(Supplementary Material) DM-SARAH: A Variance Reduction Optimization Algorithm for Machine Learning Systems

Rengang Li[†‡], Ruidong Yan[‡⋆], Zhenhua Guo[‡], Zhiyong Qiu[‡], Yaqian Zhao[‡], and Yanwei Wang[‡]
[†]*Department of Computer Science and Technology, Tsinghua University, Beijing, China*
[‡] *Inspur Electronic Information Industry Co., Ltd, Beijing, China*, [⋆]*Corresponding author*
Email: lrg22@tsinghua.edu.cn, {yanruidong, guozhenhua, qiuzhiyong, zhaoyanqian, wangyanwei}@ieisystem.com

I. PROOF OF FACT 1

Fact 1: Consider DM-SARAH with a single outer loop. Suppose each $f_i(\cdot)$ is $\mathbf{L}$-smooth, then the expectation of $||\nabla f(\mathbf{w}_k)||^2$ can be bounded for any $k \geq 0$, i.e.,

$$\sum_{k=0}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k)||^2] \leq \frac{2}{\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \sum_{k=0}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] + (\mathbf{L}\eta - 1)\sum_{k=0}^{m}\mathbb{E}[||\mathbf{v}_k||^2],$$

where $f(\mathbf{w}^*)$ is the optimal value.

Proof 1: Since each $f_i(\cdot)$ is $\mathbf{L}$-smooth, thus function f is $\mathbf{L}$-smooth. For $\mathbf{w}_{k+1}$ and $\mathbf{w}_k \in \mathbb{R}^d$, we have

$$f(\mathbf{w}_{k+1}) \leq f(\mathbf{w}_k) + \nabla f(\mathbf{w}_k)^T(\mathbf{w}_{k+1} - \mathbf{w}_k) + \frac{\mathbf{L}}{2}||\mathbf{w}_{k+1} - \mathbf{w}_k||^2. \tag{1}$$

The iteration rule is $\mathbf{w}_{k+1} = \mathbf{w}_k - \eta\mathbf{v}_k$, thus we can get

$$-\eta\mathbf{v}_k = \mathbf{w}_{k+1} - \mathbf{w}_k. \tag{2}$$

If we put (2) into (1) and take the expectation operator at both ends of the inequality (1), then we have the following inequality

$$\begin{aligned}\mathbb{E}[f(\mathbf{w}_{k+1})] &\leq \mathbb{E}[f(\mathbf{w}_k)] + \frac{\mathbf{L}}{2}\mathbb{E}[||-\eta\mathbf{v}_k||^2] + \mathbb{E}[\nabla f(\mathbf{w}_k)^T(-\eta\mathbf{v_k})]\\ &= \mathbb{E}[f(\mathbf{w}_k)] - \eta\mathbb{E}[\nabla f(\mathbf{w}_k)^T\mathbf{v_k}] + \frac{\mathbf{L}}{2}\mathbb{E}[||-\eta\mathbf{v}_k||^2]. \end{aligned}\tag{3}$$

One the other hand, we have

$$\begin{aligned}\frac{\eta}{2}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] &= \frac{\eta}{2}\mathbb{E}[\nabla f(\mathbf{w}_k)^2] - \frac{\eta}{2}2\mathbb{E}[\nabla f(\mathbf{w}_k)\mathbf{v}_k] + \frac{\eta}{2}\mathbb{E}[||\mathbf{v}_k||^2]\\ &= \frac{\eta}{2}\mathbb{E}[\nabla f(\mathbf{w}_k)^2] - \eta\mathbb{E}[\nabla f(\mathbf{w}_k)\mathbf{v}_k] + \frac{\eta}{2}\mathbb{E}[||\mathbf{v}_k||^2]. \end{aligned}\tag{4}$$

According to (4), we have

$$-\eta\mathbb{E}[\nabla f(\mathbf{w}_k)\mathbf{v}_k] = -\frac{\eta}{2}\mathbb{E}[\nabla f(\mathbf{w}_k)^2] - \frac{\eta}{2}\mathbb{E}[||\mathbf{v}_k||^2 + \frac{\eta}{2}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2]. \tag{5}$$

Combining (5) and (3), we have

$$\begin{aligned}\mathbb{E}[f(\mathbf{w}_{k+1})] &\leq \mathbb{E}[f(\mathbf{w}_k)] + \frac{\eta}{2}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] - \frac{\eta}{2}\mathbb{E}[\nabla f(\mathbf{w}_k)^2] - \frac{\eta}{2}\mathbb{E}[||\mathbf{v}_k||^2 + \frac{\mathbf{L}\eta^2}{2}\mathbb{E}[||\mathbf{v}_k||^2]\\ &= \mathbb{E}[f(\mathbf{w}_k)] + \frac{\eta}{2}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] - \frac{\eta}{2}\mathbb{E}[\nabla f(\mathbf{w}_k)^2] + (\frac{\mathbf{L}\eta^2 - \eta}{2})\mathbb{E}[||\mathbf{v}_k||^2]. \end{aligned}\tag{6}$$

If $k = 0, 1, 2, , , , m$ in inequality (6), we can get the following $(m + 1)$ inequalities:
If $k = 0$,

$$\mathbb{E}[f(\mathbf{w}_1)] \leq \mathbb{E}[f(\mathbf{w}_0)] + \frac{\eta}{2}\mathbb{E}[||\nabla f(\mathbf{w}_0) - \mathbf{v}_0||^2] - \frac{\eta}{2}\mathbb{E}[\nabla f(\mathbf{w}_0)^2] + (\frac{\mathbf{L}\eta^2 - \eta}{2})\mathbb{E}[||\mathbf{v}_0||^2].$$

Identify applicable funding agency here. If none, delete this.

If $k = 1$,

$$\mathbb{E}[f(\mathbf{w}_2)] \leq \mathbb{E}[f(\mathbf{w}_1)] + \frac{\eta}{2}\mathbb{E}[||\nabla f(\mathbf{w}_1) - \mathbf{v}_1||^2] - \frac{\eta}{2}\mathbb{E}[\nabla f(\mathbf{w}_1)^2] + (\frac{\mathbf{L}\eta^2 - \eta}{2})\mathbb{E}[||\mathbf{v}_1||^2].$$

$$\vdots$$

If $k = m$,

$$\mathbb{E}[f(\mathbf{w}_{m+1})] \leq \mathbb{E}[f(\mathbf{w}_m)] + \frac{\eta}{2}\mathbb{E}[||\nabla f(\mathbf{w}_m) - \mathbf{v}_m||^2] - \frac{\eta}{2}\mathbb{E}[\nabla f(\mathbf{w}_m)^2] + (\frac{\mathbf{L}\eta^2 - \eta}{2})\mathbb{E}[||\mathbf{v}_m||^2].$$

