Supplementary Material

A. Proof of Lemma 1

Proof. Let $1_{i\in S} = 1$ if $i \in S$ and $1_{i\in S} = 0$ otherwise. Likewise, let $1_{i,j\in S} = 1$ if $i, j \in S$ and $1_{i,j\in S} = 0$ otherwise. Note that $\mathrm{E}\left[1_{i\in S}\right] = p_i$ and $\mathrm{E}\left[1_{i,j\in S}\right] = p_{ij}$. Next, let us compute the mean of $X := \sum_{i\in S} \frac{\zeta_i}{np_i}$:

$$\mathrm{E}\left[X\right] = \mathrm{E}\left[\sum_{i\in S} \frac{\zeta_i}{np_i}\right] = \mathrm{E}\left[\sum_{i=1}^n \frac{\zeta_i}{np_i} 1_{i\in S}\right] = \sum_{i=1}^n \frac{\zeta_i}{np_i} \mathrm{E}\left[1_{i\in S}\right] = \frac{1}{n}\sum_{i=1}^n \zeta_i = \bar{\zeta}. \tag{11}$$

Let $\mathbf{A} = [a_1, \ldots, a_n] \in \mathbb{R}^{d\times n}$, where $a_i = \frac{\zeta_i}{p_i}$, and let e be the vector of all ones in $\mathbb{R}^n$. We now write the variance of X in a form which will be convenient to establish a bound:

$$\begin{aligned}
\mathrm{E}\left[\|X - \mathrm{E}\left[X\right]\|^2\right] &= \mathrm{E}\left[\|X\|^2\right] - \|\mathrm{E}\left[X\right]\|^2 \\
&= \mathrm{E}\left[\left\|\sum_{i\in S} \frac{\zeta_i}{np_i}\right\|^2\right] - \|\bar{\zeta}\|^2 \\
&= \mathrm{E}\left[\sum_{i,j} \frac{\zeta_i^\top}{np_i}\frac{\zeta_j}{np_j} 1_{i,j\in S}\right] - \|\bar{\zeta}\|^2 \\
&= \sum_{i,j} p_{ij} \frac{\zeta_i^\top}{np_i}\frac{\zeta_j}{np_j} - \sum_{i,j} \frac{\zeta_i^\top}{n}\frac{\zeta_j}{n} \\
&= \frac{1}{n^2}\sum_{i,j} (p_{ij} - p_i p_j) a_i^\top a_j \\
&= \frac{1}{n^2} e^\top \left(\left(\mathbf{P} - pp^\top\right) \circ \mathbf{A}^\top \mathbf{A}\right) e.
\end{aligned} \tag{12}$$

Since by assumption we have $\mathbf{P} - pp^\top \preceq \mathbf{Diag}\left(p \circ v\right)$, we can further bound

$$e^\top \left(\left(\mathbf{P} - pp^\top\right) \circ \mathbf{A}^\top \mathbf{A}\right) e \le e^\top \left(\mathbf{Diag}\left(p\circ v\right) \circ \mathbf{A}^\top \mathbf{A}\right) e = \sum_{i=1}^n p_i v_i \|a_i\|^2.$$

To obtain (5), it remains to combine this with (12).

Inequality (6) follows by comparing the diagonal elements of the two matrices in (4). Let us now verify the formulas for v.

- Since $\mathbf{P} - pp^\top$ is positive semidefinite (Richtárik and Takáč, 2016b), we can bound $\mathbf{P} - pp^\top \preceq n\mathbf{Diag}\left(\mathbf{P} - pp^\top\right) = \mathbf{Diag}\left(p \circ v\right)$, where $v_i = n(1 - p_i)$.

- It was shown by Qu and Richtárik (2016, Theorem 4.1) that $\mathbf{P} \preceq d\mathbf{Diag}\left(p\right)$ provided that $|S| \le d$ with probability 1. Hence, $\mathbf{P} - pp^\top \preceq \mathbf{P} \preceq d\mathbf{Diag}\left(p\right)$, which means that $v_i = d$ for all i.

- Consider now the independent sampling. Clearly,

$$\mathbf{P} - pp^\top = \begin{bmatrix} p_1(1-p_1) & 0 & \ldots & 0 \\ 0 & p_2(1-p_2) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & p_n(1-p_n) \end{bmatrix} = \mathbf{Diag}\left(p_1 v_1, \ldots, p_n v_n\right),$$

where $v_i = 1 - p_i$.

- Consider the b–nice sampling (standard uniform minibatch sampling). Direct computation shows that the probability matrix is given by

$$\mathbf{P} = \begin{bmatrix} \frac{b}{n} & \frac{b(b-1)}{n(n-1)} & \cdots & \frac{b(b-1)}{n(n-1)} \\ \frac{b(b-1)}{n(n-1)} & \frac{b}{n} & \cdots & \frac{b(b-1)}{n(n-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{b(b-1)}{n(n-1)} & \frac{b(b-1)}{n(n-1)} & \cdots & \frac{b}{n} \end{bmatrix},$$

as claimed in (3). Therefore,

$$\mathbf{P} - pp^\top = \begin{bmatrix} \frac{b}{n} - \frac{b^2}{n^2} & \frac{b(b-1)}{n(n-1)} & \cdots & \frac{b(b-1)}{n(n-1)} \\ \frac{b(b-1)}{n(n-1)} & \frac{b}{n} & \cdots & \frac{b(b-1)}{n(n-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{b(b-1)}{n(n-1)} & \frac{b(b-1)}{n(n-1)} & \cdots & \frac{b}{n} \end{bmatrix},$$

- Letting $t = \frac{(a-1)k}{a(k-1)}$ and $s = 1 - t = \frac{k-a}{a(k-1)}$ the probability matrix of the approximate independent sampling satisfies

$$\begin{aligned} \mathbf{P} - pp^\top &= \begin{bmatrix} p_1(1-p_1) & (t-1)p_1p_2 & \dots & (t-1)p_1p_k & 0 & \dots & 0 \\ (t-1)p_2p_1 & p_2(1-p_2) & \dots & (t-1)p_2p_k & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \dots & 0 \\ (t-1)p_np_1 & (t-1)p_np_2 & \dots & p_k(1-p_k) & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix} \\ &= \mathbf{Diag}\left(p_1(1-p_1(1-s)), \dots, p_k(1-p_k(1-s)), 0, \dots, 0\right) - sp_kp_k^\top \\ &\preceq \mathbf{Diag}\left(p_1(1-p_1(1-s)), \dots, p_n(1-p_n(1-s)), 0, \dots, 0\right), \end{aligned}$$

where $p_k = (p_1, \dots, p_k, 0, \dots, 0)^\top$. Therefore, $v_i = 1 - p_i(1-s)$ for $i \leq k$ and $v_i = 0$ otherwise works.

- Finally, as remarked in the introduction, the standard uniform minibatch sampling (b–nice sampling) arises as a special case of the approximate independent sampling for the choice $p_i = b/n$. Thus $k = n$, $a = b$ and hence $s = \frac{n-b}{b(n-1)}$. Based on the previous result, $v_i = 1 - \frac{b}{n}(1 - \frac{n-b}{b(n-1)}) = \frac{n-b}{n-1}$ works.

□

B. Proof of Theorem 2

We first establish a lemma we will need in order to prove Theorem 2.

Lemma 6. *Let $0 < L_1 \leq L_2 \leq \cdots \leq L_n$ be positive real numbers, $0 < b \leq n$, and consider the optimization problem*

$$\begin{aligned} &\textit{minimize}_{p \in \mathbb{R}^n} \quad \Omega(p) := \sum_{i=1}^n \frac{L_i^2}{p_i} \\ &\textit{subject to} \quad \sum_{i=1}^n p_i = b, \\ &\qquad\qquad\quad 0 \leq p_i \leq 1, \quad i = 1, 2, \dots, n. \end{aligned} \tag{13}$$

Let be the largest integer for which $0 < b + k - n \leq \frac{\sum_{i=1}^k L_i}{L_k}$ (note that the inequality holds for $k = n - b + 1$). Then (13) has the following solution:

$$p_i = \begin{cases} (b + k - n)\frac{L_i}{\sum_{j=1}^k L_j}, & \textit{if } i \leq k, \\ 1, & \textit{if } i > k. \end{cases} \tag{14}$$

Proof. The Lagrangian of the problem is

$$L(p, y, \lambda_1, .., \lambda_n, u_1, ..., u_n) = \sum_{i=1}^n \frac{L_i^2}{p_i} - \sum_{i=1}^n \lambda_i p_i - \sum_{i=1}^n u_i(1-p_i) + y\left(\sum_{i=1}^n p_i - b\right).$$

Th constraints are linear and hence KKT conditions hold. The result can be deduced from the KKT conditions. □

We can now proceed with the proof. Since n, b and $\bar{L}$ are constants, the problem is equivalent to

$$\begin{array}{ll} \text{minimize}_S & \psi(S) := \sum_{i=1}^n \frac{v_i L_i^2}{p_i} \\ \text{subject to} & v_i \quad \text{satisfies} \quad (4). \end{array}$$

In view of (6),

$$\psi(S) \overset{(6)}{\geq} \sum_{i=1}^n \frac{(1-p_i)L_i^2}{p_i} = \sum_{i=1}^n \frac{L_i^2}{p_i} - \sum_{i=1}^n L_i^2 = \Omega(p) - \sum_{i=1}^n L_i^2,$$

where function $\Omega(p)$ was defined in Lemma 6. Since $b = \mathrm{E}\left[|S|\right] = \sum_i p_i$, and $0 \leq p_i \leq 1$ for all i, then in view of Lemma 6 we have

$$\Psi(S) \geq \Omega(p^*) - \sum_{i=1}^n L_i^2,$$

where p^* is defined by (8).

On the other hand, from Lemma 1 we know that the independent sampling $S = S^*$ with probability vector p^* defined in (8) satisfies inequality (4) with $v_i = 1 - p_i$, and hence

$$\Psi(S^*) = \Omega(p^*) - \sum_{i=1}^n L_i^2.$$

Hence, it is optimal. □

C. Improvements

Let us compute α for uniform sampling.

$$\begin{aligned} \alpha \quad &\overset{(7)}{=} \quad \left(\frac{b}{n^2}\sum_{i=1}^n \frac{v_i L_i^2}{p_i}\right)/\bar{L}^2 \\ &\overset{\text{Lemma 1}}{=} \quad \left(\frac{(n-b)}{(n-1)n}\sum_{i=1}^n L_i^2\right)/\bar{L}^2 \\ &= \quad n\frac{(n-b)}{(n-1)}\sum_{i=1}^n L_i^2 / \left(\sum_{i=1}^n L_i\right)^2 \end{aligned}$$

It is easy to see that $L_{\max} \geq \bar{L}$. To prove that we have improved current best known rates, we need to show that $\bar{L}\alpha \leq L_{\max}$ and $\bar{L}^2\alpha \leq \frac{(n-b)}{(n-1)}L_{\max}^2$

Proof.

$$\begin{aligned} \bar{L}\alpha \quad &= \quad n\frac{(n-b)}{(n-1)}\frac{\sum_{i=1}^n L_i^2}{\left(\sum_{i=1}^n L_i\right)^2}\bar{L} \leq \frac{\sum_{i=1}^n L_i^2}{\left(\sum_{i=1}^n L_i\right)} = \frac{\sum_{i=1}^n L_{\max} L_i}{\left(\sum_{i=1}^n L_i\right)} = L_{\max}, \\ \bar{L}^2\alpha \quad &= \quad n\frac{(n-b)}{(n-1)}\frac{\sum_{i=1}^n L_i^2}{\left(\sum_{i=1}^n L_i\right)^2}\bar{L}^2 = \frac{(n-b)}{(n-1)}\frac{1}{n}\sum_{i=1}^n L_i^2 \leq \frac{(n-b)}{(n-1)}\frac{1}{n}\sum_{i=1}^n L_{\max}^2 = \frac{(n-b)}{(n-1)}L_{\max}^2, \end{aligned}$$

□

Let's take $b = 1$. If $L_n \gg L_i, \forall i \in [n]$, then $L_{\max} \approx n\bar{L}$ and $\alpha_{S^*} \leq 1$ and $\alpha_{S^u} \approx n$, which essentially means, that we can have in theory speedup by factor of n.

