqrt{n\kappa})\log\frac{1}{\epsilon'}\right)$	No	No	No
SVRG++[2][9]	$O\left(n\log\frac{1}{\epsilon'} + \frac{1}{\epsilon'}\right)$	$O\left((n+\kappa^2)\log\frac{1}{\epsilon'} + \frac{1}{\epsilon'}\right)$	Yes	No	No
MSVRG[11][10]	$O\left(n + \frac{1}{\epsilon'^2}\right)$	-	No	Yes	Yes
AMSVRG[10][11]	$O\left((n \wedge \sqrt{\frac{n}{\epsilon'}})\log\frac{1}{\epsilon'}\right)$	$O\left((n+\kappa \wedge n\sqrt{\kappa})\log\frac{1}{\epsilon'}\right)$	Yes	No	No
Katyusha[1][12]	$O\left(n + \sqrt{\frac{n}{\epsilon'}}\right)$	$O\left((n+\sqrt{n\kappa})\log\frac{1}{\epsilon'}\right)$	No	No	No

Table 1: Comparison of the computation cost of SCSG and other algorithms. The third column indicates whether the algorithm uses a fixed stepsize η; the fourth column indicates whether the tuning parameter depends on D_0; the last column indicates whether f_i is required to be Lipschitz or (almost) equivalently $\|\nabla f_i\|$ is required to be bounded. ϵ' denotes ϵ/D_0L.

[1] Corollary 2.2 of [4] for the general convex case and Corollary 1 of [12] for the strongly convex case. The former is stated for $\mathbb{E}\|\nabla f(x_k)\|^2$ instead of $\mathbb{E}f(x_k) - f(x^*)$ to adapt to non-convex problems, but the latter has the same rate if f is convex. The latter is proved only for SVM but is potentially extended to more general cases.

[2] For the general convex case, the stepsize is set to be $D_0/\sqrt{T}$ for given number of total steps T, which is a constant but rely on D_0 and T; for the strongly convex case, the stepsize at step t is set to be proportional to $1/t$.

[3] Section 2 of [3] for both cases.

[4] No result for the general convex case and Theorem 1 of [6] for the strongly convex case.

A.2 Communication Complexity

Note that CoCoA has an additional factor H [5] determining the iteration complexity T and hence the tradeoff between computation and communication. We discuss the details in Appendix B.1). These methods are considered in the datacenter/workers model in which there are m worker machines with an (almost) equal number of data stored in each. In contrast to SCSG, instead of sub-sampling, these methods save on communication cost by performing updates locally. Specifically, the iterate is updated in each node in parallel and sent to the datacenter; then the datacenter sums or averages the updates and broadcasts the result to each node. As a consequence, under our computational model, the communication cost is the product of the number of nodes and the iteration complexity, namely $O(mT)$. Among these methods, DANE and DiSCO are second-order methods while SCSG and CoCoA are first-order methods. We record this in column 3 to emphasize the dependence on dimension. (Another first-order method which is similar to SVRG has been developed by [7]; we do not consider this method in our comparison due to the lack of theoretical results.) In addition, we show the computation cost in column 4. As seen from Table 2, SCSG is the only method whose communication cost is free of the number of nodes. This suggests that SCSG is more scalable in the distributed setting where a large number of worker machines exist; e.g., the mobile device system. Comparing SCSG with CoCoA, we find that with the same amount of computation, SCSG is more communication-efficient when $\epsilon \gg \frac{1}{n \wedge m^2}$, in which case the former has a cost $\frac{1}{\epsilon} \log \frac{1}{\epsilon}$ while the latter has a cost at least $m^2 \log \frac{1}{\epsilon}$ (see Appendix B.1 for details). Further, if the problem is well-conditioned in the sense that $\kappa \sim \frac{1}{\epsilon}$, then the communication cost of CoCoA is $m^2 n\epsilon \log \frac{1}{\epsilon}$ which could be much larger than that of SCSG once $\epsilon \gg \frac{1}{\sqrt{n}m}$. On the other hand, both DANE and DiSCO depend on the condition number in terms of communication and depend on sample size in terms of computation if $m \ll n$. We notice that the computation cost of DANE and DiSCO match that of SCSG only when $m = \Omega(n\epsilon)$, in which case the communication cost depends on the sample size. In contrast, this tradeoff does not appear in SCSG.

	General Convex	Strongly Convex	Dim. Dependence	Comp. Cost
SCSG	$O\left((n \wedge \frac{G_n}{L\epsilon}) \log \frac{1}{\epsilon'}\right)$	$O\left((n \wedge \frac{G_n}{L\epsilon}) \log \frac{1}{\epsilon'}\right)$	$O(d)$	$O\left((n \wedge \frac{G_n}{L\epsilon} + \kappa) \log \frac{1}{\epsilon'}\right)$
CoCoA[5][13]	-	$O\left(m^2 \cdot \frac{n+\kappa}{n \wedge \frac{G_n}{L\epsilon} + \kappa} \log \frac{1}{\epsilon'}\right)$	$O(d)$	$O\left((n \wedge \frac{G_n}{L\epsilon} + \kappa) \log \frac{1}{\epsilon'}\right)$
DANE[14][14]	-	$O\left(m\kappa \log \frac{1}{\epsilon'}\right)$	$O(d^2)$	$O\left(\frac{n\kappa}{m} \log \frac{1}{\epsilon'}\right)$
DiSCO[15][15]	-	$O\left(m\sqrt{\kappa} \log \kappa \log \frac{1}{\epsilon'}\right)$	$O(d^2)$	$O\left(\frac{n}{m}\sqrt{\kappa} \log \kappa \log \frac{1}{\epsilon'}\right)$

Table 2: Comparison of communication cost between SCSG and other algorithms. The third column shows the dimension dependence and the fourth column shows the additional assumptions required other than smoothness and strong convexity. ϵ' denotes $\epsilon / D_0 L$.

[5] Section 4.4 of [13] for Lasso and and Theorem 1 of [13] for the strongly convex case.

[6] Theorem 1 of [9] for both cases.

[7] No results for the general convex case and Section 1 of [16] for Empirical Risk Minimization.

[8] Table 1 of [8] for both cases

[9] Theorem 7.1 of [2] for the general convex case and Theorem 5.1 of [2] for the strongly convex case.

[10] Corollary 13 of [11] for the general convex case and no result for the strongly convex case.

[11] Theorem 2 of [10] for the general convex case and Theorem 3 of [10] for the strongly convex case.

[12] Theorem 5.1 of [1] for the general convex case and Theorem 3.1 for the strongly convex case.

[13] No result for the general convex case and Theorem 2 of [5] for the strongly convex case.

[14] No results for the general convex case (the results in [14] only hold for quadratic programming) and Theorem 4 of [14] for the strongly convex case.

[15] No results for the general convex case and Theorem 3 of [15] for the strongly convex case. For the latter we do not take the pre-conditioning step into consideration for fair comparison and hence $\mu = L$; see [15] for details.

B Technical Proofs

B.1 Lemmas

Proof [Lemma 1] Let $W_i = I(i \in \mathcal{I})$, then it is easy to see that

$$\mathbb{E}W_i^2 = \mathbb{E}W_i = \frac{B}{n}, \quad \mathbb{E}W_i W_{i'} = \frac{B(B-1)}{n(n-1)}. \tag{1}$$

Then g can be reformulated as

$$g = \frac{1}{B}\sum_{i=1}^{n} W_i \nabla f_i(x^*).$$

Since x^* is the optimum of f, we have

$$\mathbb{E}g = \frac{1}{B}\sum_{i=1}^{n} \mathbb{E}W_i \nabla f_i(x^*) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x^*) = \nabla f(x^*) = 0,$$

and

$$\begin{aligned}
\mathbb{E}\|g\|^2 &= \frac{1}{B^2}\left(\sum_{i=1}^{n} \mathbb{E}W_i^2 \|\nabla f_i(x^*)\|^2 + \sum_{i\neq i'} \mathbb{E}W_i W_{i'} \langle \nabla f_i(x^*), \nabla f_{i'}(x^*)\rangle\right) \\
&= \frac{1}{B^2}\left(\frac{B}{n}\sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2 + \frac{B(B-1)}{n(n-1)}\sum_{i\neq i'} \langle \nabla f_i(x^*), \nabla f_{i'}(x^*)\rangle\right) \\
&= \frac{1}{B^2}\left(\left(\frac{B}{n} - \frac{B(B-1)}{n(n-1)}\right)\sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2 + \frac{B(B-1)}{n(n-1)}\left\|\sum_{i=1}^{n} \nabla f_i(x^*)\right\|^2\right) \\
&= \frac{1}{B^2}\left(\frac{B}{n} - \frac{B(B-1)}{n(n-1)}\right)\sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2 \\
&= \frac{n-B}{(n-1)B}\cdot\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2 = \frac{(n-B)G_n}{(n-1)B}.
\end{aligned}$$