If we sum up the $(m+1)$ inequalities above, then we can obtain the following equation

$$\mathbb{E}[f(\mathbf{w}_{m+1})] \leq \mathbb{E}[f(\mathbf{w}_0)] + \sum_{k=0}^{m}\frac{\eta}{2}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] - \sum_{k=0}^{m}\frac{\eta}{2}\mathbb{E}[\nabla f(\mathbf{w}_k)^2] + \sum_{k=0}^{m}(\frac{\mathbf{L}\eta^2 - \eta}{2})\mathbb{E}[||\mathbf{v}_k||^2]. \tag{7}$$

Therefore, we can get

$$\frac{\eta}{2}\sum_{k=0}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k)||^2] \leq \frac{\eta}{2}\sum_{k=0}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] + \mathbb{E}[f(\mathbf{w}_0)] + (\frac{\mathbf{L}\eta^2 - \eta}{2})\sum_{k=0}^{m}\mathbb{E}[||\mathbf{v}_k||^2] - \mathbb{E}[f(\mathbf{w}_{m+1})].$$

Multiplying both ends of the above inequality by $\frac{2}{\eta}$, we have

$$\sum_{k=0}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k)||^2] \leq \frac{2}{\eta}\mathbb{E}[f(\mathbf{w}_0)] - \frac{2}{\eta}\mathbb{E}[f(\mathbf{w}_{m+1})] + \sum_{k=0}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] + (\mathbf{L}\eta - 1)\sum_{k=0}^{m}\mathbb{E}[||\mathbf{v}_k||^2]. \tag{8}$$

Since $f(\mathbf{w}^*)$ is the optimal value, so $f(\mathbf{w}^*) \leq f(\mathbf{w}_{m+1})$ always holds. The inequality (8) can be rewritten as

$$\begin{aligned}\sum_{k=0}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k)||^2] &\leq \frac{2}{\eta}\mathbb{E}[f(\mathbf{w}_0)] - \frac{2}{\eta}\mathbb{E}[f(\mathbf{w}_{m+1})] + \sum_{k=0}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] + (\mathbf{L}\eta - 1)\sum_{k=0}^{m}\mathbb{E}[||\mathbf{v}_k||^2] \\ &\leq \frac{2}{\eta}\mathbb{E}[f(\mathbf{w}_0)] + \sum_{k=0}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] + (\mathbf{L}\eta - 1)\sum_{k=0}^{m}\mathbb{E}[||\mathbf{v}_k||^2] - \frac{2}{\eta}\mathbb{E}[f(\mathbf{w}^*)].\end{aligned}$$

II. PROOF OF FACT 2

Fact 2: Suppose each $f_i(\cdot)$ is $\mathbf{L}$-smooth. Consider $\mathbf{v}_k$ defined in DM-SARAH, for any $k \geq 1$, it holds

$$\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] = \mathbb{E}[||\nabla f(\mathbf{w}_0) - \mathbf{v}_0||^2] + \sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_k - \mathbf{v}_{k-1}||^2] - \sum_{k=1}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1})||^2].$$

Proof 2: According to $a - b = a - c + c - d + d - b$, the term $\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2]$ can be rewritten as follows.

$$\begin{aligned}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] &= \mathbb{E}[||\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1}) + \nabla f(\mathbf{w}_{k-1}) - \mathbf{v}_{k-1} + \mathbf{v}_{k-1} - \mathbf{v}_k||^2] \\ &= \mathbb{E}[||\nabla f(\mathbf{w}_{k-1}) - \mathbf{v}_{k-1} + \nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1}) + \mathbf{v}_{k-1} - \mathbf{v}_k||^2] \\ &= \mathbb{E}[||\nabla f(\mathbf{w}_{k-1}) - \mathbf{v}_{k-1}||^2] + \mathbb{E}[||\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1})||^2] + \mathbb{E}[||\mathbf{v}_{k-1} - \mathbf{v}_k||^2] \\ &+ 2\mathbb{E}[(\nabla f(\mathbf{w}_{k-1}) - \mathbf{v}_{k-1})(\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1}))] \\ &+ 2\mathbb{E}[(\nabla f(\mathbf{w}_{k-1}) - \mathbf{v}_{k-1})(\mathbf{v}_{k-1} - \mathbf{v}_k)] \\ &+ 2\mathbb{E}[(\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1})(\mathbf{v}_{k-1} - \mathbf{v}_k)].\end{aligned} \tag{9}$$

According to the definition of $\mathbf{v}_k$ in DM-SARAH, i.e.,

$$\begin{aligned}\mathbf{v}_k &= \frac{1}{b_{in}}\sum_{i\in\mathcal{B}_k}[\nabla f_i(\mathbf{w}_k) - \nabla f_i(\mathbf{w}_{k-1})] + \mathbf{v}_{k-1}, \\ \mathbf{v}_k - \mathbf{v}_{k-1} &= \frac{1}{b_{in}}\sum_{i\in\mathcal{B}_k}[\nabla f_i(\mathbf{w}_k) - \nabla f_i(\mathbf{w}_{k-1})].\end{aligned} \tag{10}$$

Applying the expectation operator to both ends of the inequality (10), we have

$$\begin{aligned}\mathbb{E}[\mathbf{v}_k - \mathbf{v}_{k-1}] &= \frac{1}{b_{in}}\mathbb{E}[\sum_{i\in\mathcal{B}_k}(\nabla f_i(\mathbf{w}_k) - \nabla f_i(\mathbf{w}_{k-1}))] \\ &= \frac{1}{b_{in}} \cdot \frac{b_{in}}{n}\sum_{i=1}^{n}[\nabla f_i(\mathbf{w}_k) - \nabla f_i(\mathbf{w}_{k-1})] \\ &= \frac{1}{n}\sum_{i=1}^{n}[\nabla f_i(\mathbf{w}_k) - \nabla f_i(\mathbf{w}_{k-1})] \\ &= \nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1}).\end{aligned} \tag{11}$$

If we add (11) to (9), then we get

$$\begin{aligned}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] &= \mathbb{E}[||\nabla f(\mathbf{w}_{k-1}) - \mathbf{v}_{k-1}||^2] + \mathbb{E}[||\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1})||^2] + \mathbb{E}[||\mathbf{v}_{k-1} - \mathbf{v}_k||^2] \\ &+ 2\mathbb{E}[(\nabla f(\mathbf{w}_{k-1}) - \mathbf{v}_{k-1})(\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1}))] \\ &- 2\mathbb{E}[(\nabla f(\mathbf{w}_{k-1}) - \mathbf{v}_{k-1})(\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1}))] - 2\mathbb{E}[||\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1})||^2] \\ &= \mathbb{E}[||\nabla f(\mathbf{w}_{k-1}) - \mathbf{v}_{k-1}||^2] - \mathbb{E}[||\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1})||^2] + \mathbb{E}[||\mathbf{v}_{k-1} - \mathbf{v}_k||^2].\end{aligned} \tag{12}$$

If $k = 1, 2, , , , m$ in inequality (12), we can get the following m inequalities:
If $k = 1$,

$$\mathbb{E}[||\nabla f(\mathbf{w}_1) - \mathbf{v}_1||^2] = \mathbb{E}[||\nabla f(\mathbf{w}_0) - \mathbf{v}_0||^2] - \mathbb{E}[||\nabla f(\mathbf{w}_1) - \nabla f(\mathbf{w}_0)||^2] + \mathbb{E}[||\mathbf{v}_1 - \mathbf{v}_0||^2].$$

