A THEORY

A.1 PROOF OF LEMMA 1

Proof. Let $m = m_k$. By differentiability of f and f_m, we have

$$f(\mathbf{w}_{k+1}) - f(\mathbf{w}_k) = \int_0^1 \langle \nabla f(\mathbf{w}_k + \tau(\mathbf{w}_{k+1} - \mathbf{w}_k)), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle d\tau$$

and

$$f_m(\mathbf{w}_{k+1}) - f_m(\mathbf{w}_k) = \int_0^1 \langle \nabla f_m(\mathbf{w}_k + \tau(\mathbf{w}_{k+1} - \mathbf{w}_k)), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle d\tau.$$

Denote, to simplify the expressions, $\mathbf{w}(\tau) \stackrel{\text{def}}{=} \mathbf{w}_k + \tau(\mathbf{w}_{k+1} - \mathbf{w}_k)$, then we get

$$\begin{aligned} f(\mathbf{w}_{k+1}) - f(\mathbf{w}_k) &= f_m(\mathbf{w}_{k+1}) - f_m(\mathbf{w}_k) + \int_0^1 \langle \nabla f(\mathbf{w}(\tau)) - \nabla f_m(\mathbf{w}(\tau)), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle d\tau \\ &= f_m(\mathbf{w}_{k+1}) - f_m(\mathbf{w}_k) + \langle \mathbf{v}_k - \nabla f_m(\mathbf{w}_k), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle \\ &\quad + \int_0^1 \langle \nabla f(\mathbf{w}(\tau)) - \nabla f_m(\mathbf{w}(\tau)) - \mathbf{v}_k + \nabla f_m(\mathbf{w}_k), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle d\tau. \end{aligned}$$

Define the subproblem solved by SABER at iteration k as $\phi_k(\mathbf{w}) = f_m(\mathbf{w}) + \langle \mathbf{v}_k - \nabla f_m(\mathbf{w}_k), \mathbf{w} - \mathbf{w}_k \rangle + \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}_k\|^2$. Then, it holds $\phi_k(\mathbf{w}_{k+1}) \leq \phi_k(\mathbf{w}_k)$ and

$$f_m(\mathbf{w}_{k+1}) - f_m(\mathbf{w}_k) + \langle \mathbf{v}_k - \nabla f_m(\mathbf{w}_k), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle \leq -\frac{1}{2\eta}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2.$$

Let us split the integral into two parts:

$$\begin{aligned} &\int_0^1 \langle \nabla f(\mathbf{w}(\tau)) - \nabla f_m(\mathbf{w}(\tau)) - \mathbf{v}_k + \nabla f_m(\mathbf{w}_k), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle d\tau \\ &\quad = \int_0^1 \langle \nabla f(\mathbf{w}(\tau)) - \nabla f_m(\mathbf{w}(\tau)) - \nabla f(\mathbf{w}_k) + \nabla f_m(\mathbf{w}_k), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle d\tau \\ &\qquad + \langle \nabla f(\mathbf{w}_k) - \mathbf{v}_k, \mathbf{w}_{k+1} - \mathbf{w}_k \rangle. \end{aligned}$$

The first part can be upper bounded using the data heterogeneity assumption:

$$\begin{aligned} &\int_0^1 \langle \nabla f(\mathbf{w}(\tau)) - \nabla f_m(\mathbf{w}(\tau)) - \nabla f(\mathbf{w}_k) + \nabla f_m(\mathbf{w}_k), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle d\tau \\ &\quad \leq \int_0^1 \|\nabla f(\mathbf{w}(\tau)) - \nabla f_m(\mathbf{w}(\tau)) - \nabla f(\mathbf{w}_k) + \nabla f_m(\mathbf{w}_k)\| \|\mathbf{w}_{k+1} - \mathbf{w}_k\| d\tau \\ &\quad \leq \int_0^1 \delta \|\mathbf{w}(\tau) - \mathbf{w}_k\| \|\mathbf{w}_{k+1} - \mathbf{w}_k\| d\tau = \delta \int_0^1 \tau \|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 d\tau = \frac{\delta}{2}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 \\ &\quad \stackrel{\eta \leq \frac{1}{4\delta}}{\leq} \frac{1}{8\eta}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2. \end{aligned}$$