■

Lemma 3 *Under Assumption **A**1 and **A**2 with μ possibly equal to 0,*

$$\begin{aligned}
\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq\ & 2L(f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y\rangle) \\
& + 2\mu\left(f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x\rangle - \frac{L}{2}\|x - y\|^2\right).
\end{aligned}$$

The same bound holds if we interchange x and y on right-hand side. In particular, when $\mu = 0$,

$$\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2L\min\{f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y\rangle, f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x\rangle\}.$$

Proof [Lemma 3] By Lemma A1 of [3], for any $x, y \in \mathbb{R}^d$,

$$\begin{aligned}
f_i(x) - f_i(y) \geq\ & \langle \nabla f_i(y), x - y\rangle + \frac{1}{2(L-\mu)}\|\nabla f_i(x) - \nabla f_i(y)\|^2 + \frac{\mu L}{2(L-\mu)}\|x - y\|^2 \\
& - \frac{\mu}{L-\mu}\langle \nabla f_i(x) - \nabla f_i(y), x - y\rangle \\
=\ & \frac{L}{L-\mu}\langle \nabla f_i(y), x - y\rangle + \frac{1}{2(L-\mu)}\|\nabla f_i(x) - \nabla f_i(y)\|^2 + \frac{\mu L}{2(L-\mu)}\|x - y\|^2 \\
& - \frac{\mu}{L-\mu}\langle \nabla f_i(x), x - y\rangle,
\end{aligned}$$

which proves the lemma. ■

Proof [**Lemma 2**] We prove that for any x,

$$\frac{G_n}{L^2} \leq \frac{2}{L^2 n} \sum \|\nabla f_i(\tilde{x}_0)\|^2 + 4\|\tilde{x}_0 - x^*\|^2.$$

In fact, by Lemma 3,

$$f_i(x^*) - f_i(x) \geq \langle \nabla f_i(x), x^* - x \rangle + \frac{1}{2L}\|\nabla f_i(x^*) - \nabla f_i(x)\|^2.$$

Summing the above inequality for all i results in

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{1}{2L}\sum_{i=1}^{n} \|\nabla f_i(x^*) - \nabla f_i(x)\|^2. \tag{2}$$

Since x^* is a minimizer, we know that $f(x^*) \leq f(x)$ and thus

$$\frac{1}{2L}\sum_{i=1}^{n} \|\nabla f_i(x^*) - \nabla f_i(x)\|^2 \leq \langle \nabla f(x), x - x^* \rangle = \langle \nabla f(x) - \nabla f(x^*), x - x^* \rangle \leq L\|x - x^*\|^2.$$

On the other hand, note that for any $a, b \in \mathbb{R}^p$,

$$\|a - b\|^2 = \|a\|^2 + \|b\|^2 - 2\langle a, b \rangle = \frac{1}{2}\|a\|^2 - \|b\|^2 + \frac{1}{2}\|a - 2b\|^2 \geq \frac{1}{2}\|a\|^2 - \|b\|^2.$$

Thus,

$$\frac{1}{4L}\sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2 - \frac{1}{2L}\sum_{i=1}^{n} \|\nabla f_i(x)\|^2 \leq L\|x - x^*\|^2.$$

The first part of the lemma is then proved by setting $x = \tilde{x}_0$. For the second part, we exchange x and x^* in equation (2) and obtain that

$$f(x) \geq f(x^*) + \frac{1}{2L}\sum_{i=1}^{n} \|\nabla f_i(x^*) - \nabla f_i(x)\|^2 \geq B + \frac{1}{2L}\sum_{i=1}^{n} \|\nabla f_i(x^*) - \nabla f_i(x)\|^2.$$

Apply the same argument as above we prove the second inequality of the lemma.

■

Lemma 4 *Let $\mathcal{I} \in \{1, \cdots, n\}$ be a random subset with size B, i be a random element of $\mathcal{I}$ and*

$$\nu = \nabla f_i(x) - \nabla f_i(\tilde{x}) + \frac{1}{B}\sum_{i \in \mathcal{I}} \nabla f_i(\tilde{x});$$

*thus for any $\beta > 1$, under Assumption **A**1 and **A**2 with μ possibly equal to 0, it holds that*

$$\begin{aligned} E\|\nu\|^2 \leq\ & 2L(1+\beta)\{\langle \nabla f(x), x - x^* \rangle - (f(x) - f(x^*))\} + 2L(1+\beta)(f(\tilde{x}) - f(x^*)) \\ & + L(1 + (\beta - 1)^{-1})\frac{(n-B)}{(n-1)B} \cdot \frac{G_n}{L} + 2\mu(1+\beta)\left(f(x) - f(x^*) - \frac{L}{2}\|x - x^*\|^2\right). \end{aligned} \tag{3}$$

In particular when $\mu = 0$,

$$\begin{aligned} E\|\nu\|^2 \leq\ & 2L(1+\beta)\{\langle \nabla f(x), x - x^* \rangle - (f(x) - f(x^*))\} + 2L(1+\beta)(f(\tilde{x}) - f(x^*)) \\ & + L(1 + (\beta - 1)^{-1})\frac{(n-B)}{(n-1)B} \cdot \frac{G_n}{L}. \end{aligned} \tag{4}$$

Proof [**Lemma 4**] Notice that for any $\beta > 0$,

$$\|x + y\|^2 \leq (1 + \beta^{-1})\|x\|^2 + (1 + \beta)\|y\|^2$$

then for $\beta > 1$ we have

$$\mathbb{E}\|\nu\|^2 = \mathbb{E}\|(\nabla f_i(x) - \nabla f_i(x^*)) - (\nabla f_i(\tilde{x}) - \nabla f_i(x^*)) + \frac{1}{B}\sum_{i \in \mathcal{I}} \nabla f_i(\tilde{x})\|^2$$

$$
\begin{aligned}
&\leq (1+\beta)\mathbb{E}\|(\nabla f_i(x)-\nabla f_i(x^*))\|^2+(1+\beta^{-1})\mathbb{E}\|(\nabla f_i(\tilde{x})-\nabla f_i(x^*))+\frac{1}{B}\sum_{i\in\mathcal{I}}\nabla f_i(\tilde{x})\|^2\\
&=(1+\beta)\mathbb{E}\|\nabla f_i(x)-\nabla f_i(x^*)\|^2+(1+\beta^{-1})\mathbb{E}\|(\nabla f_i(\tilde{x})-\nabla f_i(x^*))\\
&\quad-\frac{1}{B}\sum_{i\in\mathcal{I}}(\nabla f_i(\tilde{x})-\nabla f_i(x^*))+\frac{1}{B}\sum_{i\in\mathcal{I}}\nabla f_i(x^*)\|^2\\
&\leq(1+\beta)\mathbb{E}\|\nabla f_i(x)-\nabla f_i(x^*)\|^2+(1+\beta)\mathbb{E}\|(\nabla f_i(\tilde{x})-\nabla f_i(x^*))-\frac{1}{B}\sum_{i\in\mathcal{I}}(\nabla f_i(\tilde{x})-\nabla f_i(x^*))\|^2\\
&\quad+(1+(\beta-1)^{-1})\mathbb{E}\|\frac{1}{B}\sum_{i\in\mathcal{I}}\nabla f_i(x^*)\|^2\\
&\leq(1+\beta)\mathbb{E}\|\nabla f_i(x)-\nabla f_i(x^*)\|^2+(1+\beta)\mathbb{E}\|\nabla f_i(\tilde{x})-\nabla f_i(x^*)\|^2\\
&\quad+(1+(\beta-1)^{-1})\mathbb{E}\|\frac{1}{B}\sum_{i\in\mathcal{I}}\nabla f_i(x^*)\|^2,
\end{aligned}
$$

where the last inequality uses the fact that

$$
\frac{1}{B}\sum_{i\in\mathcal{I}}(\nabla f_i(\tilde{x})-\nabla f_i(x^*))=E(\nabla f_i(\tilde{x})-\nabla f_i(x^*)|\mathcal{I})
$$

and the trivial inequality $E(X-E(X|\mathcal{I}))^2\leq EX^2$. For the first term, by Lemma 3 we obtain that

$$
\begin{aligned}
\mathbb{E}\|\nabla f_i(x)-\nabla f_i(x^*)\|^2\leq{}&2L\mathbb{E}(f_i(x^*)-f_i(x)-\langle\nabla f_i(x),x^*-x\rangle)\\
&+2\mu\mathbb{E}\left(f_i(x)-f_i(x^*)-\langle\nabla f_i(x^*),x-x^*\rangle-\frac{L}{2}\|x-x^*\|^2\right).
\end{aligned}
$$