If $k = 2$,

$$\mathbb{E}[||\nabla f(\mathbf{w}_2) - \mathbf{v}_2||^2] = \mathbb{E}[||\nabla f(\mathbf{w}_1) - \mathbf{v}_1||^2] - \mathbb{E}[||\nabla f(\mathbf{w}_2) - \nabla f(\mathbf{w}_1)||^2] + \mathbb{E}[||\mathbf{v}_2 - \mathbf{v}_1||^2].$$

$$\vdots$$

If $k = m$,

$$\mathbb{E}[||\nabla f(\mathbf{w}_m) - \mathbf{v}_m||^2] = \mathbb{E}[||\nabla f(\mathbf{w}_m) - \mathbf{v}_m||^2] - \mathbb{E}[||\nabla f(\mathbf{w}_m) - \nabla f(\mathbf{w}_{m-1})||^2] + \mathbb{E}[||\mathbf{v}_m - \mathbf{v}_{m-1}||^2].$$

We sum the above m inequalities, then it can be obtained

$$\mathbb{E}[||\nabla f(\mathbf{w}_m) - \mathbf{v}_m||^2] = \mathbb{E}[||\nabla f(\mathbf{w}_0) - \mathbf{v}_0||^2] - \sum_{k=1}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1})||^2] + \sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_k - \mathbf{v}_{k-1}||^2]. \tag{13}$$

III. PROOF OF FACT 3

Fact 3: Suppose each $f_i(\cdot)$ is $\mathbf{L}$-smooth, then the following difference can be bounded for any $k \geq 1$, i.e.,

$$\mathbb{E}[||\mathbf{v}_k - \mathbf{v}_{k-1}||^2] - \mathbb{E}[||\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1})||^2] \leq \frac{(n - b_{in})}{b_{in}}\mathbf{L}^2\eta^2\mathbb{E}[||\mathbf{v}_{k-1}||^2].$$

Proof 3: According to the definition of $\mathbf{v}_k$ in DM-SARAH, that is,

$$\mathbf{v}_k = \frac{1}{b_{in}}\sum_{j\in\mathcal{B}_k}[\nabla f_j(\mathbf{w}_k) - f_j(\mathbf{w}_{k-1})] + \mathbf{v}_{k-1},$$

then we have

$$\mathbf{v}_k - \mathbf{v}_{k-1} = \frac{1}{b_{in}}\sum_{j\in\mathcal{B}_k}[\nabla f_j(\mathbf{w}_k) - f_j(\mathbf{w}_{k-1})].$$

Taking the expectation operator on the square of the ℓ_2-norm of the above formula, we can get

$$\mathbb{E}[||\mathbf{v}_k - \mathbf{v}_{k-1}||^2] = \mathbb{E}[||\frac{1}{b_{in}}\sum_{j\in\mathcal{B}_k}[\nabla f_j(\mathbf{w}_k) - f_j(\mathbf{w}_{k-1})||^2]. \tag{14}$$

According to the definition of full gradient, it holds

$$||\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1})||^2 = ||\frac{1}{n}\sum_{j=1}^{n}\nabla f_j(\mathbf{w}_k) - \frac{1}{n}\sum_{j=1}^{n}\nabla f_j(\mathbf{w}_{k-1})||^2. \tag{15}$$

We consider the difference between (14) and (15), i.e.,

$$\begin{aligned}&\mathbb{E}[||\mathbf{v}_k - \mathbf{v}_{k-1}||^2] - ||\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1})||^2 \\ &= \mathbb{E}[||\frac{1}{b_{in}} \sum_{j\in\mathcal{B}_k} [\nabla f_j(\mathbf{w}_k) - f_j(\mathbf{w}_{k-1})||^2] - ||\frac{1}{n}\sum_{j=1}^{n}[\nabla f_j(\mathbf{w}_k) - \nabla f_j(\mathbf{w}_{k-1})]||^2. \end{aligned} \tag{16}$$

For simplicity, let

$$\Delta := \nabla f_j(\mathbf{w}_k) - \nabla f_j(\mathbf{w}_{k-1}). \tag{17}$$

Combining (16) and (17), we can obtain the following inequality

$$\begin{aligned}\mathbb{E}[||\mathbf{v}_k - \mathbf{v}_{k-1}||^2] - ||\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1})||^2 &= \mathbb{E}\left[||\frac{1}{b_{in}}\sum_{j\in\mathcal{B}_k}\Delta||^2\right] - ||\frac{1}{n}\sum_{j=1}^{n}\Delta||^2 \\ &= \mathbb{E}\left[\frac{1}{b_{in}}\sum_{j'\in\mathcal{B}_k}\Delta_{j'}\right]\left[\frac{1}{b_{in}}\sum_{j''\in\mathcal{B}_k}\Delta_{j''}\right] - \frac{1}{n^2}\sum_{j'=1}^{n}\Delta_{j'}\sum_{j''=1}^{n}\Delta_{j''} \\ &= \frac{1}{b_{in}^2}\mathbb{E}\left[\sum_{j'\in\mathcal{B}_k}\sum_{j''\in\mathcal{B}_k}\Delta_{j'}\Delta_{j''}\right] - \frac{1}{n^2}\sum_{j'=1}^{n}\sum_{j''=1}^{n}\Delta_{j'}\Delta_{j''} \\ &= \frac{1}{b_{in}^2}\left[\frac{b_{in}}{n}\cdot\frac{(b_{in}-1)}{(n-1)}\sum_{j'=1}^{n}\sum_{j''=1}^{n}\Delta_{j'}\Delta_{j''} - \left[\frac{b_{in}}{n}\cdot\frac{(b_{in}-1)}{(n-1)}\right]\sum_{j'=j''=1}^{n}\Delta_{j'}\Delta_{j''}\right. \\ &\left. + \frac{b}{n}\sum_{j'=j''=1}^{n}\Delta_{j'}\Delta_{j''}\right] - \frac{1}{n^2}\sum_{j'=1}^{n}\sum_{j''=1}^{n}\Delta_{j'}\Delta_{j''} \\ &= \left[\frac{b_{in}-1}{b_{in}(n-1)} - \frac{1}{n^2}\right]\sum_{j'=1}^{n}\sum_{j'=1}^{n}\Delta_{j'}\Delta_{j''} + \left[\frac{n-b_{in}}{b_{in}n(n-1)}\right]\sum_{j'=j''=1}^{n}\Delta_{j'}\Delta_{j''} \\ &= \frac{(n-b_{in})}{b_{in}(n-1)n}\left[\frac{-1}{n}\sum_{j'=1}^{n}\sum_{j''=1}^{n}\Delta_{j'}\Delta_{j''} + \sum_{j'=j''=1}^{n}\Delta_{j'}\Delta_{j''}\right] \\ &= \frac{(n-b_{in})}{b_{in}(n-1)n}\left[\frac{-1}{n}||\sum_{j=1}^{n}\Delta_j||^2 + \sum_{j=1}^{n}||\Delta_j||^2\right] \\ &\leq \frac{(n-b_{in})}{b_{in}(n-1)n}\left[(n-1)\sum_{j=1}^{n}||\Delta_j||^2\right] \\ &= \frac{(n-b_{in})}{b_{in}n}\sum_{j=1}^{n}||\Delta_j||^2 \\ &= \frac{(n-b_{in})}{b_{in}n}\left[\sum_{j=1}^{n}||\nabla f_j(\mathbf{w}_k) - \nabla f_j(\mathbf{w}_{k-1})||^2\right]. \end{aligned} \tag{18}$$