D. Stochastic gradients evaluation complexity

D.1. SVRG

For SVRG, each outer loop costs $n + mb$ evaluations of stochastic gradient. If we want to obtain ϵ-solution, following must hold (Theorem 3)

$$\frac{\alpha\bar{L}n^{(2/3)}(f(x^0) - f(x^*))}{bMm\nu_2} \leq \epsilon$$

Combining these two equations with definition from Theorem 3, we get total complexity in terms of stochastic gradients evaluation

$$\frac{\mu_2\bar{L}n^{(2/3)}(f(x^0) - f(x^*))}{\epsilon\nu_2}(1 + \frac{\alpha}{3\mu_2})$$

D.2. SAGA

For SAGA, each loop costs $d + b$ evaluations of stochastic gradient. If we want to obtain ϵ-solution, following must hold (Theorem 4)

$$\frac{\alpha\bar{L}n^{(2/3)}(f(x^0) - f(x^*))}{bT\nu_2} \leq \epsilon$$

Combining these two equations with definition from Theorem 4, we get total complexity in terms of stochastic gradients evaluation

$$n + \frac{\bar{L}n^{(2/3)}(f(x^0) - f(x^*))}{\epsilon\nu_3}(1 + \alpha),$$

because of evaluation of full gradient on the start.

D.3. SARAH

For SARAH with one outer loot, each inner loop costs $2b$ evaluations of stochastic gradient. If we want to obtain ϵ-solution, following must hold (Theorem 5)

$$\frac{2\bar{L}(f(x^0) - f(x^*))\left(\sqrt{1 + \frac{4m\alpha}{b}}\right)}{m} \leq \epsilon$$

Solving this equation for m, we get

$$m \leq \frac{16\alpha\bar{L}^2(f(x^0) - f(x^*))^2 + \sqrt{16^2\alpha^2\bar{L}^4(f(x^0) - f(x^*))^4 + 16\epsilon^2\bar{L}^2(f(x^0) - f(x^*))^2b^2}}{2b\epsilon^2}$$

Combining thise equation with complexity off each inner loop we obtain total complexity in terms of stochastic gradients evaluation

$$\frac{16\alpha\bar{L}^2(f(x^0) - f(x^*))^2 + \sqrt{16^2\alpha^2\bar{L}^4(f(x^0) - f(x^*))^4 + 16\epsilon^2\bar{L}^2(f(x^0) - f(x^*))^2b^2}}{2\epsilon^2}.$$

E. Proofs for SVRG

Lemma 7. *For $c_t, c_{t+1}, \beta > 0$, suppose we have*

$$c_t = c_{t+1}(1 + \eta\beta + 2\eta^2 K) + K\eta^2\bar{L}.$$

Let η, β and c_{t+1} be chosen such that $\Gamma_t > 0$ (in Theorem (17)). The iterate x_t^{s+1} in Algorithm 5 satisfy the bound:

$$\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] \le \frac{R_t^{s+1} - R_{t+1}^{s+1}}{\Gamma_t},$$

where $R_t^{s+1} := \mathrm{E}\left[f(x_t^{s+1}) + c_t\|x_t^{s+1} - \tilde{x}^s\|^2\right]$ for $0 \le s \le S-1$.

Proof. Since f_i is L_i-smooth we have

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

Summing through all i and dividing by n we obtain

$$\mathrm{E}\left[f(x_{t+1}^{s+1})\right] \le \mathrm{E}\left[f(x_t^{s+1}) + \langle \nabla f(x_t^{s+1}), x_{t+1}^{s+1} - x_t^{s+1}\rangle + \tfrac{\bar{L}}{2}\|x_{t+1}^{s+1} - x_t^{s+1}\|^2\right].$$

Using the SVRG update in Algorithm 5 and its unbiasedness ($\mathrm{E}\left[i_t\right] v_t^{s+1} = \nabla f(x_t^{s+1})$), the right hand side above is further upper bounded by

$$\mathrm{E}\left[f(x_t^{s+1}) - \eta\|\nabla f(x_t^{s+1})\|^2 + \tfrac{\bar{L}\eta^2}{2}\|v_t^{s+1}\|^2\right]. \tag{15}$$

Consider now the Lyapunov function

$$R_t^{s+1} := \mathrm{E}\left[f(x_t^{s+1}) + c_t\|x_t^{s+1} - \tilde{x}^s\|^2\right].$$

For bounding it we will require the following:

$$\begin{aligned}
\mathrm{E}\left[\|x_{t+1}^{s+1} - \tilde{x}^s\|^2\right] &= \mathrm{E}\left[\|x_{t+1}^{s+1} - x_t^{s+1} + x_t^{s+1} - \tilde{x}^s\|^2\right] \\
&= \mathrm{E}\left[\|x_{t+1}^{s+1} - x_t^{s+1}\|^2 + \|x_t^{s+1} - \tilde{x}^s\|^2 \right.\\
&\qquad \left. + 2\langle x_{t+1}^{s+1} - x_t^{s+1}, x_t^{s+1} - \tilde{x}^s\rangle\right] \\
&= \mathrm{E}\left[\eta^2\|v_t^{s+1}\|^2 + \|x_t^{s+1} - \tilde{x}^s\|^2\right] \\
&\qquad - 2\eta\mathrm{E}\left[\langle \nabla f(x_t^{s+1}), x_t^{s+1} - \tilde{x}^s\rangle\right] \\
&\overset{(54),(55)}{\le} \mathrm{E}\left[\eta^2\|v_t^{s+1}\|^2 + \|x_t^{s+1} - \tilde{x}^s\|^2\right] \\
&\qquad + 2\eta\mathrm{E}\left[\tfrac{1}{2\beta}\|\nabla f(x_t^{s+1})\|^2 + \tfrac{1}{2}\beta\|x_t^{s+1} - \tilde{x}^s\|^2\right].
\end{aligned} \tag{16}$$

The second equality follows from the unbiasedness of the update of SVRG. Plugging Equation (15) and Equation (16) into R_{t+1}^{s+1}, we obtain the following bound:

$$\begin{aligned}
R_{t+1}^{s+1} &\le \mathrm{E}\left[f(x_t^{s+1}) - \eta\|\nabla f(x_t^{s+1})\|^2 + \tfrac{\bar{L}\eta^2}{2}\|v_t^{s+1}\|^2\right] \\
&\quad + \mathrm{E}\left[c_{t+1}\eta^2\|v_t^{s+1}\|^2 + c_{t+1}\|x_t^{s+1} - \tilde{x}^s\|^2\right] \\
&\quad + 2c_{t+1}\eta\mathrm{E}\left[\tfrac{1}{2\beta}\|\nabla f(x_t^{s+1})\|^2 + \tfrac{1}{2}\beta\|x_t^{s+1} - \tilde{x}^s\|^2\right] \\
&\le \mathrm{E}\left[f(x_t^{s+1}) - \left(\eta - \tfrac{c_{t+1}\eta}{\beta}\right)\|\nabla f(x_t^{s+1})\|^2\right] + \left(\tfrac{\bar{L}\eta^2}{2} + c_{t+1}\eta^2\right)\mathrm{E}\left[\|v_t^{s+1}\|^2\right] \\
&\quad + \left(c_{t+1} + c_{t+1}\eta\beta\right)\mathrm{E}\left[\|x_t^{s+1} - \tilde{x}^s\|^2\right].
\end{aligned} \tag{17}$$

To further bound this quantity, we use Lemma 10 to bound $\mathrm{E}\left[\|v_t^{s+1}\|^2\right]$, so that upon substituting it in Equation (17), we see that

$$\begin{aligned} R_{t+1}^{s+1} \overset{(29)}{\leq} \; & \mathrm{E}\left[f(x_t^{s+1})\right] - \left(\eta - \tfrac{c_{t+1}\eta}{\beta} - \eta^2\bar{L} - 2c_{t+1}\eta^2\right)\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] \\ & + \left[c_{t+1}\left(1 + \eta\beta + 2\eta^2 K\right) + \eta^2 K\bar{L}\right]\mathrm{E}\left[\|x_t^{s+1} - \tilde{x}^s\|^2\right] \\ \leq \; & R_t^{s+1} - \left(\eta - \tfrac{c_{t+1}\eta}{\beta} - \eta^2\bar{L} - 2c_{t+1}\eta^2\right)\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right]. \end{aligned} \tag{18}$$

The second inequality follows from the definition of c_t and R_t^{s+1}, thus concluding the proof. □

Proof of Lemma 7 and Theorem 17

Proof. Using Lemma 7 and telescoping the sum, we obtain

$$\sum_{t=0}^{m-1}\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] \leq \frac{R_0^{s+1} - R_m^{s+1}}{\gamma_n}. \tag{19}$$

This inequality in turn implies that

$$\sum_{t=0}^{m-1}\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] \leq \frac{\mathrm{E}\left[f(\tilde{x}^s) - f(\tilde{x}^{s+1})\right]}{\gamma_n}, \tag{20}$$

where we used that $R_m^{s+1} = \mathrm{E}\left[f(x_m^{s+1})\right] = \mathrm{E}\left[f(\tilde{x}^{s+1})\right]$ (since $c_m = 0$), and that $R_0^{s+1} = \mathrm{E}\left[f(\tilde{x}^s)\right]$ (since $x_0^{s+1} = \tilde{x}^s$). Now sum over all epochs to obtain

$$\frac{1}{T}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] \leq \frac{f(x^0) - f(x^*)}{T\gamma_n}. \tag{21}$$

The above inequality used the fact that $\tilde{x}^0 = x^0$. Using the above inequality and the definition of x_a in Algorithm 5, we obtain the desired result. □

Proof of Theorem 18

Proof. For our analysis, we will require an upper bound on c_0. Let $m = \lfloor Kn/(3\bar{L}^2\rfloor\mu_0), \eta = \mu_0\bar{L}/(Kn^{2/3})$. We observe that $c_0 = \frac{\mu_0^2\bar{L}^3}{Kn^{4/3}}\frac{(1+\theta)^m - 1}{\theta}$ where $\theta = 2K\eta^2 + \eta\beta$. This is obtained using the relation $c_t = c_{t+1}(1 + \eta\beta + 2K\eta^2) + \eta^2 K\bar{L}$ and the fact that $c_m = 0$. Using the specified values of β and η we have

$$\theta = 2K\eta^2 + \eta\beta = \frac{2\mu_0^2\bar{L}^2}{Kn^{4/3}} + \frac{\mu_0\bar{L}^2}{Kn} \leq \frac{3\mu_0\bar{L}^2}{Kn}. \tag{22}$$