Note that $\mathbb{E}f_i(x)=f(x)$ and $\mathbb{E}\nabla f_i(x)=\nabla f(x)$, the above expression can be simplified as

$$
\begin{aligned}
\mathbb{E}\|\nabla f_i(x)-\nabla f_i(x^*)\|^2\leq{}&2L\mathbb{E}(f(x^*)-f(x)-\langle\nabla f(x),x^*-x\rangle)\\
&+2\mu\mathbb{E}\left(f(x)-f(x^*)-\frac{L}{2}\|x-x^*\|^2\right)
\end{aligned}
$$

For the second term, we simply use Lemma 3 with $\mu=0$ and obtain that

$$
\mathbb{E}\|\nabla f_i(\tilde{x})-\nabla f_i(x^*)\|^2\leq 2L\mathbb{E}(f_i(\tilde{x})-f_i(x^*)-\langle\nabla f_i(x^*),\tilde{x}-x^*\rangle)=2L(f(\tilde{x})-f(x^*)).
$$

Combining the above results and Lemma 1, we conclude that

$$
\begin{aligned}
E\|\nu\|^2\leq{}&2L(1+\beta)\{\langle\nabla f(x),x-x^*\rangle-(f(x)-f(x^*))\}+2L(1+\beta)(f(\tilde{x})-f(x^*))\\
&+L(1+(\beta-1)^{-1})\frac{(n-B)}{(n-1)B}\cdot\frac{G_n}{L}+2\mu(1+\beta)\left(f(x)-f(x^*)-\frac{L}{2}\|x-x^*\|^2\right).
\end{aligned}
$$

■

B.2 Convergence Analysis of Non-Strongly Convex Case

Proof [**Theorem 1**] We prove Theorem 1 with

$$
C_1=\frac{1}{2\sqrt{1-\theta}(\sqrt{1-\theta}-\sqrt{\theta})},\quad C_2=\frac{1}{2(\sqrt{1-\theta}-\sqrt{\theta})^2}.
$$

It is easy to show $C_1,C_2\leq 2.5$ when $\theta<\frac{1}{5}$ via numerical calculation.

Now we state the main proof. In stage j, $x_0=\tilde{x}_{j-1}$, and we have $0\leq k\leq N_j$. In the following argument, we omit the subscript j for brevity.

$$
\begin{aligned}
\mathbb{E}\|x_{k+1}-x^*\|^2&=\mathbb{E}\|x_k-x^*-\eta\nu_k\|^2=\mathbb{E}\|x_k-x^*\|^2-2\eta\mathbb{E}\langle\nu_k,x_k-x^*\rangle+\eta^2\mathbb{E}\|\nu_k\|^2\\
&=\mathbb{E}\|x_k-x^*\|^2-2\eta\mathbb{E}\langle\nabla f(x_k),x_k-x^*\rangle+\eta^2\mathbb{E}\|\nu_k\|^2\\
&\leq\mathbb{E}\|x_k-x^*\|^2-2\eta\{1-L\eta(1+\beta)\}\mathbb{E}\langle\nabla f(x_k),x_k-x^*\rangle
\end{aligned}
$$

$$- 2L(1+\beta)\eta^2\mathbb{E}(f(x_k)-f(x^*)) + 2L(1+\beta)\eta^2\mathbb{E}(f(x_0)-f(x^*))$$
$$+ L(1+(\beta-1)^{-1})\eta^2\frac{(n-B)}{(n-1)B}\cdot\frac{G_n}{L},$$

where the last inequality uses Lemma 4. Here we restrict β such that

$$\beta < \frac{1}{\theta} - 1. \tag{5}$$

Noticing that (5) implies that $\eta < 1/L(1+\beta)$ and the convexity of f implies

$$\langle \nabla f(x_k), x_k - x^* \rangle \geq f(x_k) - f(x^*),$$

we have

$$\begin{aligned}2\eta\mathbb{E}(f(x_k)-f(x^*)) &\leq 2\eta\{1-L(1+\beta)\eta\}\mathbb{E}\langle\nabla f(x_k), x_k - x^*\rangle + 2L(1+\beta)\eta^2\mathbb{E}(f(x_k)-f(x^*)) \\ &\leq \mathbb{E}\|x_k - x^*\|^2 - \mathbb{E}\|x_{k+1}-x^*\|^2 + 2L(1+\beta)\eta^2\mathbb{E}(f(x_0)-f(x^*)) \\ &\quad + L(1+(\beta-1)^{-1})\eta^2\frac{(n-B)}{(n-1)B}\cdot\frac{G_n}{L} \end{aligned} \tag{6}$$

By definition of N (N_j),

$$\mathbb{E}(f(x_N)-f(x^*)) = \sum_{k\geq 1}\frac{\gamma^k}{A}\mathbb{E}(f(x_k)-f(x^*)),$$

and

$$\mathbb{E}\|x_N - x^*\|^2 = \sum_{k\geq 1}\frac{\gamma^k}{A}\mathbb{E}\|x_k - x^*\|^2,$$

where $A = \gamma(1-\gamma)^{-1}$ is the normalization factor. In order to be concise, let

$$\Delta(\eta,\beta,B) = L(1+(\beta-1)^{-1})\eta^2\frac{(n-B)}{(n-1)B}\cdot\frac{G_n}{L}. \tag{7}$$

Setting $k=0$ in (6),

$$\mathbb{E}\|x_1 - x^*\|^2 \leq \mathbb{E}\|x_0 - x^*\|^2 - 2\eta\{1-L(1+\beta)\eta\}\mathbb{E}(f(x_0)-f(x^*)) + \Delta(\eta,\beta,B) \leq \mathbb{E}\|x_0-x^*\|^2 + \Delta(\eta,\beta,B) \tag{8}$$

It then follows from (6) and (8) that

$$\begin{aligned}&2\eta\mathbb{E}(f(x_N)-f(x^*)) = \sum_{k\geq 1}\frac{\gamma^k}{A}2\eta\mathbb{E}(f(x_k)-f(x^*)) \\ \leq&\sum_{k\geq 1}\frac{\gamma^k}{A}(\mathbb{E}\|x_k-x^*\|^2 - \mathbb{E}\|x_{k+1}-x^*\|^2) + 2\eta^2L(1+\beta)\mathbb{E}(f(x_0)-f(x^*)) + \Delta(\eta,\beta,B) \\ =&\frac{\gamma\mathbb{E}\|x_1-x^*\|^2 - \sum_{k\geq 2}(\gamma^{k-1}-\gamma^k)\mathbb{E}\|x_k-x^*\|^2}{A} + 2\eta^2L(1+\beta)\mathbb{E}(f(x_0)-f(x^*)) + \Delta(\eta,\beta,B) \\ =&\frac{\mathbb{E}\|x_1-x^*\|^2 - \sum_{k\geq 1}(\gamma^{k-1}-\gamma^k)\mathbb{E}\|x_k-x^*\|^2}{A} + 2\eta^2L(1+\beta)\mathbb{E}(f(x_0)-f(x^*)) + \Delta(\eta,\beta,B) \\ \leq&\frac{\mathbb{E}\|x_0-x^*\|^2 - \sum_{k\geq 1}(\gamma^{k-1}-\gamma^k)\mathbb{E}\|x_k-x^*\|^2}{A} + 2\eta^2L(1+\beta)\mathbb{E}(f(x_0)-f(x^*)) + \left(1+\frac{1}{A}\right)\Delta(\eta,\beta,B) \\ =&\frac{\mathbb{E}\|x_0-x^*\|^2 - \mathbb{E}\|x_N-x^*\|^2}{A} + 2\eta^2L(1+\beta)\mathbb{E}(f(x_0)-f(x^*)) + \frac{1}{\gamma}\Delta(\eta,\beta,B) \\ =&\frac{1-\gamma}{\gamma}(\mathbb{E}\|x_0-x^*\|^2 - \mathbb{E}\|x_N-x^*\|^2) + 2\eta^2L(1+\beta)\mathbb{E}(f(x_0)-f(x^*)) + \frac{1}{\gamma}\Delta(\eta,\beta,B).\end{aligned}$$