Since $f_i(\cdot)$ has a Lipschitz continuous gradient, i.e,

$$||\nabla f_j(\mathbf{w}_k) - \nabla f_j(\mathbf{w}_{k-1})|| \leq \mathbf{L}||\mathbf{w}_k - \mathbf{w}_{k-1}||. \tag{19}$$

Therefore (18) can be bounded by the following inequality

$$\begin{aligned}
\frac{(n-b_{in})}{b_{in}n}\left[\sum_{j=1}^{n}||\nabla f_j(\mathbf{w}_k)-\nabla f_j(\mathbf{w}_{k-1})||^2\right] &\leq \frac{(n-b_{in})}{b_{in}n}\sum_{j=1}^{n}||\mathbf{L}(\mathbf{w}_k-\mathbf{w}_{k-1})||^2 \\
&= \frac{(n-b_{in})}{b_{in}n}\sum_{j=1}^{n}||\mathbf{L}(-\eta\mathbf{v}_{k-1})||^2 \\
&= \frac{(n-b_{in})}{b_{in}n}\mathbf{L}^2\eta^2\sum_{j=1}^{n}||\mathbf{v}_{k-1}||^2. \\
&= \frac{(n-b_{in})}{b_{in}}\mathbf{L}^2\eta^2||\mathbf{v}_{k-1}||^2. \qquad (20)
\end{aligned}$$

Therefore, by taking expectation operator, we have

$$\mathbb{E}[||\mathbf{v}_k-\mathbf{v}_{k-1}||^2]-\mathbb{E}[||\nabla f(\mathbf{w}_k)-\nabla f(\mathbf{w}_{k-1})||^2] \leq \frac{(n-b_{in})}{b_{in}}\mathbf{L}^2\eta^2\mathbb{E}||\mathbf{v}_{k-1}||^2. \qquad (21)$$

According to the **Fact 2**, for $k \geq 1$, we have

$$\begin{aligned}
\mathbb{E}[||\nabla f(\mathbf{w}_k)-\mathbf{v}_k||^2] &= \sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_k-\mathbf{v}_{k-1}||^2]-\sum_{k=1}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k)-\nabla f(\mathbf{w}_{k-1})||^2]+\mathbb{E}[||\nabla f(\mathbf{w}_0)-\mathbf{v}_0||^2] \\
&\leq \frac{(n-b_{in})}{b_{in}}\mathbf{L}^2\eta^2\sum_{k=1}^{m}\mathbb{E}||\mathbf{v}_{k-1}||^2+\mathbb{E}[||\nabla f(\mathbf{w}_0)-\mathbf{v}_0||^2]. \qquad (22)
\end{aligned}$$

IV. PROOF OF THEOREM 4

Theorem 4: Suppose each $f_i(\cdot)$ is $\mathbf{L}$-smooth. Consider DM-SARAH with a learning rate

$$\eta \leq \frac{2}{\mathbf{L}(\sqrt{\frac{4m(n-b_{in})}{b_{in}}+1}+1)}.$$

Let $\mathbb{E}[||\nabla f(\mathbf{w}_0)-\mathbf{v}_0||^2] \leq \mu^2$. Then the expectation of $||\nabla f(\mathbf{w}_k)||^2$ can be bounded, i.e.,

$$\mathbb{E}[||\nabla f(\tilde{\mathbf{w}}_k)||^2] \leq \frac{2}{\eta(m+1)}\mathbb{E}[f(\mathbf{w}_0)-f(\mathbf{w}^*)]+\mu^2,$$

where $f(\mathbf{w}^*)$ is the optimal value. $\tilde{\mathbf{w}}=\mathbf{w}_k$ where k is chosen uniformly at random from $\{0,1,...,m\}$.

Proof 4: According to **Fact 2** and **Fact 3**, for $k \geq 1$, we have

$$\begin{aligned}
\mathbb{E}[||\nabla f(\mathbf{w}_k)-\mathbf{v}_k||^2] &= \sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_k-\mathbf{v}_{k-1}||^2]+\mathbb{E}[||\nabla f(\mathbf{w}_0)-\mathbf{v}_0||^2]-\sum_{k=1}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k)-\nabla f(\mathbf{w}_{k-1})||^2] \\
&\leq \frac{(n-b_{in})}{b_{in}}\mathbf{L}^2\eta^2\sum_{k=1}^{m}\mathbb{E}||\mathbf{v}_{k-1}||^2+\mu^2. \qquad (23)
\end{aligned}$$

That is to say,

$$\mathbb{E}[||\nabla f(\mathbf{w}_k)-\mathbf{v}_k||^2] \leq \frac{(n-b_{in})}{b_{in}}\mathbf{L}^2\eta^2\sum_{k=1}^{m}\mathbb{E}||\mathbf{v}_{k-1}||^2+\mu^2.$$

If $k=1$,

$$\mathbb{E}[||\nabla f(\mathbf{w}_1)-\mathbf{v}_1||^2] \leq \frac{(n-b_{in})}{b_{in}}\mathbf{L}^2\eta^2\mathbb{E}||\mathbf{v}_0||^2+\mu^2.$$

If $k=2$,

$$\mathbb{E}[||\nabla f(\mathbf{w}_2)-\mathbf{v}_2||^2] \leq \frac{(n-b_{in})}{b_{in}}\mathbf{L}^2\eta^2\sum_{k=1}^{2}\mathbb{E}||\mathbf{v}_{k-1}||^2+\mu^2.$$

$$\vdots$$

If $k = m$,

$$\mathbb{E}[||\nabla f(\mathbf{w}_m) - \mathbf{v}_m||^2] \leq \frac{(n-b_{in})}{b_{in}} \mathbf{L}^2 \eta^2 \sum_{k=1}^{m} \mathbb{E}||\mathbf{v}_{k-1}||^2 + \mu^2.$$