The above inequality follows since $\mu_0 \leq 1$ and $n \geq 1$. Using the above bound on θ, we get

$$\begin{aligned} c_0 \;&=\; \frac{\mu_0^2\bar{L}^3}{n^2 K}\frac{(1+\theta)^m - 1}{\theta} = \frac{\mu_0\bar{L}((1+\theta)^m - 1)}{2\mu_0 + n^{\frac{1}{3}}} \\ &\leq\; \frac{\mu_0\bar{L}((1 + \frac{3\mu_0\bar{L}^2}{nK})^{\lfloor Kn/3\mu_0\bar{L}^2\rfloor} - 1)}{2\mu_0 + n^{\frac{1}{3}}} \\ &\leq\; n^{-\frac{1}{3}}(\mu_0\bar{L}(e-1)), \end{aligned} \tag{23}$$

wherein the second inequality follows upon noting that $(1 + \frac{1}{l})^l$ is increasing for $l > 0$ and $\lim_{l\to\infty}(1 + \frac{1}{l})^l = e$ (here e is the Euler's number). Now we can lower bound γ_n, as

$$\begin{aligned} \gamma_n \;&= \min_t\left(\eta - \tfrac{c_{t+1}\eta}{\beta} - \eta^2\bar{L} - 2c_{t+1}\eta^2\right) \\ &\geq \left(\eta - \tfrac{c_0\eta}{\beta} - \eta^2\bar{L} - 2c_0\eta^2\right) \geq \tfrac{\nu\bar{L}}{Kn^{\frac{2}{3}}}, \end{aligned} \tag{24}$$

where ν is a constant independent of n. The first inequality holds since c_t decreases with t. The second inequality holds since (a) c_0/β is upper bounded by a constant independent of n as $c_0/\beta \leq \mu_0(e-1)$ (follows from Equation (23)), (b) $\eta^2 \bar{L} \leq \mu_0 \eta$ and (c) $2c_0\eta^2 \leq 2\mu_0^2(e-1)\eta$ (follows from Equation (23)). By choosing μ_0 (independent of n) appropriately, one can ensure that $\gamma_n \geq \nu \bar{L}/(Kn^{\frac{2}{3}})$ for some universal constant ν. For example, choosing $\mu_0 = 1/4$, we have $\gamma_n \geq \nu \bar{L}/(Kn^{\frac{2}{3}})$ with $\nu = 1/40$. Substituting the above lower bound in Equation (21), we obtain the desired result. □

F. Minibatch SVRG

Proof of Theorem 3

The proofs essentially follow along the lines of Lemma 7, Theorem 17 and Theorem 18 with the added complexity of mini-batch. We first prove few intermediate results before proceeding to the proof of Theorem 3.

Lemma 8. *Suppose we have*

$$\begin{aligned} \overline{R}_t^{s+1} &:= \mathrm{E}\left[f(x_t^{s+1}) + \overline{c}_t \|x_t^{s+1} - \tilde{x}^s\|^2\right], \\ \overline{c}_t &= \overline{c}_{t+1}(1 + \eta\beta + \tfrac{2K\eta^2}{b}) + \tfrac{K\eta^2\bar{L}}{b}, \end{aligned} \tag{25}$$

for $0 \leq s \leq S-1$ *and* $0 \leq t \leq m-1$ *and the parameters* η, β *and* $\overline{c}_{t+1}$ *are chosen such that*

$$\left(\eta - \frac{\overline{c}_{t+1}\eta}{\beta} - \eta^2\bar{L} - 2\overline{c}_{t+1}\eta^2\right) \geq 0.$$

Then the iterates x_t^{s+1} *in the mini-batch version of Algorithm 5 i.e., Algorithm 1 with expected mini-batch size* b *satisfy the bound:*

$$\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] \leq \frac{\overline{R}_t^{s+1} - \overline{R}_{t+1}^{s+1}}{\left(\eta - \frac{\overline{c}_{t+1}\eta}{\beta} - \eta^2\bar{L} - 2\overline{c}_{t+1}\eta^2\right)},$$

Proof. Using essentially the same argument as the proof of Lemma 7 until Equation (17), we have

$$\begin{aligned} \overline{R}_{t+1}^{s+1} \quad \leq \quad & \mathrm{E}\left[(x_t^{s+1})\right] - \left(\eta - \tfrac{\overline{c}_{t+1}\eta}{\beta}\right)\|\nabla f(x_t^{s+1})\|^2 + \left(\tfrac{\bar{L}\eta^2}{2} + \overline{c}_{t+1}\eta^2\right)\mathrm{E}\left[\|v_t^{s+1}\|^2\right] \\ & + (\overline{c}_{t+1} + \overline{c}_{t+1}\eta\beta)\,\mathrm{E}\left[\|x_t^{s+1} - \tilde{x}^s\|^2\right]. \end{aligned} \tag{26}$$

We use Lemma 11 in order to bound $\mathrm{E}\left[\|v_t^{s+1}\|^2\right]$ in the above inequality. Substituting it in Equation (26), we see that

$$\begin{aligned} \overline{R}_{t+1}^{s+1} \quad \overset{(30)}{\leq} \quad & \mathrm{E}\left[f(x_t^{s+1})\right] - \left(\eta - \tfrac{\overline{c}_{t+1}\eta}{\beta} - \eta^2\bar{L} - 2\overline{c}_{t+1}\eta^2\right)\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] \\ & + \left[\overline{c}_{t+1}\left(1 + \eta\beta + \tfrac{2K\eta^2}{b}\right) + \tfrac{K\eta^2\bar{L}}{b}\right]\mathrm{E}\left[\|x_t^{s+1} - \tilde{x}^s\|^2\right] \\ \overset{(25)}{\leq} \quad & \overline{R}_t^{s+1} - \left(\eta - \tfrac{\overline{c}_{t+1}\eta}{\beta} - \eta^2\bar{L} - 2\overline{c}_{t+1}\eta^2\right)\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right]. \end{aligned} \tag{27}$$

The second inequality follows from the definition of $\overline{c}_t$ and $\overline{R}_t^{s+1}$, thus concluding the proof. □

The following theorem provides convergence rate of mini-batch`SVRG`.

Theorem 9. *Let* $\overline{\gamma}_n$ *denote the following quantity:*

$$\overline{\gamma}_n := \min_{0 \leq t \leq m-1} \quad \left(\eta - \tfrac{\overline{c}_{t+1}\eta}{\beta} - \eta^2\bar{L} - 2\overline{c}_{t+1}\eta^2\right).$$

Suppose $\overline{c}_m = 0$, $\overline{c}_t = \overline{c}_{t+1}(1 + \eta\beta + \frac{2K\eta^2}{b}) + \frac{K\eta^2\bar{L}}{b}$ *for* $t \in \{0, \ldots, m-1\}$ *and* $\overline{\gamma}_n > 0$. *Then for the output* x_a *of mini-batch version of Algorithm 5 with mini-batch size* b, *we have*

$$\mathrm{E}\left[\|\nabla f(x_a)\|^2\right] \leq \frac{f(x^0) - f(x^*)}{T\overline{\gamma}_n},$$

where x^* *is an optimal solution to* (1).

Proof. Using Lemma 8 and telescoping the sum, we obtain

$$\sum_{t=0}^{m-1} \mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] \leq \frac{\overline{R}_0^{s+1} - \overline{R}_m^{s+1}}{\overline{\gamma}_n}.$$

This inequality in turn implies that

$$\sum_{t=0}^{m-1} \mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] \leq \frac{\mathrm{E}\left[f(\tilde{x}^s) - f(\tilde{x}^{s+1})\right]}{\overline{\gamma}_n},$$

where we used that $\overline{R}_m^{s+1} = \mathrm{E}\left[f(x_m^{s+1})\right] = \mathrm{E}\left[f(\tilde{x}^{s+1})\right]$ (since $\overline{c}_m = 0$), and that $\overline{R}_0^{s+1} = \mathrm{E}\left[f(\tilde{x}^s)\right]$. Now sum over all epochs and using the fact that $\tilde{x}^0 = x^0$, we get the desired result. □

We now present the proof of Theorem 3 using the above results.

Proof of Theorem 3. We first observe that using the specified values of $\beta = \bar{L}/n^{1/3}$, $\eta = \mu_2 b\bar{L}/(Kn^{2/3})$ and $\eta = \lfloor nK/(b\bar{L}^2\mu_2)\rfloor$we obtain

$$\overline{\theta} := \frac{2K\eta^2}{b} + \eta\beta = \frac{2\mu_2^2 b\bar{L}^2}{Kn^{4/3}} + \frac{\bar{L}^2\mu_2 b}{Kn} \leq \frac{3\mu_2\bar{L}^2 b}{Kn}.$$

The above inequality follows since $\mu_2 \leq 1$ and $n \geq 1$. For our analysis, we will require the following bound on $\overline{c}_0$:

$$\begin{aligned}\overline{c}_0 &= \frac{\mu_2^2 b^2\bar{L}^3}{Kbn^{4/3}} \frac{(1+\overline{\theta})^m - 1}{\overline{\theta}} = \frac{\mu_2 b\bar{L}((1+\overline{\theta})^m - 1)}{2b\mu_2 + bn^{1/3}} \\ &\leq n^{-1/3}(\mu_2\bar{L}(e-1)),\end{aligned} \tag{28}$$

wherein the first equality holds due to the relation $\overline{c}_t = \overline{c}_{t+1}(1 + \eta\beta + \frac{2K\eta^2}{b}) + \frac{K\eta^2\bar{L}}{b}$, and the inequality follows upon again noting that $(1+1/l)^l$ is increasing for $l > 0$ and $\lim_{l\to\infty}(1+\frac{1}{l})^l = e$. Now we can lower bound $\overline{\gamma}_n$, as

$$\begin{aligned}\overline{\gamma}_n &= \min_t\left(\eta - \tfrac{\overline{c}_{t+1}\eta}{\beta} - \eta^2\bar{L} - 2\overline{c}_{t+1}\eta^2\right) \\ &\geq \left(\eta - \tfrac{\overline{c}_0\eta}{\beta} - \eta^2\bar{L} - 2\overline{c}_0\eta^2\right) \geq \frac{b\bar{L}\nu_2}{Kn^{2/3}},\end{aligned}$$

where ν_2 is a constant independent of n. The first inequality holds since $\overline{c}_t$ decreases with t. The second one holds since (a) $\overline{c}_0/\beta$ is upper bounded by a constant independent of n as $\overline{c}_0/\beta \leq \mu_2(e-1)$ (due to Equation(28)), (b) $\eta^2\bar{L} \leq \mu_2\eta$ (as $b \leq {}^{K}/_{\bar{L}^2 n^{2/3}}$) and (c) $2\overline{c}_0\eta^2 \leq 2\mu_2^2(e-1)\eta$ (again due to Equation (28) and the fact $b \leq {}^{K}/_{\bar{L}^2 n^{2/3}}$). By choosing an appropriately small constant μ_2 (independent of n), one can ensure that $\overline{\gamma}_n \geq \bar{L}b\nu_2/(Kn^{2/3})$ for some universal constant ν_2. For example, choosing $\mu_2 = 1/4$, we have $\overline{\gamma}_n \geq \bar{L}b\nu_2/(Kn^{2/3})$ with $\nu_2 = 1/40$. Substituting the above lower bound in Theorem 9, we obtain the desired result. □

LEMMAS

Lemma 10. *For the intermediate iterates v_t^{s+1} computed by Algorithm 5, we have the following:*

$$\mathrm{E}\left[\|v_t^{s+1}\|^2\right] \leq 2\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] + 2K\mathrm{E}\left[\|x_t^{s+1} - \tilde{x}^s\|^2\right]. \tag{29}$$

Proof. The proof simply follows from the proof of Lemma 11 with $S_t = \{i_t\}$. □

We now present a result to bound the variance of mini-batch SVRG.