This implies that

$$2\eta\mathbb{E}(f(\tilde{x}_j)-f(x^*)) \leq \frac{1-\gamma}{\gamma}(\mathbb{E}\|\tilde{x}_{j-1}-x^*\|^2 - \mathbb{E}\|\tilde{x}_j-x^*\|^2) + 2\eta^2L(1+\beta)\mathbb{E}(f(\tilde{x}_{j-1})-f(x^*)) + \frac{1}{\gamma}\Delta(\eta,\beta,B). \tag{9}$$

By convexity of f, we have

$$
\begin{aligned}
&2\eta\{1-\eta L(1+\beta)\}T\mathbb{E}(f(\bar{x}_T)-f(x^*)) \\
\leq&\sum_{j=1}^{T}\left\{2\eta\mathbb{E}(f(\tilde{x}_j)-f(x^*))-2\eta^2L(1+\beta)\mathbb{E}(f(\tilde{x}_{j-1})-f(x^*))\right\}+2\eta^2L(1+\beta)(f(\tilde{x}_0)-f(x^*))+\frac{T}{\gamma}\Delta(\eta,\beta,B) \\
\leq&\frac{1-\gamma}{\gamma}\|\tilde{x}_0-x^*\|^2+2\eta^2L(1+\beta)(f(\tilde{x}_0)-f(x^*))+\frac{T}{\gamma}\Delta(\eta,\beta,B) \\
\leq&\left(\frac{1}{\gamma}-1+\eta^2L^2(1+\beta)\right)\|\tilde{x}_0-x^*\|^2+\frac{T}{\gamma}\Delta(\eta,\beta,B) \\
\leq&\frac{1}{\gamma}\|\tilde{x}_0-x^*\|^2+\frac{T}{\gamma}\Delta(\eta,\beta,B),
\end{aligned}
$$

where the third inequality uses the smoothness of f, i.e.,

$$f(\tilde{x}_0)-f(x^*)\leq\frac{L}{2}\|\tilde{x}_0-x^*\|^2,$$

and the fourth inequality uses the fact that

$$\eta^2L^2(1+\beta)\leq\{\eta L(1+\beta)\}^2\leq 1.$$

Then

$$
\begin{aligned}
\mathbb{E}(f(\bar{x}_T)-f(x^*))&\leq\frac{1}{T}\cdot\frac{D_0}{2\eta\{1-\eta L(1+\beta)\}\gamma}+\frac{1}{2\eta\{1-\eta L(1+\beta)\}\gamma}\Delta(\eta,\beta,B) \\
&=\frac{1}{T}\cdot\frac{D_0}{2\eta\{1-\eta L(1+\beta)\}\gamma}+\frac{\beta}{\beta-1}\frac{\eta L}{2\{1-\eta L(1+\beta)\}\gamma}\frac{(n-B)}{(n-1)B}\cdot\frac{G_n}{L} \\
&=\frac{1}{T}\cdot\frac{LD_0}{2\theta\{1-\theta(1+\beta)\}\gamma}+\frac{\beta}{\beta-1}\frac{\theta}{2\{1-\theta(1+\beta)\}\gamma}\frac{(n-B)}{(n-1)B}\cdot\frac{G_n}{L} \\
&\triangleq J_1+J_2.
\end{aligned}
$$

We distinguish two cases:

1. If $B=n$, then $J_2=0$. Setting $\beta=1$ implies that
$$\mathbb{E}(f(\tilde{x}_T)-f(x^*))\leq\frac{1}{T}\cdot\frac{LD_0}{2\theta(1-2\theta)\gamma}.$$
2. If $B<n$, then optimizing J_2 over the set $\mathcal{B}=\{\beta:\beta>1,\theta(1+\beta)<1\}$ produces $\beta=\sqrt{(1-\theta)/\theta}$, in which case (5) is satisfied and
$$J_1=\frac{1}{T}\cdot\frac{LD_0}{2\theta\sqrt{1-\theta}(\sqrt{1-\theta}-\sqrt{\theta})\gamma}$$
and
$$J_2=\frac{\theta}{2(\sqrt{1-\theta}-\sqrt{\theta})^2\gamma}\frac{(n-B)}{(n-1)B}\cdot\frac{G_n}{L}.$$

■

Based on Theorem 1, we can obtain the following result.

Corollary 3 *Under the settings of Theorem 1, assume $\theta\in(0,\frac{1}{2})$ and*

$$B\geq\min\left\{n,\frac{2C_2\theta}{\gamma\epsilon}\cdot\frac{G_n}{L}\right\}\tag{10}$$

and

$$T\geq\frac{2C_1D_0}{\theta\gamma}\cdot\frac{L}{\epsilon}.\tag{11}$$

Then

$$\mathbb{E}(f(\bar{x}_T)-f(x^*))\leq\epsilon.$$

Proof [**Corollary 3**] It is easy to verify that under (20) and (21),

$$\frac{1}{T}\cdot\frac{C_1LD_0}{\theta\gamma}\leq\frac{\epsilon}{2}$$

and

$$C_2\frac{\theta}{\gamma}\frac{(n-B)}{(n-1)B}\cdot\frac{G_n}{L}\leq\frac{\epsilon}{2}.$$

Therefore,

$$\mathbb{E}(f(\bar{x}_T)-f(x^*))\leq\epsilon$$

. ■

As a direct consequence, we obtain the computation and communication complexity of SCSG for non-strongly convex case.

Proof [**Corollary 1**] It is known from Corollary 3 that

$$\mathbb{E}(f(\bar{x}_T)-f(x^*))\leq\epsilon.$$

The computation cost is

$$\mathbb{E}C_{\text{comp}}=\mathbb{E}\sum_{j=1}^{T}(B+N_j)=\mathbb{E}2BT=O\left(D_0\min\left\{\frac{nL}{\epsilon},\frac{G_n}{\epsilon^2}\right\}\right).$$

By Lemma 2, we know that

$$\frac{G_n}{L^2}\leq\frac{2}{L^2}\cdot\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\tilde{x}_0)\|^2+4D_0=O(1)$$

and as a result,

$$\mathbb{E}C_{\text{comp}}=O\left(\min\left\{n,\frac{G_n}{L^2}\cdot\frac{L}{\epsilon}\right\}\frac{D_0L}{\epsilon}\right).$$

Similarly the communication cost is

$$\mathbb{E}C_{\text{comm}}=\mathbb{E}BT=O\left(\min\left\{n,\frac{G_n}{L^2}\cdot\frac{L}{\epsilon}\right\}\frac{D_0L}{\epsilon}\right).$$

■

B.3 Convergence Analysis of Strongly Convex Case With Assumption A2

Similarly to the last section, we first establish a slightly more accurate version of Theorem 2 as follows.

Theorem 3 *Let $D_0=\|\tilde{x}_0-x^*\|^2$. If $\eta=\frac{\theta}{L+\mu}$ for some $\theta\in(0,\frac{1}{2})$, and one of the following assumptions hold:*

(i) $\gamma>1-\eta\mu$, $m\geq\log\left(\frac{1}{1-\gamma}\right)\Big/\log\left(\frac{\gamma}{1-\eta\mu}\right)$;

(ii) $\gamma=1-\eta\mu$, $m\geq\frac{1}{2L\mu\eta^2}=\frac{(\kappa+1)^2}{2\kappa\theta^2}$.

where $\kappa=L/\mu$ is the condition number. Then

1. *If $B=n$,*

$$\mathbb{E}(f(\tilde{x}_T)-f(x^*))\leq\left(\frac{2L\eta}{1-2\mu\eta}\right)^T\cdot\frac{5LD_0}{\theta}=\left(1-\frac{(L+\mu)(1-2\theta)}{L+\mu-2\mu\theta}\right)^T\cdot\frac{5LD_0}{\theta}$$

2. *If* $B < n$,

$$\mathbb{E}(f(\tilde{x}_T)-f(x^*)) \leq \frac{5LD_0}{\theta}\cdot\left(1-\frac{(L+\mu)(1-\theta-\sqrt{\theta(1-\theta)})}{L+\mu-\mu(\theta+\sqrt{\theta(1-\theta)})}\right)^T+\frac{4\theta}{(\sqrt{1-\theta}-\sqrt{\theta})^2}\cdot\frac{(n-B)}{(n-1)B}\cdot\frac{G_n}{L}.$$

Remark 1 *Note that when* $\theta < \frac{1}{2}$,

$$\frac{(L+\mu)(1-2\theta)}{L+\mu-2\mu\theta} \geq 1-2\theta,$$

and thus Part 1 can be simplified by a slightly weaker bound as in Theorem 2:

$$(2\theta)^T \cdot \frac{5LD_0}{\theta} = 10LD_0 \cdot (2\theta)^{T-1}.$$

Similarly,

$$1-\frac{(L+\mu)(1-\theta-\sqrt{\theta(1-\theta)})}{L+\mu-\mu(\theta+\sqrt{\theta(1-\theta)})} \leq 1-\theta-\sqrt{\theta(1-\theta)} = \frac{L\sqrt{\theta}(\sqrt{\theta}+\sqrt{1-\theta})}{L+\mu\sqrt{1-\theta}(\sqrt{1-\theta}-\sqrt{\theta})} \leq \sqrt{2\theta},$$

and thus Part 2 can be simplified by a slightly weaker bound as in Theorem 2

$$(\sqrt{2\theta})^T \cdot \frac{5LD_0}{\theta} = 10LD_0 \cdot (2\theta)^{\frac{T}{2}-1}.$$

Furthermore, Part 2 of the Theorem implies that the constant C_3 *in Theorem 2 is*

$$C_3 = \frac{4}{(\sqrt{1-\theta}-\sqrt{\theta})^2},$$

and it is easy to see that $C_3 \leq 20$ *if* $\theta < \frac{1}{5}$.