We sum the above m inequalities and subtract the term $(1-\mathbf{L}\eta)\sum_{k=0}^{m}\mathbb{E}[||\mathbf{v}_k||^2]$ at both ends, thus we can get

$$\begin{aligned}
\sum_{k=1}^{m} \mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] - (1-\mathbf{L}\eta)\sum_{k=0}^{m}\mathbb{E}[||\mathbf{v}_k||^2] &\leq \frac{(n-b_{in})}{b_{in}} \mathbf{L}^2\eta^2 [m\mathbb{E}[||\mathbf{v}_0||^2] \\
&\quad + (m-1)\mathbb{E}[||\mathbf{v}_1||^2] + ... + \mathbb{E}[||\mathbf{v}_{m-1}||^2]] - (1-\mathbf{L}\eta)\sum_{k=0}^{m}\mathbb{E}[||\mathbf{v}_k||^2] + m\mu^2 \\
&\leq \frac{(n-b_{in})}{b_{in}} \mathbf{L}^2\eta^2 [m\mathbb{E}[||\mathbf{v}_0||^2] + m\mathbb{E}[||\mathbf{v}_1||^2] + ... + m\mathbb{E}[||\mathbf{v}_{m-1}||^2]] - (1-\mathbf{L}\eta)\sum_{k=0}^{m}\mathbb{E}[||\mathbf{v}_k||^2] + m\mu^2 \\
&= \left[\frac{(n-b_{in})}{b_{in}} \mathbf{L}^2\eta^2 m - (1-\mathbf{L}\eta)\right]\sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_{k-1}||^2] + m\mu^2. \qquad (24)
\end{aligned}$$

On the other hand, let the coefficient of (24) be 0, i.e.,

$$\frac{(n-b_{in})}{b_{in}} \mathbf{L}^2 m\eta^2 + \mathbf{L}\eta - 1 = 0. \qquad (25)$$

Without loss of generality, let $A = \frac{(n-b_{in})\mathbf{L}^2 m}{b_{in}}$ and $B = \mathbf{L}$, then the root is as follow.

$$\begin{aligned}
\eta &= \frac{-B + \sqrt{B^2 + 4A}}{2A} \\
&= \frac{-\mathbf{L} + \sqrt{\mathbf{L}^2 + 4\frac{(n-b_{in})\mathbf{L}^2 m}{b_{in}}}}{2\frac{(n-b_{in})\mathbf{L}^2 m}{b_{in}}} \\
&= \frac{\sqrt{4b_{in}(n-b_{in})m + b_{in}^2} - b_{in}}{2(n-b_{in})\mathbf{L}m} \\
&= \frac{2}{\mathbf{L}\left(\sqrt{\frac{4(n-b_{in})m}{b_{in}} + 1} + 1\right)}. \qquad (26)
\end{aligned}$$

When $\eta \leq \frac{2}{\mathbf{L}(\sqrt{\frac{4(n-b_{in})m}{b_{in}}+1}+1)}$, then the first item of (24) is negative. In other words,

$$\begin{aligned}
\sum_{k=0}^{m} \mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] - (1-\mathbf{L}\eta)\sum_{k=0}^{m}\mathbb{E}[||\mathbf{v}_k||^2] &\leq \left[\frac{(n-b_{in})}{b_{in}} \mathbf{L}^2\eta^2 m - (1-\mathbf{L}\eta)\right]\sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_{k-1}||^2] + (m+1)\mu^2 \\
&\leq 0 + (m+1)\mu^2 = (m+1)\mu^2. \qquad (27)
\end{aligned}$$

According to the **Fact 1** and inequality (27), we have

$$\begin{aligned}
\sum_{k=0}^{m} \mathbb{E}[||\nabla f(\mathbf{w}_k)||^2] &\leq \frac{2}{\eta}\mathbb{E}[f(\mathbf{w}_0)] + \sum_{k=0}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] + (\mathbf{L}\eta - 1)\sum_{k=0}^{m}\mathbb{E}[||\mathbf{v}_k||^2] - \frac{2}{\eta}\mathbb{E}[f(\mathbf{w}^*)]. \\
&\leq \frac{2}{\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + (m+1)\mu^2. \qquad (28)
\end{aligned}$$

If $\tilde{\mathbf{w}} = \mathbf{w}_k$, where k is chosen uniformaly at random from $\{0, 1, ..., m\}$, then

$$\begin{aligned}
\mathbb{E}[||\nabla f(\tilde{\mathbf{w}})||^2] &= \frac{1}{m+1}\sum_{k=0}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k)||^2] \\
&\leq \frac{2}{\eta(m+1)}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \mu^2. \qquad (29)
\end{aligned}$$

V. PROOF REMARK 5

Remark 5: Suppose each $f_i(\cdot)$ is $\mathbf{L}$-smooth. Let $b_{out} = \log n$, $b_{in} = (1/4)n$ and $\eta = 2/(\mathbf{L}(\sqrt{12m+1}+1))$ where m is the total number of iterations in inner loops. Then $||\nabla f(\tilde{\mathbf{w}}_k)||^2$ converges sublinearly in expectation and the total complexity of DM-SARAH to achieve a ϵ-approximate solution is $\mathcal{O}(\log n + \frac{n\mathbf{L}^2}{2\epsilon^2})$.

Proof 5: Let $b_{out} = \log n$, $b_{in} = \frac{1}{4}n$ and $\eta = 2/(\mathbf{L}(\sqrt{12m+1}+1))$ where m is the total number of iterations in inner loops. According to the **Theorem 4**, we have

$$\mathbb{E}[||\nabla f(\tilde{\mathbf{w}}_k)||^2] \leq \frac{2}{\frac{2}{\mathbf{L}(\sqrt{12m+1}+1)}(m+1)}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \mu^2 \leq \epsilon$$

Therefore, we have $m = \mathcal{O}(\mathbf{L}^2/\epsilon^2)$. The complexity analysis of DM-SARAH is as follows:

- It needs $b_{out} = \log n$ gradients in one outer loop, i.e., its complexity is $\mathcal{O}(\log n)$;
- It needs $2mb_{in}$ gradients in m iterations of inner loops, i.e., its complexity is $\mathcal{O}(n\mathbf{L}^2/(2\epsilon^2))$;
- It needs $b_{out} + 2msb_{in}$ gradients in one outer loop, i.e., its total complexity is $\mathbf{O}(\log n + \frac{n\mathbf{L}^2}{2\epsilon^2})$.

VI. PROOF OF LEMMA 1

Lemma 1: Suppose each $f_i(\cdot)$ is $\mathbf{L}$-smooth, then the expectation of $||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2$ can be bounded for any $k \geq 1$,

$$\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] \leq \frac{1}{b_{in}}\mathbf{L}^2\eta^2\sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_{k-1}||^2] + \mu^2.$$

Proof 6: Consider the definition of $\mathbf{v}_k$ in the DM-SARAH algrithm, it holds

$$\mathbf{v}_k = \frac{1}{b_{in}}\sum_{i\in\mathcal{B}_k}[\nabla f_i(\mathbf{w}_k) - \nabla f_i(\mathbf{w}_{k-1})] + \mathbf{v}_{k-1},$$
$$||\mathbf{v}_k - \mathbf{v}_{k-1}||^2 = ||\frac{1}{b_{in}}\sum_{i\in\mathcal{B}_k}[\nabla f_i(\mathbf{w}_k) - \nabla f_i(\mathbf{w}_{k-1})]||^2.$$