Lemma 11. *Let v_t^{s+1} be computed by the mini-batch version of Algorithm 5 i.e., Algorithm 1 with sampling S. Then,*

$$\mathrm{E}\left[\|v_t^{s+1}\|^2\right] \leq 2\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] + \tfrac{2K}{b}\mathrm{E}\left[\|x_t^{s+1} - \tilde{x}^s\|^2\right]. \tag{30}$$

Proof. For the simplification, we use the following notation:

$$\zeta_t^{s+1} = \sum_{i_t \in S_t} \frac{1}{np_{i_t}} \left(\nabla f_{i_t}(x_t^{s+1}) - \nabla f_{i_t}(\tilde{x}^s) \right).$$

We use the definition of v_t^{s+1} to get

$$\begin{aligned}
\mathrm{E}\left[\|v_t^{s+1}\|^2\right] &= \mathrm{E}\left[\|\zeta_t^{s+1} + \nabla f(\tilde{x}^s)\|^2\right] \\
&= \mathrm{E}\left[\|\zeta_t^{s+1} + \nabla f(\tilde{x}^s) - \nabla f(x_t^{s+1}) + \nabla f(x_t^{s+1})\|^2\right] \\
&\leq 2\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] + 2\mathrm{E}\left[\|\zeta_t^{s+1} - \mathrm{E}\left[\zeta_t^{s+1}\right]\|^2\right] \\
&= 2\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] \\
&\quad + 2\mathrm{E}\left[\left\| \sum_{i_t \in S_t} \left(\frac{1}{np_{i_t}} \left(\nabla f_{i_t}(x_t^{s+1}) - \nabla f_{i_t}(\tilde{x}^s) \right) - \mathrm{E}\left[\zeta_t^{s+1}\right] \right) \right\|^2 \right].
\end{aligned}$$

The first inequality follows from fact that $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$ and the fact that $\mathrm{E}\left[\zeta_t^{s+1}\right] = \nabla f(x_t^{s+1}) - \nabla f(\tilde{x}^s)$. From the above inequality, we get

$$\begin{aligned}
\mathrm{E}\left[\|v_t^{s+1}\|^2\right] &\overset{(1)}{\leq} 2\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] + 2\sum_{i=1}^{n} \frac{v_i p_i}{n^2 p_i^2} \left\| \left(\nabla f_i(x_t^{s+1}) - \nabla f_i(\tilde{x}^s) \right) \right\|^2 \\
&\overset{(52),(7)}{\leq} 2\mathrm{E}\left[\|\nabla f(x_t^{s+1})\|^2\right] + \frac{2K}{b}\mathrm{E}\left[\|x_t^{s+1} - \tilde{x}^s\|^2\right].
\end{aligned}$$

□

G. Proofs for SAGA

Lemma 12. *For $c_t, c_{t+1}, \beta > 0$, suppose we have*

$$c_t = c_{t+1}(1 - \tfrac{d}{n} + \eta\beta + 2\frac{K\eta^2}{b}) + \frac{K\eta^2\bar{L}}{b}.$$

Also let η, β and c_{t+1} be chosen such that $\Gamma_t > 0$. Then, the iterates $\{x^t\}$ of Algorithm 6 satisfy the bound

$$\mathrm{E}\left[\|\nabla f(x^t)\|^2\right] \leq \frac{R^t - R^{t+1}}{\Gamma_t},$$

where $R^t := \mathrm{E}\left[f(x^t)\right] + c_t \max_{i\in[n]} \mathrm{E}\left[\|x^t - \alpha_i^t\|^2\right]$.

Proof. Since f is $\bar{L}$-smooth we have

$$\mathrm{E}\left[f(x^{t+1})\right] \leq \mathrm{E}\left[f(x^t) + \langle \nabla f(x^t), x^{t+1} - x^t\rangle + \tfrac{\bar{L}}{2}\|x^{t+1} - x^t\|^2\right].$$

We first note that the update in Algorithm 6 is unbiased i.e., $\mathrm{E}\left[v^t\right] = \nabla f(x^t)$. By using this property of the update on the right hand side of the inequality above, we get the following:

$$\mathrm{E}\left[f(x^{t+1})\right] \leq \mathrm{E}\left[f(x^t) - \eta\|\nabla f(x^t)\|^2 + \tfrac{\bar{L}\eta^2}{2}\|v^t\|^2\right]. \tag{31}$$

Here we used the fact that $x^{t+1} - x^t = -\eta v^t$ (see Algorithm 2). Consider now the Lyapunov function

$$R^t := \mathrm{E}\left[f(x^t)\right] + c_t \max_{i\in[n]} \mathrm{E}\left[\|x^t - \alpha_i^t\|^2\right].$$

For bounding R^{t+1} we need the following:

$$\mathrm{E}\left[\|x^{t+1} - \alpha_i^{t+1}\|^2\right] = \frac{d}{n}\mathrm{E}\left[\|x^{t+1} - x^t\|^2\right] + \frac{n-d}{n}\underbrace{\mathrm{E}\left[\|x^{t+1} - \alpha_i^t\|^2\right]}_{T_1}, \tag{32}$$

The above equality follows from the definition of α_i^{t+1} and the definition of randomness of index j_t in Algorithm 6 and Algorithm 2. The term T_1 in (32) can be bounded as follows

$$\begin{aligned}
T_1 &= \mathrm{E}\left[\|x^{t+1} - x^t + x^t - \alpha_i^t\|^2\right] \\
&= \mathrm{E}\left[\|x^{t+1} - x^t\|^2 + \|x^t - \alpha_i^t\|^2\right] + 2\langle x^{t+1} - x^t, x^t - \alpha_i^t\rangle] \\
&= \mathrm{E}\left[\|x^{t+1} - x^t\|^2 + \|x^t - \alpha_i^t\|^2\right] - 2\eta\mathrm{E}\left[\langle \nabla f(x^t), x^t - \alpha_i^t\rangle\right] \\
&\overset{(54),(55)}{\leq} \mathrm{E}\left[\|x^{t+1} - x^t\|^2 + \|x^t - \alpha_i^t\|^2\right] + 2\eta\mathrm{E}\left[\tfrac{1}{2\beta}\|\nabla f(x^t)\|^2 + \tfrac{1}{2}\beta\|x^t - \alpha_i^t\|^2\right] \\
&\leq \mathrm{E}\left[\|x^{t+1} - x^t\|^2\right] + \max_{i\in[n]}\mathrm{E}\left[\|x^t - \alpha_i^t\|^2\right] + 2\eta\mathrm{E}\left[\tfrac{1}{2\beta}\|\nabla f(x^t)\|^2\right] + \eta\beta\max_{i\in[n]}\mathrm{E}\left[\|x^t - \alpha_i^t\|^2\right].
\end{aligned} \tag{33}$$

The second equality again follows from the unbiasedness of the update of `SAGA`. The last inequality follows from a simple application of Cauchy-Schwarz and Young's inequality. Plugging (31) and (33) into R^{t+1}, we obtain the following bound:

$$\begin{aligned}
R^{t+1} \leq\; & \mathrm{E}\left[f(x^t) - \eta\|\nabla f(x^t)\|^2 + \tfrac{\bar{L}\eta^2}{2}\|v^t\|^2\right] \\
& + \mathrm{E}\left[c_{t+1}\|x^{t+1} - x^t\|^2\right] + c_{t+1}\frac{n-d}{n}\max_{i\in[n]}\mathrm{E}\left[\|x^t - \alpha_i^t\|^2\right] \\
& + \frac{2(n-1)c_{t+1}\eta}{n}\mathrm{E}\left[\tfrac{1}{2\beta}\|\nabla f(x^t)\|^2\right] + \tfrac{1}{2}\beta\max_{i\in[n]}\mathrm{E}\left[\|x^t - \alpha_i^t\|^2\right] \\
\leq\; & \mathrm{E}\left[f(x^t) - \left(\eta - \tfrac{c_{t+1}\eta}{\beta}\right)\|\nabla f(x^t)\|^2\right] + \left(\tfrac{\bar{L}\eta^2}{2} + c_{t+1}\eta^2\right)\mathrm{E}\left[\|v^t\|^2\right] \\
& + \left(\frac{n-d}{n}c_{t+1} + c_{t+1}\eta\beta\right)\max_{i\in[n]}\mathrm{E}\left[\|x^t - \alpha_i^t\|^2\right],
\end{aligned} \tag{34}$$

where we use that $\|x^t - \alpha^t_{i_{\max}}\|^2 \leq \max_{i\in[n]} \|x^t - \alpha^t_i\|^2$ To further bound the quantity in (34), we use Lemma 13 to bound $\mathrm{E}\left[\|v^t\|^2\right]$, so that upon substituting it into (34), we obtain

$$\begin{aligned} R^{t+1} &\overset{(36)}{\leq} \mathrm{E}\left[f(x^t)\right] - \left(\eta - \frac{c_{t+1}\eta}{\beta} - \eta^2 \bar{L} - 2c_{t+1}\eta^2\right) \mathrm{E}\left[\|\nabla f(x^t)\|^2\right] \\ &\quad + \left[c_{t+1}\big(1 - \frac{d}{n} + \eta\beta + 2\frac{K\eta^2}{b}\big) + \frac{K\eta^2\bar{L}}{b}\right] \max_{i\in[n]} \mathrm{E}\left[\|x^t - \alpha^t_i\|^2\right] \\ &\leq R^t - \big(\eta - \tfrac{c_{t+1}\eta}{\beta} - \eta^2\bar{L} - 2c_{t+1}\eta^2\big)\mathrm{E}\left[\|\nabla f(x^t)\|^2\right]. \end{aligned} \tag{35}$$

The second inequality follows from the definition of c_t i.e., $c_t = c_{t+1}\big(1 - \frac{d}{n} + \eta\beta + 2\frac{K\eta^2}{b}\big) + \frac{K\eta^2\bar{L}}{b}$ and R^t specified in the statement, thus concluding the proof. □

The following lemma provides a bound on the variance of the update used in Minibatch `SAGA` algorithm. More specifically, it bounds the quantity $\mathrm{E}\left[\|v^t\|^2\right]$.

Lemma 13. *Let v^t be computed by Algorithm 2. Then,*

$$\mathrm{E}\left[\|v^t\|^2\right] \leq 2\mathrm{E}\left[\|\nabla f(x^t)\|^2\right] + \frac{2K}{b} \max_{i\in[n]} \mathrm{E}\left[\|x^t - \alpha^t_i\|^2\right]. \tag{36}$$

Proof. For ease of exposition, we use the notation

$$\zeta^t_i := \frac{1}{np_i}\left(\nabla f_i(x^t) - \nabla f_i(\alpha^t_i)\right).$$

Using the convexity of $\|\cdot\|^2$ and the definition of v^t we get

$$\begin{aligned} \mathrm{E}\left[\|v^t\|^2\right] &= \mathrm{E}\left[\|\sum_{i\in S_t} \zeta^t_i + \tfrac{1}{n}\sum_{i=1}^n \nabla f(\alpha^t_i)\|^2\right] \\ &= \mathrm{E}\left[\|\sum_{i\in S_t} \zeta^t_i + \tfrac{1}{n}\sum_{i=1}^n \nabla f(\alpha^t_i) - \nabla f(x^t) + \nabla f(x^t)\|^2\right] \\ &\leq 2\mathrm{E}\left[\|\nabla f(x^t)\|^2\right] + 2\mathrm{E}\left[\|\sum_{i\in S_t}\zeta^t_i - \mathrm{E}\left[\zeta^t\right]\|^2\right] \\ &\overset{(1)}{\leq} 2\mathrm{E}\left[\|\nabla f(x^t)\|^2\right] + 2\sum_{i=1}^n \mathrm{E}\left[p_{i_t}\|\zeta^t_{i_t}\|^2\right]. \end{aligned}$$