Proof [**Theorem 3**] Similar to the proof of Theorem 1, if β satisfies (5), then $1-L\eta(1+\beta) > 0$ and

$$\begin{aligned}
\mathbb{E}\|x_{k+1}-x^*\|^2 &\leq \{1-\eta^2\mu L(1+\beta)\}\mathbb{E}\|x_k-x^*\|^2 - 2\eta\{1-L\eta(1+\beta)\}\mathbb{E}\langle\nabla f(x_k), x_k-x^*\rangle \\
&\quad - 2(L-\mu)(1+\beta)\eta^2\mathbb{E}(f(x_k)-f(x^*)) + 2L(1+\beta)\eta^2\mathbb{E}(f(x_0)-f(x^*)) \\
&\quad + L(1+(\beta-1)^{-1})\eta^2\frac{(n-B)}{(n-1)B}\cdot\frac{G_n}{L} \qquad (12)\\
&\leq (1-\mu\eta)\mathbb{E}\|x_k-x^*\|^2 - 2\eta\{1-\mu\eta(1+\beta)\}\mathbb{E}(f(x_k)-f(x^*)) \\
&\quad + 2L(1+\beta)\eta^2\mathbb{E}(f(x_0)-f(x^*)) + L(1+(\beta-1)^{-1})\eta^2\frac{(n-B)}{(n-1)B}\cdot\frac{G_n}{L}, \qquad (13)
\end{aligned}$$

where the last inequality is from the strong convexity of f, i.e.

$$\langle\nabla f(x_k), x_k-x^*\rangle \geq (f(x_k)-f(x^*)) + \frac{\mu}{2}\|x_k-x^*\|^2.$$

Setting $k=0$, we have

$$\begin{aligned}
\mathbb{E}\|x_1-x^*\|^2 &\leq (1-\mu\eta)\mathbb{E}\|x_0-x^*\|^2 - 2\eta\left(1-(L+\mu)\eta(1+\beta)\right)\mathbb{E}(f(x_0)-f(x^*)) + \Delta(\eta,\beta,B) \\
&\leq (1-\mu\eta)\mathbb{E}\|x_0-x^*\|^2 + \Delta(\eta,\beta,B), \qquad (14)
\end{aligned}$$

where $\Delta(\eta,\beta,B)$ is defined by (7). By definition of N (the stage index j is omitted for brevity),

$$\mathbb{E}(f(x_N)-f(x^*)) = \frac{1}{A(m,\gamma)}\sum_{k=1}^{m}\frac{\gamma^k}{(1-\mu\eta)^k}\mathbb{E}(f(x_k)-f(x^*)),$$

and

$$\mathbb{E}\|x_N-x^*\|^2 = \frac{1}{A(m,\gamma)}\sum_{k=1}^{m}\frac{\gamma^k}{(1-\mu\eta)^k}\mathbb{E}\|x_k-x^*\|^2,$$

where $A(m,\gamma)$ is the normalization factor such that

$$A(m,\gamma)=\sum_{k=1}^{m}\frac{\gamma^k}{(1-\mu\eta)^k}.$$

It then follows from (13) and (14) that

$$\begin{aligned}
&2\eta\{1-\mu\eta(1+\beta)\}\mathbb{E}(f(x_N)-f(x^*))=\frac{1}{A(m,\gamma)}\sum_{k=1}^{m}\frac{\gamma^k}{(1-\mu\eta)^k}2\eta\{1-\mu\eta(1+\beta)\}\mathbb{E}(f(x_k)-f(x^*))\\
\leq&\frac{1}{A(m,\gamma)}\sum_{k=1}^{m}\gamma^k\left(\frac{\mathbb{E}\|x_k-x^*\|^2}{(1-\mu\eta)^{k-1}}-\frac{\mathbb{E}\|x_{k+1}-x^*\|^2}{(1-\mu\eta)^k}\right)+2\eta^2L(1+\beta)\mathbb{E}(f(x_0)-f(x^*))+\Delta(\eta,\beta,B)\\
=&\frac{1}{A(m,\gamma)}\left\{\gamma\mathbb{E}\|x_1-x^*\|^2-\sum_{k=2}^{m}\frac{\gamma^{k-1}-\gamma^k}{(1-\mu\eta)^{k-1}}\mathbb{E}\|x_k-x^*\|^2\right\}+2\eta^2L(1+\beta)\mathbb{E}(f(x_0)-f(x^*))+\Delta(\eta,\beta,B)\\
=&\frac{1}{A(m,\gamma)}\left\{\mathbb{E}\|x_1-x^*\|^2-\sum_{k=1}^{m}\frac{\gamma^{k-1}-\gamma^k}{(1-\mu\eta)^{k-1}}\mathbb{E}\|x_k-x^*\|^2\right\}+2\eta^2L(1+\beta)\mathbb{E}(f(x_0)-f(x^*))+\Delta(\eta,\beta,B)\\
=&\frac{1}{A(m,\gamma)}\left\{\mathbb{E}\|x_1-x^*\|^2-\frac{A(m,\gamma)(1-\mu\eta)(1-\gamma)}{\gamma}\mathbb{E}\|x_N-x^*\|^2\right\}+2\eta^2L(1+\beta)\mathbb{E}(f(x_0)-f(x^*))+\Delta(\eta,\beta,B)\\
\leq&\frac{1-\mu\eta}{A(m,\gamma)}\left\{\mathbb{E}\|x_0-x^*\|^2-\frac{A(m,\gamma)(1-\gamma)}{\gamma}\mathbb{E}\|x_N-x^*\|^2\right\}\\
&+\left(1+\frac{1}{A(m,\gamma)}\right)\Delta(\eta,\beta,B)+2\eta^2L(1+\beta)\mathbb{E}(f(x_0)-f(x^*)).
\end{aligned}$$

The above inequality can be rewritten as

$$\begin{aligned}
&2\eta\{1-\mu\eta(1+\beta)\}\mathbb{E}(f(x_N)-f(x^*))+\frac{(1-\mu\eta)(1-\gamma)}{\gamma}\mathbb{E}\|x_N-x^*\|^2 &(15)\\
\leq&2\eta^2L(1+\beta)\mathbb{E}(f(x_0)-f(x^*))+\frac{1-\mu\eta}{A(m,\gamma)}\mathbb{E}\|x_0-x^*\|^2+\left(1+\frac{1}{A(m,\gamma)}\right)\Delta(\eta,\beta,B). &(16)
\end{aligned}$$