Since $f_i(\cdot)$ is $\mathbf{L}$-smooth and $\mathbf{w}_k = \mathbf{w}_{k-1} - \eta\mathbf{v}_{k-1}$, then the above equality satisfies the following condition.

$$\begin{aligned}
||\mathbf{v}_k - \mathbf{v}_{k-1}||^2 &= ||\frac{1}{b_{in}}\sum_{i\in\mathcal{B}_k}[\nabla f_i(\mathbf{w}_k) - \nabla f_i(\mathbf{w}_{k-1})]||^2 \\
&\leq ||\frac{1}{b_{in}}\sum_{i=1}^{b_{in}}\mathbf{L}(\mathbf{w}_k - \mathbf{w}_{k-1})||^2 \\
&= \frac{1}{b_{in}^2}\cdot\mathbf{L}^2\sum_{i=1}^{b_{in}}||\mathbf{w}_k - \mathbf{w}_{k-1}||^2 \\
&= \frac{1}{b_{in}^2}\cdot\mathbf{L}^2\sum_{i=1}^{b_{in}}||-\eta\mathbf{v}_{k-1}||^2 \\
&= \frac{1}{b_{in}^2}\cdot\mathbf{L}^2\eta^2\sum_{i=1}^{b_{in}}||\mathbf{v}_{k-1}||^2.
\end{aligned} \tag{30}$$

On the other hand, according to the **Fact 2** and (30), we have

$$\begin{aligned}
\mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] &= \sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_k - \mathbf{v}_{k-1}||^2] - \sum_{k=1}^{m}\mathbb{E}[||\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k-1})||^2 + \mathbb{E}[||\nabla f(\mathbf{w}_0) - \mathbf{v}_0||^2] \\
&\leq \sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_k - \mathbf{v}_{k-1}||^2] + \mathbb{E}[||\nabla f(\mathbf{w}_0) - \mathbf{v}_0||^2] \\
&\leq \sum_{k=1}^{m}\frac{1}{b_{in}^2}\cdot\mathbf{L}^2\eta^2\sum_{i=1}^{b_{in}}\mathbb{E}[||\mathbf{v}_{k-1}||^2] + \mu^2 \\
&= \frac{1}{b_{in}^2}\cdot\mathbf{L}^2\eta^2\sum_{k=1}^{m}b_{in}\mathbb{E}[||\mathbf{v}_{k-1}||^2] + \mu^2 \\
&= \frac{1}{b_{in}}\cdot\mathbf{L}^2\eta^2\sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_{k-1}||^2] + \mu^2.
\end{aligned} \tag{31}$$

VII. PROOF OF LEMMA 2

Lemma 2: Suppose each $f_i(\cdot)$ is $\mathbf{L}$-smooth. Learning rate

$$\eta \leq \frac{1}{2\mathbf{L}^3(\sqrt{\frac{4m}{b_{in}}+1}+1)}.$$

Then for any $k \geq 1$, it holds,

$$\left[\frac{1}{b_{in}} \cdot \mathbf{L}^2\eta^2 m - (1-\mathbf{L}\eta)\right]\sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_{k-1}||^2] \leq 0.$$

Proof 7: Consider the following inequality

$$\begin{aligned}
&\frac{1}{b_{in}}\mathbf{L}^2\eta^2\sum_{t=0}^{m}\sum_{k=1}^{t}\mathbb{E}[||\mathbf{v}_{k-1}||^2] - (1-\mathbf{L}\eta)\sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_k||^2] = \frac{1}{b_{in}}\cdot\mathbf{L}^2\eta^2[\mathbb{E}[||\mathbf{v}_0||^2] + \mathbb{E}[||\mathbf{v}_0||^2] + \mathbb{E}[||\mathbf{v}_1||^2] \\
&\quad + ... + \mathbb{E}[||\mathbf{v}_0||^2] + ... + \mathbb{E}[||\mathbf{v}_{m-1}||^2]] - (1-\mathbf{L}\eta)\sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_k||^2] \\
&\quad = \frac{1}{b_{in}}\cdot\mathbf{L}^2\eta^2[m\mathbb{E}[||\mathbf{v}_0||^2] + (m-1)\mathbb{E}[||\mathbf{v}_1||^2] + ... + \mathbb{E}[||\mathbf{v}_{m-1}||^2]] - (1-\mathbf{L}\eta)\sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_k||^2] \\
&\quad \leq \frac{1}{b_{in}}\cdot\mathbf{L}^2\eta^2[m\mathbb{E}[||\mathbf{v}_0||^2] + m\mathbb{E}[||\mathbf{v}_1||^2] + ... + m\mathbb{E}[||\mathbf{v}_{m-1}||^2]] - (1-\mathbf{L}\eta)\sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_k||^2] \\
&\quad \leq \frac{1}{b_{in}}\cdot\mathbf{L}^2\eta^2 m\sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_{k-1}||^2] - (1-\mathbf{L}\eta)\sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_{k-1}||^2] \\
&\quad = \left[\frac{1}{b_{in}}\mathbf{L}^2\eta^2 m - (1-\mathbf{L}\eta)\right]\sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_{k-1}||^2].
\end{aligned} \tag{32}$$

Without loss of generality, let

$$\frac{1}{b_{in}}\cdot\mathbf{L}^2\eta^2 m + \mathbf{L}\eta - 1 = 0.$$

The root of above equality is as follow.

$$\begin{aligned}
\eta &= \frac{\sqrt{\mathbf{L}^2 + \frac{4\mathbf{L}^2 m}{b_{in}}} - \mathbf{L}}{\frac{2\cdot 4\mathbf{L}^2 m}{b_{in}}} \\
&= \frac{(\mathbf{L}\sqrt{1+\frac{4m}{b_{in}}} - \mathbf{L})}{(\frac{8\mathbf{L}^2 m}{b_{in}})} \\
&= \frac{(\mathbf{L}\sqrt{1+\frac{4m}{b_{in}}} - \mathbf{L})(\mathbf{L}\sqrt{1+\frac{4m}{b_{in}}} + \mathbf{L})}{(\frac{8\mathbf{L}^2 m}{b_{in}})(\mathbf{L}\sqrt{1+\frac{4m}{b_{in}}} + \mathbf{L})} \\
&= \frac{1}{2\mathbf{L}^3(\sqrt{1+\frac{4m}{b_{in}}}+1)}.
\end{aligned}$$

When $\eta \leq \frac{1}{2\mathbf{L}^3(\sqrt{1+\frac{4m}{b_{in}}}+1)}$, it holds

$$\left[\frac{1}{b_{in}} \cdot \mathbf{L}^2\eta^2 m - (1-\mathbf{L}\eta)\right]\sum_{k=1}^{m}\mathbb{E}[||\mathbf{v}_{k-1}||^2] \leq 0. \tag{33}$$

VIII. PROOF OF THEOREM 6

Theorem 6: Suppose each $f_i(\cdot)$ is $\mathbf{L}$-smooth. Consider DM-SARAH with $\eta \leq \frac{1}{2\mathbf{L}^3(\sqrt{\frac{4m}{b_{in}}+1}+1)}$. Let $b_{out} = \log n$ where $m = n$. If there is a constant $\sigma > 0$ such that $\mathbb{E}[||\nabla f_i(\mathbf{w}_0)||] \leq \sigma$, then the expectation of $||\nabla f(\mathbf{w}_k)||^2$ can be bounded, i.e.,

$$\mathbb{E}[||\nabla f(\tilde{\mathbf{w}}_k)||^2] \leq \frac{2}{(m+1)\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \frac{\sigma}{\log m} + \mu^2,$$

where $f(\mathbf{w}^*)$ is the optimal value.