The first inequality follows from the fact that $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ and that $\mathrm{E}\left[\zeta^t\right] = \nabla f(x^t) - \frac{1}{n}\sum_{i=1}^n \nabla f(\alpha^t_i)$.v

$$\begin{aligned} \mathrm{E}\left[\|v^t\|^2\right] &\leq 2\mathrm{E}\left[\|\nabla f(x^t)\|^2\right] + 2\sum_{i=1}^n \mathrm{E}\left[\frac{p_i}{n^2p_i^2}\|\nabla f_i(x^t) - \nabla f_i(\alpha^t_i)\|^2\right] \\ &\overset{(52),(5)}{\leq} 2\mathrm{E}\left[\|\nabla f(x^t)\|^2\right] + 2\sum_{i=1}^n \mathrm{E}\left[\frac{v_iL_i^2}{n^2p_i}\|x^t - \alpha^t_i\|^2\right] \\ &\overset{(7)}{\leq} 2\mathrm{E}\left[\|\nabla f(x^t)\|^2\right] + \frac{2K}{b}\max_{i\in[n]}\mathrm{E}\left[\|x^t - \alpha^t_i\|^2\right]. \end{aligned} \tag{37}$$

The last inequality follows from L_i-smoothness of f_i and using properties of S sampling, thus concluding the proof. □

PROOF OF THEOREM 19

Proof. We apply telescoping sums to the result of Lemma 12 to obtain

$$\gamma_n \sum_{t=0}^{T-1} \mathrm{E}\left[\|\nabla f(x^t)\|^2\right] \quad \leq \textstyle\sum_{t=0}^{T-1} \Gamma_t \mathrm{E}\left[\|\nabla f(x^t)\|^2\right] \leq R^0 - R^T. \tag{38}$$

The first inequality follows from the definition of γ_n. This inequality in turn implies the bound

$$\sum_{t=0}^{T-1} \mathrm{E}\left[\|\nabla f(x^t)\|^2\right] \leq \frac{\mathrm{E}\left[f(x^0) - f(x^T)\right]}{\gamma_n}, \tag{39}$$

where we used that $R^T = \mathrm{E}\left[f(x^T)\right]$ (since $c_T = 0$), and that $R^0 = \mathrm{E}\left[f(x^0)\right]$ (since $\alpha_i^0 = x^0$ for $i \in [n]$). Using inequality (39), the optimality of x^*, and the definition of x_a in Algorithm 6, we obtain the desired result. □

PROOF OF THEOREM 20 AND THEOREM 4

Proof. With the values of $\mu_3 = 1/3, \nu_3 = 12$ $\eta = b\bar{L}/(3Kn^{2/3})$, $d = b\bar{L}^2/K$ and $\beta = \bar{L}/n^{1/3}$, let us first establish an upper bound on c_t. Let θ denote $\frac{\bar{L}^2 b}{Kn} - \eta\beta - 2K\eta^2/b$. Observe that $\theta < 1$ and $\theta \geq 4\bar{L}^2 b/(9Kn)$. This is due to the specific values of η and β and lower bound of K. Also, we have $c_t = c_{t+1}(1-\theta) + K\eta^2\bar{L}/b$. Using this relationship, it is easy to see that $c_t = K\eta^2\bar{L}\frac{1-(1-\theta)^{T-t}}{b\theta}$. Therefore, we obtain the bound

$$c_t = K\eta^2\bar{L}\frac{1-(1-\theta)^{T-t}}{b\theta} \leq \frac{K\eta^2\bar{L}}{b\theta} \leq \frac{\bar{L}}{4n^{1/3}}, \tag{40}$$

for all $0 \leq t \leq T$, where the inequality follows from the definition of η and the fact that $\theta \geq 4\bar{L}^2 b/(9Kn)$. Using the above upper bound on c_t we can conclude that

$$\gamma_n = \min_t \left(\eta - \frac{c_{t+1}\eta}{\beta} - \eta^2\bar{L} - 2c_{t+1}\eta^2\right) \geq \frac{\bar{L}b}{12Kn^{2/3}},$$

upon using the following inequalities: (i) $c_{t+1}\eta/\beta \leq \eta/4$, (ii) $\eta^2 L \leq \eta/3$ and (iii) $2c_{t+1}\eta^2 \leq \eta/6$, which hold due to the upper bound on c_t in (40) and if $b \leq {}^{K}/{}_{\bar{L}^2} n^{2/3}$. Substituting this bound on γ_n in Theorem 19, we obtain the desired result. □

Theorem 20 is special case with $b = 1$ and $d = 1$.

SARAH-non-convex

This lemmas are modification of lemmas appeared in (Nguyen et al., 2017b) for importance sampling with mini-batch.

Lemma 14. *Consider* `SARAH`*, then we have*

$$\begin{aligned}\sum_{t=0}^{m} \mathrm{E}\left[\|\nabla f(x^t)\|^2\right] \quad &\leq \quad \frac{2}{\eta}[f(x^0) - f(x^*)] + \sum_{t=0}^{m} \mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] \\ &\qquad -(1-\bar{L}\eta)\sum_{t=0}^{m} \mathrm{E}\left[\|v^t\|^2\right],\end{aligned} \tag{41}$$

where x_* is an optimal solution of (1).

Proof. By $\bar{L}$-smoothness of f and $x^{t+1} = x^t - \eta v^t$, we have

$$\begin{aligned}\mathrm{E}\left[f(x^{t+1})\right] \quad &\leq \quad \mathrm{E}\left[f(x^t)\right] - \eta\mathrm{E}\left[\nabla f(x^t)^\top v^t\right] + \frac{\bar{L}\eta^2}{2}\mathrm{E}\left[\|v^t\|^2\right] \\ &= \quad \mathrm{E}\left[f(x^t)\right] - \frac{\eta}{2}\mathrm{E}\left[\|\nabla f(x^t)\|^2\right] + \frac{\eta}{2}\mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] \\ &\qquad - \left(\frac{\eta}{2} - \frac{\bar{L}\eta^2}{2}\right)\mathrm{E}\left[\|v^t\|^2\right],\end{aligned}$$

where the last equality follows from the fact $r^\top q = \frac{1}{2}\left[\|r\|^2 + \|q\|^2 - \|r-q\|^2\right]$, for any $r, q \in \mathbb{R}^d$.

By summing over $t = 0, \ldots, m$, we have

$$\begin{aligned} \mathrm{E}\left[f(x^{m+1})\right] \quad &\leq \quad \mathrm{E}\left[f(x^0)\right] - \frac{\eta}{2}\sum_{t=0}^{m}\mathrm{E}\left[\|\nabla f(x^t)\|^2\right] + \frac{\eta}{2}\sum_{t=0}^{m}\mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] \\ &\quad - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\sum_{t=0}^{m}\mathrm{E}\left[\|v^t\|^2\right], \end{aligned}$$

which is equivalent to ($\eta > 0$):

$$\begin{aligned} \sum_{t=0}^{m}\mathrm{E}\left[\|\nabla f(x^t)\|^2\right] \quad &\leq \quad \frac{2}{\eta}\mathrm{E}\left[f(x^0) - f(x^{m+1})\right] + \sum_{t=0}^{m}\mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] \\ &\quad -(1 - \bar{L}\eta)\sum_{t=0}^{m}\mathrm{E}\left[\|v^t\|^2\right] \\ &\leq \quad \frac{2}{\eta}[f(x^0) - f(x^*)] + \sum_{t=0}^{m}\mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] \\ &\quad -(1 - \bar{L}\eta)\sum_{t=0}^{m}\mathrm{E}\left[\|v^t\|^2\right], \end{aligned}$$

where the last inequality follows since x^* is an optimal solution of (1). (Note that x^0 is given.)

□

Lemma 15. *Consider v^t defined in* SARAH*, then for any $t \geq 1$,*

$$\mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] = \sum_{j=1}^{t}\mathrm{E}\left[\|v^j - v^{j-1}\|^2\right] - \sum_{j=1}^{t}\mathrm{E}\left[\|\nabla f(x^j) - \nabla f(x^{j-1})\|^2\right].$$

Proof. Let $\mathcal{F}_j = \sigma(x^0, i_1, i_2, \ldots, i_{j-1})$ be the σ-algebra generated by $x^0, i_1, i_2, \ldots, i_{j-1}$; $\mathcal{F}_0 = \mathcal{F}_1 = \sigma(x^0)$. Note that $\mathcal{F}_j$ also contains all the information of $x^0, \ldots, x^j$ as well as $v^0, \ldots, v^{j-1}$. For $j \geq 1$, we have

$$\begin{aligned} \mathrm{E}\left[\|\nabla f(x^j) - v^j\|^2 | \mathcal{F}_j\right] \quad &= \quad \mathrm{E}\left[\|[\nabla f(x^{j-1}) - v^{j-1}] + [\nabla f(x^j) - \nabla f(x^{j-1})]\right. \\ &\qquad \left. -[v^j - v^{j-1}]\|^2 | \mathcal{F}_j\right] \\ &= \quad \|\nabla f(x^{j-1}) - v^{j-1}\|^2 + \|\nabla f(x^j) - \nabla f(x^{j-1})\|^2 \\ &\qquad +\mathrm{E}\left[\|v^j - v^{j-1}\|^2 | \mathcal{F}_j\right] \\ &\qquad +2(\nabla f(x^{j-1}) - v^{j-1})^\top(\nabla f(x^j) - \nabla f(x^{j-1})) \\ &\qquad -2(\nabla f(x^{j-1}) - v^{j-1})^\top \mathrm{E}\left[v^j - v^{j-1} | \mathcal{F}_j\right] \\ &\qquad -2(\nabla f(x^j) - \nabla f(x^{j-1}))^\top \mathrm{E}\left[v^j - v^{j-1} | \mathcal{F}_j\right] \\ &= \quad \|\nabla f(x^{j-1}) - v^{j-1}\|^2 - \|\nabla f(x^j) - \nabla f(x^{j-1})\|^2 \\ &\qquad +\mathrm{E}\left[\|v^j - v^{j-1}\|^2 | \mathcal{F}_j\right], \end{aligned}$$

where the last equality follows from

$$\begin{aligned} \mathrm{E}\left[v^j - v^{j-1} | \mathcal{F}_j\right] \quad &= \quad \mathrm{E}\left[\sum_{i \in I_j}\frac{1}{np_i}\nabla f_i(x^j) - \nabla f_i(x^{j-1})] \Big| \mathcal{F}_j\right] \\ &= \quad \sum_{i=1}^{n}\frac{p_i}{np_i}[\nabla f_i(x^j) - \nabla f_i(x^{j-1})] = \nabla f(x^j) - \nabla f(x^{j-1}). \end{aligned}$$

By taking expectation for the above equation, we have

$$\begin{aligned}\mathrm{E}\left[\|\nabla f(x^j)-v^j\|^2\right] &= \mathrm{E}\left[\|\nabla f(x^{j-1})-v^{j-1}\|^2\right]-\mathrm{E}\left[\|\nabla f(x^j)-\nabla f(x^{j-1})\|^2\right]\\ &\quad+\mathrm{E}\left[\|v^j-v^{j-1}\|^2\right].\end{aligned}$$

Note that $\|\nabla f(x^0)-v^0\|^2=0$. By summing over $j=1,\dots,t$ $(t\geq 1)$, we have

$$\mathrm{E}\left[\|\nabla f(x^t)-v^t\|^2\right]=\sum_{j=1}^{t}\mathrm{E}\left[\|v^j-v^{j-1}\|^2\right]-\sum_{j=1}^{t}\mathrm{E}\left[\|\nabla f(x^j)-\nabla f(x^{j-1})\|^2\right].$$

□

With the above Lemmas, we can derive the following upper bound for $\mathrm{E}\left[\|\nabla f(x^t)-v^t\|^2\right]$.