Define λ as

$$\lambda=\max\left\{\frac{L\eta(1+\beta)}{1-\mu\eta(1+\beta)},\frac{\gamma}{(1-\gamma)A(m,\gamma)}\right\} \tag{17}$$

then

$$\begin{aligned}
&2\eta\{1-\mu\eta(1+\beta)\}\mathbb{E}(f(\tilde{x}_j)-f(x^*))+\frac{(1-\mu\eta)(1-\gamma)}{\gamma}\mathbb{E}\|\tilde{x}_j-x^*\|^2-\frac{1}{1-\lambda}\left(1+\frac{1}{A(m,\gamma)}\right)\Delta(\eta,\beta,B)\\
=&2\eta\{1-\mu\eta(1+\beta)\}\mathbb{E}(f(x_N)-f(x^*))+\frac{(1-\mu\eta)(1-\gamma)}{\gamma}\mathbb{E}\|x_N-x^*\|^2-\frac{1}{1-\lambda}\left(1+\frac{1}{A(m,\gamma)}\right)\Delta(\eta,\beta,B)\\
\leq&\lambda\left[2\eta\{1-\mu\eta(1+\beta)\}\mathbb{E}(f(x_0)-f(x^*))+\frac{(1-\mu\eta)(1-\gamma)}{\gamma}\mathbb{E}\|x_0-x^*\|^2-\frac{1}{1-\lambda}\left(1+\frac{1}{A(m,\gamma)}\right)\Delta(\eta,\beta,B)\right]\\
=&\lambda\left[2\eta\{1-\mu\eta(1+\beta)\}\mathbb{E}(f(\tilde{x}_{j-1})-f(x^*))+\frac{(1-\mu\eta)(1-\gamma)}{\gamma}\mathbb{E}\|\tilde{x}_{j-1}-x^*\|^2-\frac{1}{1-\lambda}\left(1+\frac{1}{A(m,\gamma)}\right)\Delta(\eta,\beta,B)\right].
\end{aligned}$$

This implies that

$$\begin{aligned}
&2\eta\{1-\mu\eta(1+\beta)\}\mathbb{E}(f(\tilde{x}_T)-f(x^*))+\frac{(1-\mu\eta)(1-\gamma)}{\gamma}\mathbb{E}\|\tilde{x}_j-x^*\|^2\\
\leq&\lambda^T\left[2\eta\{1-\mu\eta(1+\beta)\}\mathbb{E}(f(\tilde{x}_0)-f(x^*))+\frac{(1-\mu\eta)(1-\gamma)}{\gamma}\mathbb{E}\|\tilde{x}_0-x^*\|^2\right]+\frac{1}{1-\lambda}\left(1+\frac{1}{A(m,\gamma)}\right)\Delta(\eta,\beta,B)\\
\leq&\lambda^T\left[2\eta\{1-\mu\eta(1+\beta)\}\mathbb{E}(f(\tilde{x}_0)-f(x^*))+\frac{(1-\mu\eta)(1-\gamma)}{\gamma}\mathbb{E}\|\tilde{x}_0-x^*\|^2\right]+\frac{2}{1-\lambda}\Delta(\eta,\beta,B),
\end{aligned}$$

where the last inequality uses the fact that $A(m,\gamma) \geq \gamma/(1-\mu\eta) \geq 1$ and hence

$$\begin{aligned}\mathbb{E}(f(\tilde{x}_T) - f(x^*)) &\leq \lambda^T \left[\mathbb{E}(f(\tilde{x}_0) - f(x^*)) + \frac{(1-\mu\eta)(1-\gamma)}{2\eta\gamma\{1-\mu\eta(1+\beta)\}}\mathbb{E}\|\tilde{x}_0 - x^*\|^2\right] \\ &\quad + \frac{1}{\eta(1-\lambda)\{1-\mu\eta(1+\beta)\}}\Delta(\eta,\beta,B) \\ &\triangleq \lambda^T J_1 + J_2. \end{aligned} \tag{18}$$

Now we prove that under both conditions (i) and (ii),

$$\frac{L\eta(1+\beta)}{1-\mu\eta(1+\beta)} \geq \frac{\gamma}{(1-\gamma)A(m,\gamma)}.$$

(i) Since

$$m \geq \log\left(\frac{1}{1-\gamma}\right) \Big/ \log\left(\frac{\gamma}{1-\mu\eta}\right)$$

then it can be verified that

$$m \geq \log\left(\frac{\gamma-(1-\mu\eta)}{1-\gamma}\frac{1-\mu\eta(1+\beta)}{L\eta(1+\beta)} + 1\right) \Big/ \log\left(\frac{\gamma}{1-\mu\eta}\right) \tag{19}$$

by noticing that

$$\frac{\gamma-(1-\mu\eta)}{1-\gamma}\frac{1-\mu\eta(1+\beta)}{L\eta(1+\beta)} + 1 \leq \left(\frac{\gamma-(1-\mu\eta)}{1-\gamma} + 1\right)\frac{1-\mu\eta(1+\beta)}{L\eta(1+\beta)} = \frac{\mu}{L}\frac{1-\mu\eta(1+\beta)}{(1+\beta)(1-\gamma)} \leq \frac{1}{2(1-\gamma)},$$

where the last inequality uses the fact that $\beta \geq 1$. Then

$$\frac{\gamma}{(1-\gamma)A(m,\gamma)} = \frac{\gamma}{(1-\gamma)\sum_{k=1}^m \left(\frac{\gamma}{1-\eta\mu}\right)^k} = \frac{1-\eta\mu}{1-\gamma}\cdot\frac{\frac{\gamma}{1-\eta\mu}-1}{\left(\frac{\gamma}{1-\eta\mu}\right)^m - 1} \geq \frac{L\eta(1+\beta)}{1-\mu\eta(1+\beta)}.$$

(ii)

$$m \geq \frac{1}{2L\mu\eta^2} \geq \frac{(1-2\mu\eta)(1-\mu\eta)}{2L\mu\eta^2}$$

then

$$\frac{\gamma}{(1-\gamma)A(m,\gamma)} \leq \frac{1-\eta\mu}{m(1-\gamma)} = \frac{1-\eta\mu}{m\eta\mu} \leq \frac{2L\eta}{1-2\mu\eta} \leq \frac{L\eta(1+\beta)}{1-\mu\eta(1+\beta)}.$$

Therefore, in both cases,

$$\lambda = \frac{L\eta(1+\beta)}{1-\mu\eta(1+\beta)}$$

Now we distinguish two cases:

1. If $B = n$, then $J_2 = 0$. Set $\beta = 1$ then

$$\lambda = \frac{2L\eta}{1-2\mu\eta}$$

and

$$J_1 \leq \frac{L}{2}\left(1 + \frac{(1-\mu\eta)(1-\gamma)}{L\eta\gamma(1-2\mu\eta)}\right)D_0.$$

Note that $\mu \leq (\mu+L)/2 \leq L$ and $\gamma \geq 1-\mu\eta$,

$$J_1 \leq \frac{L}{2}\left(1 + \frac{2}{\theta(1-\theta)}\right)D_0 \leq \frac{5LD_0}{\theta}.$$

Therefore,

$$\mathbb{E}(f(\tilde{x}_T) - f(x^*)) \leq \left(\frac{2L\eta}{1-2\mu\eta}\right)^T \cdot \frac{5LD_0}{\theta} = \left(1 - \frac{(L+\mu)(1-2\theta)}{L+\mu-2\mu\theta}\right)^T \cdot \frac{5LD_0}{\theta}.$$

2. If $B < n$, then

$$J_2 \leq \frac{4L\eta}{1-(L+\mu)\eta(1+\beta)} \cdot \frac{\beta}{\beta-1} \cdot \frac{(n-B)}{(n-1)B} \cdot \frac{G_n}{L} \leq \frac{4\theta}{1-\theta(1+\beta)} \cdot \frac{\beta}{\beta-1} \cdot \frac{(n-B)}{(n-1)B} \cdot \frac{G_n}{L}.$$

As in the smooth case, set $\beta = \sqrt{(1-\theta)/\theta}$, then (5) is satisfied and

$$J_2 \leq \frac{4\theta}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \cdot \frac{(n-B)}{(n-1)B} \cdot \frac{G_n}{L}.$$

In addition,

$$J_1 \leq \frac{L}{2}\left(1 + \frac{(1-\mu\eta)(1-\gamma)}{L\eta\gamma\{1-\mu\eta(1+\sqrt{(1-\theta)/\theta})\}}\right) D_0.$$

Similar to the case $B = n$, we have

$$J_1 \leq \frac{L}{2}\left(1 + \frac{4(1-\mu\eta)(1-\gamma)}{\theta\gamma}\right) D_0 \leq \frac{L}{2}\left(1+\frac{4}{\theta}\right) D_0 \leq \frac{5LD_0}{\theta},$$

where the first inequality uses the fact that

$$1-\mu\eta(1+\sqrt{(1-\theta)/\theta}) \geq 1 - \frac{L+\mu}{2}\eta(1+\sqrt{(1-\theta)/\theta}) = 1 - \frac{\theta+\sqrt{\theta(1-\theta)}}{2} \geq \frac{1}{2}$$

and the second inequality uses $\gamma \geq 1-\mu\eta$. Therefore,

$$\mathbb{E}(f(\tilde{x}_T) - f(x^*)) \leq \frac{5LD_0}{\theta} \cdot \left(1 - \frac{(L+\mu)\sqrt{1-\theta}(\sqrt{1-\theta}-\sqrt{\theta})}{L+\mu-\mu(\theta+\sqrt{\theta(1-\theta)})}\right)^T + \frac{4\theta}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \cdot \frac{(n-B)}{(n-1)B} \cdot \frac{G_n}{L}.$$

■

Based on Theorem 3, we can obtain the following result.