Proof 8: In DM-SARAH algorithm, $\tilde{\mathbf{w}}^s = \mathbf{w}_k^s$ and $\mathbf{w}_0^s = \mathbf{w}^{s-1}$ $(s \geq 1)$ where k is chosen uniformly at random from $\{0, 1, 2, ..., m\}$, i.e.,

$$\mathbb{E}[||\nabla f(\tilde{\mathbf{w}}_k)||^2] = \frac{1}{m+1} \sum_{k=0}^{m} \mathbb{E}[||\nabla f(\mathbf{w}_k)||^2]. \tag{34}$$

According to **Theorem 4** and **Fact 1**, we have

$$\sum_{k=0}^{m} \mathbb{E}[||\nabla f(\mathbf{w}_k)||^2] \leq \frac{2}{\eta}\mathbb{E}[f(\mathbf{w}_0)] + \sum_{k=0}^{m} \mathbb{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] + (\mathbf{L}\eta - 1)\sum_{k=0}^{m} \mathbb{E}[||\mathbf{v}_k||^2] - \frac{2}{\eta}\mathbb{E}[f(\mathbf{w}^*)] + (m+1)\mu^2. \tag{35}$$

Combining (34) and (35), it holds

$$\mathbb{E}[||\nabla f(\tilde{\mathbf{w}}_k)||^2] \leq \frac{1}{m+1}\left[\frac{2}{\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + (m+1)\mu^2 + \sum_{k=0}^{m} \mathbf{E}[||\nabla f(\mathbf{w}_k) - \mathbf{v}_k||^2] + (\mathbf{L}\eta - 1)\sum_{k=0}^{m} \mathbb{E}[||\mathbf{v}_k||^2]\right].$$

According to **Lemma 1**, it can be rewritten as follows.

$$\begin{aligned}\mathbb{E}[||\nabla f(\tilde{\mathbf{w}}_k)||^2] \leq& \frac{2}{(m+1)\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \mathbb{E}[||\nabla f(\mathbf{w}_0) - \mathbf{v}_0||^2] + \mu^2 \\ &+ \frac{1}{m+1}\left[\frac{1}{b_{in}}\mathbf{L}^2\eta^2 \sum_{t=0}^{m}\sum_{k=1}^{t} \mathbb{E}[||\mathbf{v}_{k-1}||^2] + (\mathbf{L}\eta - 1)\sum_{k=0}^{m} \mathbf{E}[||\mathbf{v}_k||^2]\right].\end{aligned} \tag{36}$$

From the **Lemma 2**, we have

$$\frac{1}{m+1}[\frac{1}{b_{in}}\mathbf{L}^2\eta^2 \sum_{t=0}^{m}\sum_{k=1}^{t} \mathbb{E}[||\mathbf{v}_{k-1}||^2] + (\mathbf{L}\eta - 1)\sum_{k=0}^{m} \mathbf{E}[||\mathbf{v}_k||^2]] \leq 0.$$

Therefore, the (36) can be rewritten as

$$\mathbb{E}[||\nabla f(\tilde{\mathbf{w}}_k)||^2] \leq \frac{2}{(m+1)\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \mathbb{E}[||\nabla f(\mathbf{w}_0) - \mathbf{v}_0||^2] + \mu^2. \tag{37}$$

From the [1], we have

$$\begin{aligned}\mathbb{E}[||\mathbf{v}_0 - \nabla f(\mathbf{w}_0)||^2] &= \mathbb{E}\left[||\frac{1}{b_{out}}\sum_{i=1}^{b_{out}} \nabla f_i(\mathbf{w}_0; \xi_i) - \nabla f(\mathbf{w}_0)||^2\right] \\ &= \frac{1}{b_{out}}\left[\mathbf{E}[||\nabla f_i(\mathbf{w}_0; \xi_i)||] - ||\nabla f(\mathbf{w}_0)||^2\right].\end{aligned} \tag{38}$$

Combining (37) and (38), we have

$$\begin{aligned}\mathbb{E}[||\nabla f(\tilde{\mathbf{w}}_k)||^2] &\leq \frac{2}{(m+1)\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \frac{1}{b_{out}}\left[\mathbf{E}[||\nabla f_i(\mathbf{w}_0; \xi_i)||] - ||\nabla f(\mathbf{w}_0)||^2\right] + \mu^2 \\ &\leq \frac{2}{(m+1)\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \frac{1}{b_{out}}\mathbf{E}[||\nabla f_i(\mathbf{w}_0; \xi_i)||] + \mu^2 \\ &\leq \frac{2}{(m+1)\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \frac{1}{b_{out}}\sigma + \mu^2 \\ &= \frac{2}{(m+1)\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \frac{\sigma}{\log m} + \mu^2.\end{aligned}$$

IX. PROOF OF REMARK 7

Remark 7: Suppose each $f_i(\cdot)$ is $\mathbf{L}$-smooth. Let $b_{out} = \log n$, $b_{in} = (1/4)n$ and $\eta = 1/(2\mathbf{L}^3(\sqrt{17}+1))$ where $m = n$ is the total number of iterations in inner loops. Then the total complexity of DM-SARAH to achieve a ϵ-approximate solution is $\mathcal{O}(\log n + n \cdot 2^{\sigma/\epsilon - 1})$.

Proof 9: Let $b_{out} = \log n$, $b_{in} = \frac{1}{4}n$ and $\eta = 1/(2\mathbf{L}^3(\sqrt{17}+1))$ where $m = n$ is the total number of iterations in inner loops. According to the **Theorem 6**, we have

$$\mathbb{E}[||\nabla f(\tilde{\mathbf{w}}_k)||^2] \leq \frac{2}{\frac{1}{2\mathbf{L}^3(\sqrt{17}+1)}(m+1)}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \frac{\sigma}{\log m} + \mu^2 \leq \epsilon.$$

Therefore, we have $m = \mathcal{O}(2^{\sigma/\epsilon})$. The complexity analysis of DM-SARAH is as follows:

- It needs $b_{out} = \log n$ gradients in one outer loop, i.e., its complexity is $\mathcal{O}(\log n)$;
- It needs $2mb_{in}$ gradient in m iterations of inner loops, i.e., its complexity is $\mathcal{O}(n \cdot 2^{\sigma/\epsilon - 1})$;
- It needs $b_{out} + 2mb_{in}$ gradient in one outer loop, i.e., its total complexity is $\mathbf{O}(\log n + n \cdot 2^{\sigma/\epsilon - 1})$.

Proof is complete.