Lemma 16. *Consider v^t defined in* `SARAH`*. Then for any $t\geq 1$,*

$$\mathrm{E}\left[\|\nabla f(x^t)-v^t\|^2\right]\leq\frac{1}{b}K\eta^2\sum_{j=1}^{t}\mathrm{E}\left[\|v^{j-1}\|^2\right].$$

Proof. Let

$$\xi_t=\frac{1}{np_i}\left(\nabla f_t(x^j)-\nabla f_t(x^{j-1})\right) \tag{42}$$

We have

$$\begin{aligned}&\mathrm{E}\left[\|v^j-v^{j-1}\|^2|\mathcal{F}_j\right]-\|\nabla f(x^j)-\nabla f(x^{j-1})\|^2\\
=\;&\mathrm{E}\left[\Big\|\sum_{i\in I_j}\frac{1}{np_i}[\nabla f_i(x^j)-\nabla f_i(x^{j-1})]\Big\|^2\Big|\mathcal{F}_j\right]-\Big\|\frac{1}{n}\sum_{i=1}^{n}[\nabla f_i(x^j)-\nabla f_i(x^{j-1})]\Big\|^2\\
=\;&\mathrm{E}\left[\Big\|\sum_{i\in I_j}\xi_i\Big\|^2\Big|\mathcal{F}_j\right]-\Big\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\Big\|^2\\
\overset{(1)}{\leq}\;&\sum_{i=1}^{n}v_ip_i\|\xi_i\|^2\\
=\;&\sum_{i=1}^{n}\frac{v_ip_i}{p_i^2n^2}\|\nabla f_i(x^j)-\nabla f_i(x^{j-1})\|^2\\
\overset{(52),(7)}{\leq}\;&\frac{1}{b}K\eta^2\|v^{j-1}\|^2.\end{aligned}$$

Hence, by taking expectation, we have

$$\mathrm{E}\left[\|v^j-v^{j-1}\|^2\right]-\mathrm{E}\left[\|\nabla f(x^j)-\nabla f(x^{j-1})\|^2\right]\leq\frac{1}{b}K\eta^2\mathrm{E}\left[[\|v^{j-1}\|^2\right].$$

By Lemma 15, for $t\geq 1$,

$$\begin{aligned}\mathrm{E}\left[\|\nabla f(x^t)-v^t\|^2\right] &= \sum_{j=1}^{t}\mathrm{E}\left[\|v^j-v^{j-1}\|^2\right]-\sum_{j=1}^{t}\mathrm{E}\left[\|\nabla f(x^j)-\nabla f(x^{j-1})\|^2\right]\\ &\leq \frac{1}{b}K\eta^2\sum_{j=1}^{t}\mathrm{E}\left[\|v^{j-1}\|^2\right].\end{aligned}$$

This completes the proof. □

PROOF OF THEOREM 5

Proof. By Lemma 16, we have

$$\mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] \leq \frac{1}{b}K\eta^2 \sum_{j=1}^{t} \mathrm{E}\left[\|v^{j-1}\|^2\right].$$

Note that $\|\nabla f(x^0) - v^0\|^2 = 0$. Hence, by summing over $t = 0, \ldots, m$ ($m \geq 1$), we have

$$\begin{aligned} \sum_{t=0}^{m} \mathrm{E}\left[\|v^t - \nabla f(x^t)\|^2\right] \quad \leq \quad & \frac{1}{b}K\eta^2 \Big[m\mathrm{E}\left[\|v^0\|^2\right] \\ & +(m-1)\mathrm{E}\left[\|v^1\|^2\right] + \cdots + \mathrm{E}\left[\|v^{m-1}\|^2\right] \Big]. \end{aligned} \tag{43}$$

We have

$$\begin{aligned} & \sum_{t=0}^{m} \mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] - (1 - \bar{L}\eta) \sum_{t=0}^{m} \mathrm{E}\left[\|v^t\|^2\right] \\ & \leq \frac{1}{b}K\eta^2 \Big[m\mathrm{E}\left[\|v^0\|^2\right] + (m-1)\mathrm{E}\left[\|v^1\|^2\right] + \cdots + \mathrm{E}\left[\|v^{m-1}\|^2\right] \Big] \\ & \quad -(1 - \bar{L}\eta) \Big[\mathrm{E}\left[\|v^0\|^2\right] + \mathrm{E}\left[\|v^1\|^2\right] + \cdots + \mathrm{E}\left[\|v^m\|^2\right] \Big] \\ & \leq \left[\frac{1}{b}K\eta^2 m - (1 - \bar{L}\eta) \right] \sum_{t=1}^{m} \mathrm{E}\left[\|v^{t-1}\|^2\right] \overset{(10)}{\leq} 0 \end{aligned} \tag{44}$$

since

$$\eta = \frac{2}{\bar{L}\left(\sqrt{1 + \frac{4Km}{\bar{L}^2 b}} + 1\right)}$$

is a root of equation

$$\frac{1}{b}K\eta^2 m - (1 - \bar{L}\eta) = 0.$$

Therefore, by Lemma 14, we have

$$\begin{aligned} \sum_{t=0}^{m} \mathrm{E}\left[\|\nabla f(x^t)\|^2\right] \quad \leq \quad & \frac{2}{\eta}[f(x^0) - f(x^*)] + \sum_{t=0}^{m} \mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] \\ & -(1 - \bar{L}\eta) \sum_{t=0}^{m} \mathrm{E}\left[\|v^t\|^2\right] \\ \overset{(44)}{\leq} \tfrac{2}{\eta} \quad & [f(x^0) - f(x^*)]. \end{aligned}$$

If x_a is chosen uniformly at random from $\{x^t\}_{t=0}^m$, then

$$\mathrm{E}\left[\|\nabla f(x_a)\|^2\right] = \frac{1}{m+1} \sum_{t=0}^{m} \mathrm{E}\left[\|\nabla f(x^t)\|^2 \right] \leq \frac{2}{\eta(m+1)}[f(x^0) - f(x^*)] \Bigg].$$

This concludes the proof. □

H. One Sample Importance Sampling

H.1. SVRG

In this section, we introduce SVRG algorithm with batch size equal to 1.

Theorem 17. *Let $c_m = 0$, $\eta = \eta > 0$, $\beta = \beta > 0$, and $c_t = c_{t+1}(1 + \eta\beta + 2K\eta^2) + K\eta^2\bar{L}$ such that $\Gamma_t > 0$ for $0 \le t \le m-1$. Define the quantity $\gamma_n := \min_t \Gamma_t$. Further, let T be a multiple of m. Then for the output x_a of Algorithm 5 we have*

$$\mathrm{E}\left[\|\nabla f(x_a)\|^2\right] \le \frac{f(x^0) - f(x^*)}{T\gamma_n}, \tag{45}$$

where x^* is an optimal solution to (1) and $\Gamma_t = \left(\eta - \frac{c_{t+1}\eta}{\beta} - \eta^2\bar{L} - 2c_{t+1}\eta^2\right)$.

Theorem 18. *Let $\eta = \bar{L}\mu_0/(Kn^{\frac{2}{3}})$ $(0 < \mu_0 < 1)$, $\beta = \bar{L}/n^{\frac{1}{3}}$, $m = \lfloor Kn/(3\bar{L}^2\mu_0)\rfloor$ and T is some multiple of m. Then there exists universal constants $\mu_0, \nu > 0$ such that we have the following: $\gamma_n \ge \frac{\bar{L}\nu}{K} n^{\frac{2}{3}}$ in Theorem 17 and*

$$\mathrm{E}\left[\|\nabla f(x_a)\|^2\right] \quad \le \frac{Kn^{\frac{2}{3}}[f(x^0) - f(x^*)]}{\bar{L}T\nu}, \tag{46}$$

where x^* is an optimal solution to the problem in (1) and x_a is the output of Algorithm 5.

Comparing Theorem 17 to the previous result in (Reddi et al., 2016a), we can see improvement in constant, if we assume different L_i-smooth constants for different functions. If the all L_i's are the same then our result is the same as previous result for uniform sampling, because then $\alpha = \frac{n-1}{n-1} = 1$.

H.2. SAGA

Here, we provide similar analysis as for SVRG with the same result. We provide more generalized improved form of theorems which appeared in (Reddi et al., 2016b).

Theorem 19. *Let $c_T = 0$, $\beta > 0$, and $c_t = c_{t+1}(1 - \frac{1}{n} + \eta\beta + 2K\eta^2) + K\eta^2\bar{L}$ be such that $\Gamma_t > 0$ for $0 \le t \le T-1$. Define the quantity $\gamma_n := \min_{0 \le t \le T-1} \Gamma_t$. Then the output x_a of Algorithm 6 satisfies the bound*

$$\mathrm{E}\left[\|\nabla f(x_a)\|^2\right] \le \frac{f(x^0) - f(x^*)}{T\gamma_n},$$

where x^* is an optimal solution to (1) and $\Gamma_t = \left(\eta - \frac{c_{t+1}\eta}{\beta} - \eta^2\bar{L} - 2c_{t+1}\eta^2\right)$.

Theorem 20. *Let $\eta = \bar{L}/(3Kn^{2/3})$ and $\beta = \bar{L}/n^{1/3}$. Then, $\gamma_n \ge \frac{\bar{L}}{12Kn^{2/3}}$ and we have the bound*

$$\mathrm{E}\left[\|\nabla f(x_a)\|^2\right] \quad \le \frac{12Kn^{2/3}[f(x^0) - f(x^*)]}{\bar{L}T},$$

where x^* is an optimal solution to the problem in (1) and x_a is the output of Algorithm 6.

We can see that exactly same conclusions apply here as for SVRG and results can be interpreted in the same way.

Algorithm 5 SVRG$\left(x^0, T, m, \{p_i\}_{i=0}^n, \eta\right)$

```
Input: $\tilde{x}^0 = x_m^0 = x^0 \in \mathbb{R}^d$, epoch length $m$, step sizes $\{\eta_i > 0\}_{i=0}^{m-1}$, $S = \lceil T/m \rceil$
for $s = 0$ to $S-1$ do
    $x_0^{s+1} = x_m^s$
    $g^{s+1} = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\tilde{x}^s)$
    for $t = 0$ to $m-1$ do
        With $\{p_i\}_{i=0}^n$ randomly pick $i_t$ from $\{1, \ldots, n\}$
        $v_t^{s+1} = \frac{1}{np_{i_t}}(\nabla f_{i_t}(x_t^{s+1}) - \nabla f_{i_t}(\tilde{x}^s)) + g^{s+1}$
        $x_{t+1}^{s+1} = x_t^{s+1} - \eta v_t^{s+1}$
    end for
    $\tilde{x}^{s+1} = x_m^{s+1}$
end for
Output: Iterate $x_a$ chosen uniformly random from $\{\{x_t^{s+1}\}_{t=0}^m\}_{s=0}^S$.
```

Algorithm 6 SAGA$\big(x^0, T, \{p_i\}_{i=0}^n, \eta\big)$

```
Input: $x^0 \in \mathbb{R}^d$, $\alpha_i^0 = x^0$ for $i \in [n]$, number of iterations $T$, step size $\eta > 0$
$g^0 = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\alpha_i^0)$
for $t = 0$ to $T - 1$ do
    Randomly pick $i_t$ from $[n]$ with $\{p_i\}_{i=0}^n$
    Randomly uniformly pick $i_t$ from $[n]$
    $v^t = \frac{1}{np_{i_t}}(\nabla f_{i_t}(x^t) - \nabla f_{i_t}(\alpha_{i_t}^t)) + g^t$
    $x^{t+1} = x^t - \eta v^t$
    $\alpha_{j_t}^{t+1} = x^t$ and $\alpha_j^{t+1} = \alpha_j^t$ for $j \neq j_t$
    $g^{t+1} = g^t - \frac{1}{n}(\nabla f_{j_t}(\alpha_{j_t}^t) - \nabla f_{j_t}(\alpha_{j_t}^{t+1}))$
end for
Output: Iterate $x_a$ chosen uniformly random from $\{x^t\}_{t=0}^T$.
```

I. SARAH : Convex Case

I.1. Main result

Consider Algorithm 7, which is an arbitrary sampling variant of the SARAH method..