Corollary 4 *Under the settings of Theorem 3, assume $\theta \in (0, \frac{1}{2} - \theta_0)$ and*

$$B \geq \min\left\{n, \frac{2C_3\theta}{\epsilon} \cdot \frac{G_n}{L}\right\} \tag{20}$$

where C_3 is defined in Theorem 3 and

$$T \geq \log\left(\frac{10D_0L}{\theta\epsilon}\right) \Big/ \log\left(\sqrt{\frac{1}{2\theta}}\right), \tag{21}$$

$$\mathbb{E}(f(\tilde{x}_T) - f(x^*)) \leq \epsilon.$$

Proof [**Corollary 4**] By remark 1, it is easy to verify that under (20) and (21),

$$\frac{5LD_0}{\theta} \cdot \left(1 - \frac{(L+\mu)\sqrt{1-\theta}(\sqrt{1-\theta}-\sqrt{\theta})}{L+\mu-\mu(\theta+\sqrt{\theta(1-\theta)})}\right)^T \leq \frac{5LD_0}{\theta}(\sqrt{2\theta})^T \leq \frac{\epsilon}{2},$$

and

$$\frac{4\theta}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \cdot \frac{(n-B)}{(n-1)B} \cdot \frac{G_n}{L} \leq \frac{C_3\theta}{B}\frac{G_n}{L} \leq \frac{\epsilon}{2}.$$

Therefore,

$$\mathbb{E}(f(\tilde{x}_T) - f(x^*)) \leq \epsilon.$$

■

Proof [**Corollary 2**] It is known from Corollary 4 that

$$\mathbb{E}(f(\tilde{x}_T) - f(x^*)) \leq \epsilon.$$

The computation cost is

$$\mathbb{E}C_{\text{comp}} = \mathbb{E}\sum_{j=1}^{T}(B+N_j) \le T(B+m).$$

Now we consider two conditions separately. By definition, $m = \frac{(\kappa+1)^2}{2\kappa\theta^2} = O(\frac{\kappa}{\theta^2})$ and hence

$$\mathbb{E}C_{\text{comp}} = O\left(\min\left\{n+\frac{\kappa}{\theta^2}, \frac{G_n\theta}{L\epsilon}+\frac{\kappa}{\theta^2}\right\}\log\frac{D_0L}{\theta\epsilon}\right).$$

Recalling that

$$\frac{G_n}{L^2} \le \frac{2}{L^2}\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\tilde{x}_0)\|^2 + 4D_0,$$

we conclude that

$$\mathbb{E}C_{\text{comp}} = O\left(\min\left\{n+\kappa, \frac{G_n}{L^2}\cdot\frac{L}{\epsilon}+\kappa\right\}\log\frac{D_0L}{\epsilon}\right).$$

In contrast, the communication cost does not depend on m and hence

$$\mathbb{E}C_{\text{comm}} = \mathbb{E}BT = O\left(\left(n\wedge\frac{G_n}{L^2}\cdot\frac{L}{\epsilon}\right)\log\frac{D_0L}{\epsilon}\right).$$

■

C Miscellanies

C.1 Convergence Analysis for Strongle Convex Case (Under Assumption A3)

In some applications, Assumption **A**2 might not be valid but Assumption **A**3 is. Since $\bar{\mu}$ is generally hard to estimate in cases without a L_2-regularization term, for which Assumption **A**2 holds and better results can be obtained from Theorem 2, we sample N_j from a geometric distribution as in the general convex case. The following theorem provides a similar result to Theorem 2, assuming only **A**1 and **A**3 but requiring more stringent conditions on the parameters.

Theorem 4 *Assume **A**1 and **A**3. Let* N_j

$$\zeta = \frac{1}{\bar{\mu}\eta}\frac{1-\gamma}{\gamma} + 2\eta L.$$

1. *If* $B = n$ *and* $\zeta < 1$*, then*

$$\mathbb{E}(f(\tilde{x}_T) - f(x^*)) \le \zeta^T(f(\tilde{x}_0) - f(x^*));$$

2. *If* $B < n$ *and* $\zeta + \alpha\eta L < 1$ *for some* $\alpha > 0$*, then*

$$\mathbb{E}(f(\tilde{x}_T) - f(x^*)) \le \tilde{\zeta}^T(f(\tilde{x}_0) - f(x^*)) + \frac{1}{1-\tilde{\zeta}}\cdot\frac{1+\alpha}{2\alpha}\eta G_n\frac{n-B}{(n-1)B}.$$

where $\tilde{\zeta} = \zeta + \alpha\eta L$.

Proof [**Theorem 4**] With an identical proof to Theorem 1, we reach the equation (9):

$$2\eta\mathbb{E}(f(\tilde{x}_j)-f(x^*)) \le \frac{1-\gamma}{\gamma}(\mathbb{E}\|\tilde{x}_{j-1}-x^*\|^2-\mathbb{E}\|\tilde{x}_j-x^*\|^2)+2\eta^2L(1+\beta)\mathbb{E}(f(\tilde{x}_{j-1})-f(x^*))+\frac{1}{\gamma}\Delta(\eta,\beta,B)$$

Instead of summing over j, we use the strong convexity of f and obtain that

$$2\eta\mathbb{E}(f(\tilde{x}_j)-f(x^*)) \le \frac{1-\gamma}{\gamma}\cdot\frac{2}{\bar{\mu}}\mathbb{E}(f(\tilde{x}_{j-1})-f(x^*))+2\eta^2L(1+\beta)\mathbb{E}(f(\tilde{x}_{j-1})-f(x^*))+\frac{1}{\gamma}\Delta(\eta,\beta,B).$$

Rearranging the terms we have

$$\mathbb{E}(f(\tilde{x}_j) - f(x^*)) \le (\zeta + (\beta-1)\eta L)\,\mathbb{E}(f(\tilde{x}_{j-1}) - f(x^*)) + \frac{1}{2\eta\gamma}\Delta(\eta,\beta,B).$$

We distinguish two cases:

1. If $B = n$, we set $\beta = 1$ and then
$$\mathbb{E}(f(\tilde{x}_j) - f(x^*)) \leq \zeta \mathbb{E}(f(\tilde{x}_{j-1}) - f(x^*))$$
which proves Part 1 of the theorem;
2. If $B < n$, we set $\beta = 1 + \alpha$ and then
$$\mathbb{E}(f(\tilde{x}_j) - f(x^*)) \leq \tilde{\zeta} \mathbb{E}(f(\tilde{x}_{j-1}) - f(x^*)) + \frac{1+\alpha}{2\alpha} \eta G_n \frac{n-B}{(n-1)B}.$$
Then a standard argument shows that
$$\mathbb{E}(f(\tilde{x}_T) - f(x^*)) \leq \tilde{\zeta}^T (f(\tilde{x}_0) - f(x^*)) + \frac{1}{1-\tilde{\zeta}} \cdot \frac{1+\alpha}{2\alpha} \eta G_n \frac{n-B}{(n-1)B}.$$

■

The requirement that $\zeta < 1$ implies that
$$\frac{1}{\bar{\mu}\eta} \frac{1-\gamma}{\gamma} < 1, \quad 2\eta L < 1,$$
and these results together imply that
$$\frac{\gamma}{1-\gamma} > \frac{2L}{\bar{\mu}}.$$
Denote by $\bar{\kappa} = \frac{L}{\bar{\mu}}$ the condition number of f, then
$$\mathbb{E}N_j = \frac{1}{1-\gamma} > 1 + 2\bar{\kappa}.$$
This plays a similar role to m as in the previous subsection with κ replaced by $\tilde{\kappa}$. Similar to Corollary 2, we can derive the computation and communication cost as follows.