X. PROOF OF THEOREM 8

Theorem 8: Suppose each $f_i(\cdot)$ is $\mathbf{L}$-smooth and convex. Let $\mathbb{E}[||\nabla f_i(\mathbf{w}_0)||] \leq \sigma$ and $b_{out} = \sqrt{m+1}$, then for any $k \geq 0$, it holds

$$\begin{aligned}\mathbb{E}[||\nabla f(\tilde{\mathbf{w}}_k)||^2] \leq & \frac{2}{(m+1)\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] \\ & + \left[\frac{4\mathbf{L}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + 2\sigma^2}{\sqrt{m+1}}\right] + \mu^2,\end{aligned}$$

where $f(\mathbf{w}^*)$ is the optimal value.

Proof 10: On the one hand, similar to the proof of **Theorem 6**, we also have

$$\mathbb{E}[||\nabla f(\tilde{\mathbf{w}}_k)||^2] \leq \frac{2}{(m+1)\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \mathbb{E}[||\nabla f(\mathbf{w}_0) - \mathbf{v}_0||^2] + \mu^2.$$

On the other hand, it can be rewritten as follows.

$$\begin{aligned}\mathbb{E}[||\nabla f(\tilde{\mathbf{w}}_k)||^2] \leq & \frac{2}{(m+1)\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] \\ & + \left[\frac{4\mathbf{L}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + 2\mathbb{E}[||\nabla f_i(\mathbf{w}^*;\xi_i)||^2]}{b_{out}} - \frac{||\nabla f(\mathbf{w}_0)||^2}{b_{out}}\right] + \mu^2 \\ \leq & \frac{2}{(m+1)\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \left[\frac{4\mathbf{L}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + 2\mathbb{E}[||\nabla f_i(\mathbf{w}^*;\xi_i)||^2]}{b_{out}}\right] + \mu^2 \\ \leq & \frac{2}{(m+1)\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \left[\frac{4\mathbf{L}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + 2\mathbb{E}[||\nabla f_i(\mathbf{w}_0;\xi_i)||^2]}{b_{out}}\right] + \mu^2 \\ \leq & \frac{2}{(m+1)\eta}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \left[\frac{4\mathbf{L}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + 2\sigma^2}{\sqrt{m+1}}\right] + \mu^2.\end{aligned} \tag{39}$$

XI. PROOF OF REMARK 9

Remark 9: Suppose each function $f_i(\cdot)$ is $\mathbf{L}$-smooth and convex. Let $b_{out} = \sqrt{m+1}$, $b_{in} = (1/4)n$ and $\eta = 1/(2\mathbf{L}^3(\sqrt{17}+1))$ where $m = n$ is the total number of iterations in inner loops. Then the total complexity of DM-SARAH to achieve a ϵ-approximate solution is $\mathcal{O}(\frac{\sigma^2}{\epsilon} + \frac{n\sigma^4}{2\epsilon^2})$.

Proof 11: Let $b_{out} = \sqrt{m+1}$, $b_{in} = \frac{1}{4}n$ and $\eta = 1/(2\mathbf{L}^3(\sqrt{17}+1))$ where $m = n$ is the total number of iterations in inner loops. According to the **Theorem 8**, we have

$$\mathbb{E}[||\nabla f(\tilde{\mathbf{w}}_k)||^2] \leq \frac{2}{\frac{1}{2\mathbf{L}^3(\sqrt{17}+1)}(m+1)}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \left[\frac{4\mathbf{L}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + 2\sigma^2}{\sqrt{m+1}}\right] + \mu^2 \leq \epsilon.$$

Therefore, we have $m = \mathcal{O}(\sigma^4/\epsilon^2)$. The complexity analysis of DM-SARAH is as follows:

- It needs $b_{out} = \sqrt{m+1}$ gradients in one outer loop, i.e., its complexity is $\mathcal{O}(\frac{\sigma^2}{\epsilon})$;
- It needs $2mb_{in}$ gradient in m iterations of inner loops, i.e., its complexity is $\mathcal{O}(\frac{n\sigma^4}{2\epsilon^2})$;
- It needs $b_{out} + 2mb_{in}$ gradient in one outer loop, i.e., its total complexity is $\mathcal{O}(\frac{\sigma^2}{\epsilon} + \frac{n\sigma^4}{2\epsilon^2})$.

Proof is complete.

XII. PROOF OF REMARK 10

Remark 10: Suppose each function $f_i(\cdot)$ is α-strongly convex and $\mathbf{L}$-smooth. Let $b_{out} = \sqrt{m+1}$, $b_{in} = (1/4)n$ and $\eta = 1/(2\mathbf{L}^3(\sqrt{17}+1))$ where $m = n$ is the total number of iterations in inner loops. Then the total complexity of DM-SARAH to achieve a ϵ-approximate solution is $\mathcal{O}(\frac{\sigma^2\kappa}{\epsilon} + \frac{n\sigma^2\kappa}{2\epsilon^2})$.

Proof 12: Let $b_{out} = \sqrt{m+1}$, $b_{in} = \frac{1}{4}n$ and $\eta = 1/(2\mathbf{L}^3(\sqrt{17}+1))$ where $m = n$ is the total number of iterations in inner loops. According to the Corollary, we have

$$\begin{aligned}\mathbb{E}[||\nabla f(\tilde{\mathbf{w}}_k)||^2] \leq & \frac{2}{\frac{1}{2\mathbf{L}^3(\sqrt{17}+1)}(m+1)}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \frac{2\sigma^2}{\sqrt{m+1}}\left[\frac{\mathbf{L}}{\alpha} + 1\right] + \mu^2 \\ = & \frac{4\mathbf{L}^3(\sqrt{17}+1)}{m+1}\mathbb{E}[f(\mathbf{w}_0) - f(\mathbf{w}^*)] + \frac{2\sigma^2}{\sqrt{m+1}}[\kappa + 1] + \mu^2 \leq \epsilon.\end{aligned}$$

Therefore, we have $m = \mathcal{O}(\sigma^4\kappa^2/\epsilon^2)$ where $\kappa = \mathbf{L}/\alpha$ is a condition number. The complexity analysis of DM-SARAH is as follows:

- It needs $b_{out} = \sqrt{m+1}$ gradients in one outer loop, i.e., its complexity is $\mathcal{O}(\frac{\sigma^2\kappa}{\epsilon})$;
- It needs $2mb_{in}$ gradient in m iterations of inner loops, i.e., its complexity is $\mathcal{O}(\frac{n\sigma^2\kappa}{2\epsilon^2})$;
- It needs $sb_{out} + 2msb_{in}$ gradient in s outer loops, i.e., its total complexity is $\mathcal{O}(\frac{\sigma^2\kappa}{\epsilon} + \frac{n\sigma^2\kappa}{2\epsilon^2})$.

Proof is complete.

References

[1] L. M. Nguyen, N. H. Nguyen, D. T. Phan, J. R. Kalagnanam, and K. Scheinberg, "When does stochastic gradient algorithm work well?" *stat*, vol. 1050, p. 18, 2018.