Algorithm 7 SARAH

```
1:  Parameters: the learning rate $\eta > 0$ and the inner loop size $m$.
2:  Initialize: $\tilde{x}_0$
3:  Iterate:
4:  for $s = 1, 2, \dots$ do
5:      $x_0 = \tilde{x}_{s-1}$
6:      $v^0 = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x^0)$
7:      $x_1 = x_0 - \eta v^0$
8:      Iterate:
9:      for $t = 1, \dots, m - 1$ do
10:         Sample $i_t$ at random from $[n]$ with probability $\{p_i\}_{i=1}^n$
11:         $v^t = \frac{1}{np_i}(\nabla f_{i_t}(x^t) - \nabla f_{i_t}(x^{t-1})) + v^{t-1}$
12:         $x_{t+1} = x^t - \eta v^t$
13:     end for
14:     Set $\tilde{x}_s = x^t$ with $t$ chosen uniformly at random from $\{0, 1, \dots, m\}$
15: end for
```

Note, that only 10-th and 11-th row are changed comparing to classic SARAH algorithm presented in (Nguyen et al., 2017a). We do not sample uniformly anymore and also in the 11-th row of Algorithm 7, where we use factor $\frac{1}{np_i}$ in order to stay unbiased in outer cycle.

Then using similar analysis used in (Nguyen et al., 2017a) and additional lemmas we can prove following theorems with p_i in Algorithm 7 to be $\frac{L_i}{\sum_{j=1}^n L_i}$

Theorem 21. *Suppose that $f_i(x)$ are L_i-smooth and convex, $f(x)$ is μ strongly convex. Consider v^t defined in* SARAH *(Algorithm 7) with $\eta < 2/\bar{L}$, where $\bar{L} = \frac{1}{n}\sum_{j=1}^n L_i$. Then, for any $t \geq 1$,*

$$\begin{aligned}
\mathrm{E}\left[\|v^t\|^2\right] \quad &\leq \quad \left[1 - \left(\tfrac{2}{\eta\bar{L}} - 1\right)\mu^2\eta^2\right] \mathrm{E}\left[\|v^{t-1}\|^2\right] \\
&\leq \quad \left[1 - \left(\tfrac{2}{\eta\bar{L}} - 1\right)\mu^2\eta^2\right]^t \mathrm{E}\left[\|\nabla f(x^0)\|^2\right].
\end{aligned}$$

By choosing $\eta = \mathcal{O}(1/\bar{L})$, we obtain the linear convergence of $\|v^t\|^2$ in expectation with the rate $(1 - 1/\kappa^2)$, where $\kappa = \frac{\bar{L}}{\mu}$ is condition number, This is improvement over previous result in (Nguyen et al., 2017a), because of $\frac{\bar{L}}{\mu} \leq \frac{L_{\max}}{\mu}$. Below we show that a better convergence rate could be obtained under a stronger convexity assumption for each single $f_i(x)$.

Theorem 22. *Suppose that $f_i(x)$ are L_i-smooth and μ strongly convex. Consider v^t defined by in `SARAH` (Algorithm 7) with $\eta \le 2/(\mu+\bar{L})$. Then the following bound holds, $\forall\, t \ge 1$,*

$$\begin{aligned} \mathrm{E}\left[\|v^t\|^2\right] &\le \left(1 - \tfrac{2\mu\bar{L}\eta}{\mu+\bar{L}}\right)\mathrm{E}\left[\|v^{t-1}\|^2\right] \\ &\le \left(1 - \tfrac{2\mu\bar{L}\eta}{\mu+\bar{L}}\right)^t \mathrm{E}\left[\|\nabla f(x^0)\|^2\right]. \end{aligned}$$

By setting $\eta = \mathcal{O}(1/\bar{L})$, we derive the linear convergence with the rate of $(1-1/\kappa)$, where $\hat{\kappa} = \frac{\bar{L}}{\mu}$ which is an improvement over the previous result of (Nguyen et al., 2017a), because if we take the optimal stepsize $\nu = \frac{2}{\mu+\bar{L}}$ than we can easily prove that $\frac{2\mu\bar{L}\eta}{\mu+\bar{L}}$ is greater than $\frac{2\mu L_{\max}\eta}{\mu+L_{\max}}$, with optimal step size, where $L_{\max} = \max_i\{L_i\}$.

I.2. Lemmas

We start with modification of lemmas in (Nguyen et al., 2017a), which we later use in the proofs of Theorem 22 and Theorem 21. The first Lemma 23 bounds the sum of expected values of $\|\nabla f(x^t)\|^2$. The second, Lemma 24, bounds $\mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right]$.

Lemma 23. *Suppose that $f_i(x)$'s are L_i-smooth. Consider `SARAH` (Algorithm 7). Then, we have*

$$\begin{aligned} \sum_{t=0}^{m} \mathrm{E}\left[\|\nabla f(x^t)\|^2\right] &\le \frac{2}{\eta}\mathrm{E}\left[f(x^0) - f(x^*)\right] \\ &+ \sum_{t=0}^{m} \mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] - (1-\bar{L}\eta)\sum_{t=0}^{m}\mathrm{E}\left[\|v^t\|^2\right]. \end{aligned} \tag{47}$$

Lemma 24. *Suppose that $f_i(x)$'s are L_i-smooth. Consider `SARAH` (Algorithm 7). Then for any $t \ge 1$,*

$$\mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] = \sum_{j=1}^{t}\mathrm{E}\left[\|v^j - v^{j-1}\|^2\right] - \sum_{j=1}^{t}\mathrm{E}\left[\|\nabla f(x^j) - \nabla f(x^{j-1})\|^2\right].$$

Lemma 25. *Suppose that $f_i(x)$'s are L_i-smooth and convex. Consider `SARAH` (Algorithm 7) with $\eta < 2/\bar{L}$. Then we have that for any $t \ge 1$,*

$$\begin{aligned} \mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] &\le \frac{\eta\bar{L}}{2-\eta\bar{L}}\left[\mathrm{E}\left[\|v^0\|^2\right] - \mathrm{E}\left[\|v^t\|^2\right]\right] \\ &\le \frac{\eta\bar{L}}{2-\eta\bar{L}}\mathrm{E}\left[\|v^0\|^2\right], \end{aligned}$$

where $\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L_i$.

PROOF OF LEMMA 23

Proof. By Lemma 26 and $x^{t+1} = x^t - \eta v^t$, we have

$$\begin{aligned} \mathrm{E}\left[f(x^{t+1})\right] &\le \mathrm{E}\left[f(x^t)\right] - \eta\mathrm{E}\left[\nabla f(x^t)^\top v^t\right] + \frac{\bar{L}\eta^2}{2}\mathrm{E}\left[\|v^t\|^2\right] \\ &= \mathrm{E}\left[f(x^t)\right] - \frac{\eta}{2}\mathrm{E}\left[\|\nabla f(x^t)\|^2\right] + \frac{\eta}{2}\mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] \\ &\quad - \left(\frac{\eta}{2} - \frac{\bar{L}\eta^2}{2}\right)\mathrm{E}\left[\|v^t\|^2\right], \end{aligned}$$

where the last equality follows from the fact $a^\top b = \frac{1}{2}\left[\|a\|^2 + \|b\|^2 - \|a-b\|^2\right]$.

By summing over $t = 0, \dots, m$, we have

$$\begin{aligned}\mathrm{E}\left[f(x_{m+1})\right] &\leq \mathrm{E}\left[f(x^0)\right] - \frac{\eta}{2}\sum_{t=0}^{m}\mathrm{E}\left[\|\nabla f(x^t)\|^2\right] + \frac{\eta}{2}\sum_{t=0}^{m}\mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] \\ &\quad - \left(\frac{\eta}{2} - \frac{\bar{L}\eta^2}{2}\right)\sum_{t=0}^{m}\mathrm{E}\left[\|v^t\|^2\right],\end{aligned}$$

which is equivalent to ($\eta > 0$):

$$\begin{aligned}\sum_{t=0}^{m}\mathrm{E}\left[\|\nabla f(x^t)\|^2\right] &\leq \frac{2}{\eta}\mathrm{E}\left[f(x^0) - f(x_{m+1})\right] + \\ &\quad \sum_{t=0}^{m}\mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] - (1 - \bar{L}\eta)\sum_{t=0}^{m}\mathrm{E}\left[\|v^t\|^2\right] \\ &\leq \frac{2}{\eta}\mathrm{E}\left[f(x^0) - f(x^*)\right] + \sum_{t=0}^{m}\mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] \\ &\quad -(1 - \bar{L}\eta)\sum_{t=0}^{m}\mathrm{E}\left[\|v^t\|^2\right],\end{aligned}$$

where the last inequality follows since x^* is a global minimizer of (1). □

PROOF OF LEMMA 24

Proof. Let $\mathcal{F}_j$ be σ algebra that contains all the information of $x^0, \dots, x^j$ as well as $v^0, \dots, v^{j-1}$. For $j \geq 1$, we have

$$\begin{aligned}&\mathrm{E}\left[\|\nabla f(x^j) - v^j\|^2 | \mathcal{F}_j\right] = \\ &\mathrm{E}\left[\|[\nabla f(x^{j-1}) - v^{j-1}] + [\nabla f(x^j) - \nabla f(x^{j-1})] - [v^j - v^{j-1}]\|^2 | \mathcal{F}_j\right] \\ &= \|\nabla f(x^{j-1}) - v^{j-1}\|^2 + \|\nabla f(x^j) - \nabla f(x^{j-1})\|^2 \\ &\quad + \mathrm{E}\left[\|v^j - v^{j-1}\|^2 | \mathcal{F}_j\right] \\ &\quad + 2(\nabla f(x^{j-1}) - v^{j-1})^\top(\nabla f(x^j) - \nabla f(x^{j-1})) \\ &\quad - 2(\nabla f(x^{j-1}) - v^{j-1})^\top\mathrm{E}\left[v^j - v^{j-1} | \mathcal{F}_j\right] \\ &\quad - 2(\nabla f(x^j) - \nabla f(x^{j-1}))^\top\mathrm{E}\left[v^j - v^{j-1} | \mathcal{F}_j\right] \\ &= \|\nabla f(x^{j-1}) - v^{j-1}\|^2 - \|\nabla f(x^j) - \nabla f(x^{j-1})\|^2 \\ &\quad + \mathrm{E}\left[\|v^j - v^{j-1}\|^2 | \mathcal{F}_j\right],\end{aligned}$$

where the last equality follows from

$$\mathrm{E}\left[v^j - v^{j-1} | \mathcal{F}_j\right] = \mathrm{E}\left[\tfrac{1}{np_{i_j}}\left(\nabla f_{i_j}(x^j) - \nabla f_{i_j}(x^{j-1})\right) | \mathcal{F}_j\right] = \nabla f(x^j) - \nabla f(x^{j-1}).$$

By taking expectation for the above equation, we have

$$\begin{aligned}\mathrm{E}\left[\|\nabla f(x^j) - v^j\|^2\right] &= \mathrm{E}\left[\|\nabla f(x^{j-1}) - v^{j-1}\|^2\right] - \mathrm{E}\left[\|\nabla f(x^j) - \nabla f(x^{j-1})\|^2\right] \\ &\quad + \mathrm{E}\left[\|v^j - v^{j-1}\|^2\right].\end{aligned}$$

Note that $\|\nabla f(x^0) - v^0\|^2 = 0$. By summing over $j = 1, \dots, t$ ($t \geq 1$), we have

$$\mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] = \textstyle\sum_{j=1}^{t}\mathrm{E}\left[\|v^j - v^{j-1}\|^2\right] - \sum_{j=1}^{t}\mathrm{E}\left[\|\nabla f(x^j) - \nabla f(x^{j-1})\|^2\right].$$