Corollary 5 *Assume **A**1 and **A**3. Let $\bar{\kappa} = \frac{L}{\bar{\mu}}$ denote the condition number of f. Set γ to satisfy*
$$\frac{1}{1-\gamma} \geq 1 + c\bar{\kappa},$$
for some $c > 8$ and set $\eta = \frac{\theta}{L}$ for any θ with
$$\frac{1}{c\theta} + 2\theta < 1.$$
Further, select the batch size B and number of stages T such that
$$B = \min\left\{n, \left\lceil \frac{C_4}{\epsilon} \cdot \frac{G_n}{L} \right\rceil \right\}, \quad T = \left\lceil C_5 \log\left(\frac{(f(\tilde{x}_0) - f(x^*))}{\epsilon} \right) \right\rceil,$$
for some large enough constants C_4, C_5 which only depend on c and θ. Then $\mathbb{E}(f(\tilde{x}_T) - f(x^*)) \leq \epsilon$ and
$$\mathbb{E}C_{\text{comp}} = O\left(\left(n \wedge \frac{G_n}{L^2} \cdot \frac{L}{\epsilon} + \bar{\kappa} \right) \log \frac{D_0 L}{\epsilon} \right), \quad \mathbb{E}C_{\text{comm}} = O\left(\left(n \wedge \frac{G_n}{L^2} \cdot \frac{L}{\epsilon} \right) \log \frac{D_0 L}{\epsilon} \right).$$

Proof [Corollary 5] Since $\frac{1}{1-\gamma} \geq 1 + c\bar{\kappa}$ and $\eta = \frac{\theta}{L}$, it holds that
$$\zeta \leq \frac{1}{c\theta} + 2\theta.$$
Since $c > 8$, there exists $\theta > 0$ such that $\zeta < 1$. Similarly, there exists $\alpha > 0$ such that
$$\{\theta : \frac{1}{c\theta} + (2+\alpha)\theta < 1\}$$
is a non-empty set. Then under our settings, for sufficiently large C_4 and C_5, $\mathbb{E}(f(\tilde{x}_T) - f(x^*)) \leq \epsilon$. Similar to Corollary 1 and Corollary 2, we prove that
$$\mathbb{E}C_{\text{comp}} = O\left(\left(n \wedge \frac{G_n}{L\epsilon} + \bar{\kappa} \right) \log \frac{D_0 L}{\epsilon} \right), \quad \mathbb{E}C_{\text{comm}} = O\left(\left(n \wedge \frac{G_n}{L\epsilon} \right) \log \frac{D_0 L}{\epsilon} \right).$$

■

C.2 Communication Cost of CoCoA (Table 2)

CoCoA has an additional factor H [5] determining the iteration complexity T and hence the tradeoff between computation and communication. For given H, the computation cost of CoCoA is HT. Under our notation, Theorem 2 of [5] implies that the iteration complexity

$$T = \Omega\left(\frac{m}{1-\Theta}\frac{n+\kappa}{n}\log\frac{D_0 L}{\epsilon}\right) = \Omega\left(m\log\frac{1}{\tilde{\epsilon}}\right).$$

To match the computation cost of CoCoA to SCSG, we assume

$$H = O\left(\frac{1}{m}\left(n \wedge \frac{G_n}{L\epsilon} + \kappa\right)\right).$$

Then equation (5) of [5] implies that

$$\Theta \geq \left(1 - \frac{n}{n+\kappa}\cdot\frac{m}{n}\right)^H \leq \exp\left\{-m\cdot\frac{n\wedge\frac{G_n}{L\epsilon}+\kappa}{n+\kappa}\right\} = \left(\left(1-\frac{m}{n+\kappa}\right)^{\frac{n+\kappa}{m}}\right)^{\frac{n\wedge\frac{G_n}{L\epsilon}+\kappa}{n+\kappa}}.$$

In most applications, m/n is bounded away from 1. Suppose $m/n \leq \zeta < 1$, then

$$\Theta \geq \left((1-\zeta)^{\frac{1}{\zeta}}\right)^{\frac{n+\kappa}{n\wedge\frac{G_n}{L\epsilon}+\kappa}} \triangleq \exp\left\{-c\cdot\frac{n\wedge\frac{G_n}{L\epsilon}+\kappa}{n+\kappa}\right\}$$

where $c = -\zeta^{-1}\log(1-\zeta)$. Note that for any $a > 0$, $1/(1-e^{-a}) \geq a^{-1}$, we conclude that

$$\frac{1}{1-\Theta} \geq c^{-1}\cdot\frac{n+\kappa}{n\wedge\frac{G_n}{L\epsilon}+\kappa}$$

and hence

$$T = \Omega\left(m\cdot\frac{n+\kappa}{n\wedge\frac{G_n}{L\epsilon}+\kappa}\log\frac{1}{\tilde{\epsilon}}\right).$$

As a consequence, we obtain that the communication cost of CoCoA is at least

$$m^2\cdot\frac{n+\kappa}{n\wedge\frac{G_n}{L\epsilon}+\kappa}\log\frac{1}{\tilde{\epsilon}}.$$

C.3 Bounding M_1 and M_2 for (Multi-Class) Logistic Regression (Equation (3))

In Section 4 we claim that $M_1 = 2, M_2 = 1$ for (multi-class) logistic regression. Here we establish this claim. Denote by x the concatenation of $x_1, \ldots, x_{K-1}$ as in Section 5. Then

$$f_i(x) = \log\left(1 + \sum_{k=1}^{K-1} e^{a_i^T x_k}\right) - \sum_{k=1}^{K-1} I(y_i = k) a_i^T x_k.$$

For any $k = 1, \ldots, K-1$,

$$\frac{\partial f_i(x)}{\partial x_k} = \left(\frac{e^{a_i^T x_k}}{1+\sum_{k=1}^{K-1} e^{a_i^T x_k}} - I(y_i = k)\right)\cdot a_i$$

and thus

$$\nabla f_i(x) = H_i(x) \otimes a_i \Longrightarrow \|\nabla f_i(x)\|^2 = \|H_i(x)\|^2 \cdot \|a_i\|^2,$$

where

$$H_i(x) = \left(\frac{e^{a_i^T x_1}}{1+\sum_{k=1}^{K-1} e^{a_i^T x_k}} - I(y_i = 1), \ldots, \frac{e^{a_i^T x_{K-1}}}{1+\sum_{k=1}^{K-1} e^{a_i^T x_k}} - I(y_i = K-1)\right)^T.$$

It is easy to see that for any i and x

$$\|H_i(x)\|^2 \leq \|H_i(x)\|_1 \leq 2.$$

This entails that

$$G_n \leq \frac{1}{n}\sum_{i=1}^{n}\|a_i\|^2.$$

On the other hand, for any $k = 1, \ldots, K-1$,

$$\frac{\partial^2 f_i}{\partial x_k \partial x_k^T} = (H_i(x)_k - H_i(x)_k^2)a_i a_i^T,$$

and for any $k \neq k'$,

$$\frac{\partial^2 f_i}{\partial x_k \partial x_{k'}^T} = -H_i(x)_k H_i(x)_{k'} a_i a_i^T.$$

Thus,

$$\nabla^2 f_i(x) = (\mathrm{diag}(H_i(x)) - H_i(x)H_i(x)^T) \otimes a_i a_i^T.$$

As a consequence,

$$L \leq \max_i \lambda_{\max}(\nabla^2 f_i(x)) \leq \lambda_{\max}(\mathrm{diag}(H_i(x)) - H_i(x)H_i(x)^T) \cdot \sup\|a_i\|^2 \leq \sup\|a_i\|^2.$$

References

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

[2] Z. Allen-Zhu and Y. Yuan. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *International Conference on Machine Learning*, 2016.

[3] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[4] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[5] M. Jaggi, V. Smith, M. Takác, J. Terhorst, S. Krishnan, T. Hofmann, and M. I. Jordan. Communication-efficient distributed dual coordinate ascent. In *Advances in Neural Information Processing Systems*, pages 3068–3076, 2014.

[6] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[7] J. Konečnỳ, B. McMahan, and D. Ramage. Federated optimization: Distributed optimization beyond the datacenter. *ArXiv e-prints abs/1511.03575*, 2015.

[8] H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.

[9] Q. Lin, Z. Lu, and L. Xiao. An accelerated proximal coordinate gradient method. In *Advances in Neural Information Processing Systems*, pages 3059–3067, 2014.

[10] A. Nitanda. Accelerated stochastic gradient descent for minimizing finite sums. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 195–203, 2016.

[11] S. J. Reddi, A. Hefny, S. Sra, B. Póczós, and A. J. Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, 2016.

[12] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical Programming*, 127(1):3–30, 2011.

[13] S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *International Conference on Machine Learning*, pages 64–72, 2014.

[14] O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *International Conference on Machine Learning*, volume 32, pages 1000–1008, 2014.

[15] Y. Zhang and X. Lin. Disco: Distributed optimization for self-concordant empirical loss. In *International Conference on Machine Learning*, pages 362–370, 2015.

[16] Y. Zhang and L. Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *International Conference on Machine Learning*, volume 951, page 2015, 2015.