□

PROOF OF LEMMA 25

Proof. For $j \geq 1$, we have

$$\begin{aligned}
\mathrm{E}\left[\|v^j\|^2|\mathcal{F}_j\right] &= \mathrm{E}\left[\|v^{j-1} - \frac{1}{np_{i_j}}(\nabla f_{i_j}(x^{j-1}) - \nabla f_{i_j}(x^j))\|^2|\mathcal{F}_j\right] \\
&= \|v^{j-1}\|^2 + \mathrm{E}\left[\frac{1}{n^2p_{i_j}^2}\|\nabla f_{i_j}(x^{j-1}) - \nabla f_{i_j}(x^j)\|^2|\mathcal{F}_j\right] \\
&\quad -\mathrm{E}\left[\tfrac{2}{\eta np_{i_j}}(\nabla f_{i_j}(x^{j-1}) - \nabla f_{i_j}(x^j))^\top(x^{j-1} - x^j)|\mathcal{F}_j\right] \\
&\overset{(52)}{\leq} \|v^{j-1}\|^2 + \mathrm{E}\left[\frac{1}{n^2p_{i_j}^2}\|\nabla f_{i_j}(x^{j-1}) - \nabla f_{i_j}(x^j)\|^2|\mathcal{F}_j\right] \\
&\quad -\mathrm{E}\left[\tfrac{2}{L_{i_j}\eta np_{i_j}}\|\nabla f_{i_j}(x^{j-1}) - \nabla f_{i_j}(x^j)\|^2|\mathcal{F}_j\right] \\
&= \|v^{j-1}\|^2 + \left(1 - \tfrac{2}{\eta\bar{L}}\right)\mathrm{E}\left[\left\|\frac{1}{np_{i_j}}\left(\nabla f_{i_j}(x^{j-1}) - \nabla f_{i_j}(x^j)\right)\right\|^2|\mathcal{F}_j\right] \\
&= \|v^{j-1}\|^2 + \left(1 - \tfrac{2}{\eta\bar{L}}\right)\mathrm{E}\left[\|v^j - v^{j-1}\|^2|\mathcal{F}_j\right],
\end{aligned}$$

The consequent equality follows from definition of p_i's and the last equality follows from definition of `SARAH` . Taking expectation, we get

$$\mathrm{E}\left[\|v^j - v^{j-1}\|^2\right] \leq \frac{\eta\bar{L}}{2-\eta\bar{L}}\left[\mathrm{E}\left[\|v^{j-1}\|^2\right] - \mathrm{E}\left[\|v^j\|^2\right]\right],$$

when $\eta < 2/\bar{L}$.

By summing the above inequality over $j = 1, \dots, t$ $(t \geq 1)$, we have

$$\sum_{j=1}^t \mathrm{E}\left[\|v^j - v^{j-1}\|^2\right] \leq \frac{\eta\bar{L}}{2-\eta\bar{L}}\left[\mathrm{E}\left[\|v^0\|^2\right] - \mathrm{E}\left[\|v^t\|^2\right]\right]. \tag{48}$$

By Lemma 24, we have

$$\mathrm{E}\left[\|\nabla f(x^t) - v^t\|^2\right] \leq \textstyle\sum_{j=1}^t \mathrm{E}\left[\|v^j - v^{j-1}\|^2\right] \overset{(48)}{\leq} \frac{\eta\bar{L}}{2-\eta\bar{L}}\left[\mathrm{E}\left[\|v^0\|^2\right] - \mathrm{E}\left[\|v^t\|^2\right]\right].$$

□

PROOF OF THEOREM 21

Proof. For $t \geq 1$, we have

$$\begin{aligned}
\|\nabla f(x^t) - \nabla f(x^{t-1})\|^2 &= \left\|\frac{1}{n}\sum_{i=1}^n[\nabla f_i(x^t) - \nabla f_i(x^{t-1})]\right\|^2 \\
&= \left\|\sum_{i=1}^n p_i\frac{1}{np_i}[\nabla f_i(x^t) - \nabla f_i(x^{t-1})]\right\|^2 \\
&\overset{(56)}{\leq} \sum_{i=1}^n p_i\|\tfrac{1}{np_i}\left(\nabla f_i(x^t) - \nabla f_i(x^{t-1})\right)\|^2 \\
&= \mathrm{E}\left[\left\|\tfrac{1}{np_i}\left(\nabla f_{i_t}(x^t) - \nabla f_{i_t}(x^{t-1})\right)\right\|^2|\mathcal{F}_t\right].
\end{aligned} \tag{49}$$

Using the proof of Lemma 25, for $t \geq 1$, we have

$$\begin{aligned}
\mathrm{E}\left[\|v^t\|^2 | \mathcal{F}_t\right] &\leq \|v^{t-1}\|^2 + \left(1 - \tfrac{2}{\eta \bar{L}}\right) \mathrm{E}\left[\|\nabla f_{i_t}(x^{t-1}) - \nabla f_{i_t}(x^t)\|^2 | \mathcal{F}_t\right] \\
&\overset{(49)}{\leq} \|v^{t-1}\|^2 + \left(1 - \tfrac{2}{\eta \bar{L}}\right) \|\nabla f(x^t) - \nabla f(x^{t-1})\|^2 \\
&\leq \|v^{t-1}\|^2 + \left(1 - \tfrac{2}{\eta \bar{L}}\right) \mu^2 \eta^2 \|v^{t-1}\|^2.
\end{aligned}$$

Note that $1 - \frac{2}{\eta \bar{L}} < 0$ since $\eta < 2/\bar{L}$. The last inequality follows by the strong convexity of f, that is, $\mu\|x^t - x^{t-1}\| \leq \|\nabla f(x^t) - \nabla f(x^{t-1})\|$ and the fact that $x^t = x^{t-1} - \eta v^{t-1}$. By taking the expectation and applying recursively, we have

$$\begin{aligned}
\mathrm{E}\left[\|v^t\|^2\right] &\leq \left[1 - \left(\tfrac{2}{\eta \bar{L}} - 1\right) \mu^2 \eta^2\right] \mathrm{E}\left[\|v^{t-1}\|^2\right] \\
&\leq \left[1 - \left(\tfrac{2}{\eta \bar{L}} - 1\right) \mu^2 \eta^2\right]^t \mathrm{E}\left[\|v^0\|^2\right] \\
&= \left[1 - \left(\tfrac{2}{\eta \bar{L}} - 1\right) \mu^2 \eta^2\right]^t \mathrm{E}\left[\|\nabla f(x^0)\|^2\right].
\end{aligned}$$

□

PROOF OF THEOREM 22

Proof. We obviously have $\mathrm{E}\left[\|v^0\|^2 | \mathcal{F}_0\right] = \|\nabla f(x_0)\|^2$. For $t \geq 1$, we have

$$\begin{aligned}
\mathrm{E}\left[\|v^t\|^2 | \mathcal{F}_t\right] &= \mathrm{E}\left[\|v^{t-1} - \tfrac{1}{np_{i_t}}(\nabla f_{i_t}(x^{t-1}) - \nabla f_{i_t}(x^t))\|^2 | \mathcal{F}_t\right] \\
&= \|v^{t-1}\|^2 + \mathrm{E}\left[\tfrac{1}{n^2 p_{i_t}^2}\|\nabla f_{i_t}(x^{t-1}) - \nabla f_{i_t}(x^t)\|^2 \mathcal{F}_t\right] \\
&\quad - \mathrm{E}\left[\tfrac{2}{\eta n p_{i_t}}(\nabla f_{i_t}(x^{t-1}) - \nabla f_{i_t}(x^t))^\top (x^{t-1} - x^t) | \mathcal{F}_t\right] \\
&= \|v^{t-1}\|^2 + \mathrm{E}\left[\|\tfrac{1}{np_{i_t}}\left(\nabla f_{i_t}(x^{t-1}) - \nabla f_{i_t}(x^t)\right)\|^2 | \mathcal{F}_t\right] \\
&\quad - \tfrac{2}{\eta}(\nabla f(x^{t-1}) - \nabla f(x^t))^\top (x^{t-1} - x^t) \\
&\overset{(53),(51)}{\leq} \|v^{t-1}\|^2 + \mathrm{E}\left[\|\tfrac{1}{np_{i_t}}\left(\nabla f_{i_t}(x^{t-1}) - \nabla f_{i_t}(x^t)\right)\|^2 | \mathcal{F}_t\right] \\
&\quad - \tfrac{2\mu\bar{L}\eta}{\mu+\bar{L}}\|v^{t-1}\|^2 - \tfrac{2}{\eta(\mu+\bar{L})}\|\nabla f_{i_t}(x^{t-1}) - \nabla f_{i_t}(x^t)\|^2 \\
&\leq \left(1 - \tfrac{2\mu\bar{L}\eta}{\mu+\bar{L}}\right)\|v^{t-1}\|^2 \\
&\quad + \mathrm{E}\left[\|\tfrac{1}{np_{i_t}}\left(\nabla f_{i_t}(x^{t-1}) - \nabla f_{i_t}(x^t)\right)\|^2 | \mathcal{F}_t\right] - \|\nabla f(x^{t-1}) - \nabla f(x^t)\|^2 \\
&= \left(1 - \tfrac{2\mu\bar{L}\eta}{\mu+\bar{L}}\right)\|v^{t-1}\|^2 \\
&\quad - \mathrm{E}\left[\|\tfrac{1}{np_{i_t}}\left(\nabla f_{i_t}(x^{t-1}) - \nabla f_{i_t}(x^t)\right) - \nabla f(x^{t-1}) - \nabla f(x^t)\|^2 | \mathcal{F}_t\right] \\
&\leq \left(1 - \tfrac{2\mu\bar{L}\eta}{\mu+\bar{L}}\right)\|v^{t-1}\|^2, \qquad (50)
\end{aligned}$$

where in the first two equalities, we used definition of SARAH . The first inequality follows from fact that $f(x)$ is $\bar{L}$-smooth and μ strongly convex, thus following inequality holds (inequality from (Nesterov, 2013))

$$(\nabla f(x) - \nabla f(x'))^\top (x - x') \geq \frac{\mu \bar{L}}{\mu + \bar{L}}\|x - x'\|^2 + \frac{1}{\mu + \bar{L}}\|\nabla f(x) - \nabla f(x')\|^2, \qquad (51)$$

The second one uses assumption that $\eta \leq \frac{2}{\mu+\bar{L}}$, thus $\eta = \frac{2}{\mu+\bar{L}}$ is optimal step size under this analysis. By taking the expectation and applying recursively, the desired result is achieved. □

J. Technical Lemmas

Lemma 26. *Let f_i's be function, which are L_i-smooth, then $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ is $\bar{L}$-smooth, where $\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L_i$.*

Proof. For each function f_i we have by definition of L_i-smoothness, $\forall x, y \in \mathbb{R}^d$

$$f_i(x) \leq f_i(y) + \nabla f_i(y)^\top (x - y) + \frac{L_i}{2}\|x - y\|^2 \tag{52}$$

Summing through all i's and dividing by n, we get

$$f(x) \leq f(y) + \nabla f(y)^\top (x - y) + \frac{\bar{L}}{2}\|x - y\|^2 \tag{53}$$

□

Lemma 27 (Cauchy-Schwarz inequality)**.** *For all $x, y \in \mathbb{R}^d$ we have*

$$|\langle x, y\rangle| \leq \|x\|\|y\|. \tag{54}$$

Lemma 28 (Young's inequality)**.** *For $a, b \in \mathbb{R}$ and $\beta > 0$ we have*

$$ab \leq \frac{a^2\beta}{2} + \frac{b^2}{2\beta}. \tag{55}$$

Lemma 29 (Jensen's inequality)**.** *Let X be a random variable and $g(x)$ be a convex function. Then*

$$g(\mathrm{E}[X]) \leq \mathrm{E}[g(X)]. \tag{56}$$