For the second part, we use Young's inequality

$$\langle \nabla f(\mathbf{w}_k) - \mathbf{v}_k, \mathbf{w}_{k+1} - \mathbf{w}_k \rangle \leq 2\eta \|\nabla f(\mathbf{w}_k) - \mathbf{v}_k\|^2 + \frac{1}{8\eta}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2.$$

Thus, combining the bounds on the integral with the initial inequalities, we obtain

$$\begin{aligned} f(\mathbf{w}_{k+1}) - f(\mathbf{w}_k) &\leq -\frac{1}{2\eta}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 + \frac{1}{8\eta}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 + 2\eta\|\mathbf{v}_k - \nabla f(\mathbf{w}_k)\|^2 \\ &\quad + \frac{1}{8\eta}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 \\ &= -\frac{1}{4\eta}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 + 2\eta\|\mathbf{v}_k - \nabla f(\mathbf{w}_k)\|^2. \end{aligned}$$

□

A.2 PROOF OF LEMMA 2

Proof. Denote by j the index sampled to update $\mathbf{v}_{k+1}$, i.e., $j = m_{k+1}$. With probability p, we have $\mathbf{v}_{k+1} = \nabla f(\mathbf{w}_{k+1})$, so

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{w}_{k+1}) - \mathbf{v}_{k+1}\|^2\right] \\
&= p \cdot 0 + (1-p)\mathbb{E}\left[\|\nabla f(\mathbf{w}_{k+1}) - \mathbf{v}_k - \nabla f_j(\mathbf{w}_{k+1}) + \nabla f_j(\mathbf{w}_k)\|^2\right] \\
&= (1-p)\mathbb{E}\left[\|\nabla f(\mathbf{w}_k) - \mathbf{v}_k + \nabla f(\mathbf{w}_{k+1}) - \nabla f(\mathbf{w}_k) - \nabla f_j(\mathbf{w}_{k+1}) - \nabla f_j(\mathbf{w}_k)\|^2\right] \\
&= (1-p)\mathbb{E}\left[\|\nabla f(\mathbf{w}_k) - \mathbf{v}_k\|^2\right] \\
&\quad + (1-p)\mathbb{E}\left[2\langle \nabla f(\mathbf{w}_k) - \mathbf{v}_k, \nabla f(\mathbf{w}_{k+1}) - \nabla f(\mathbf{w}_k) - \nabla f_j(\mathbf{w}_{k+1}) - \nabla f_j(\mathbf{w}_k)\rangle\right] \\
&\quad + (1-p)\mathbb{E}\left[\|\nabla f(\mathbf{w}_{k+1}) - \nabla f(\mathbf{w}_k) - \nabla f_j(\mathbf{w}_{k+1}) - \nabla f_j(\mathbf{w}_k)\|^2\right].
\end{aligned}$$

Since j is sampled after we have produced $\mathbf{w}_{k+1}$, it is independent of $\mathbf{w}_{k+1}$, and it holds

$$\mathbb{E}\left[\langle \nabla f(\mathbf{w}_k) - \mathbf{v}_k, \nabla f(\mathbf{w}_{k+1}) - \nabla f(\mathbf{w}_k) - \nabla f_j(\mathbf{w}_{k+1}) - \nabla f_j(\mathbf{w}_k)\rangle\right] = 0.$$

Moreover, by second-order data heterogeneity, we have

$$\mathbb{E}\left[\|\nabla f(\mathbf{w}_{k+1}) - \nabla f(\mathbf{w}_k) - \nabla f_j(\mathbf{w}_{k+1}) - \nabla f_j(\mathbf{w}_k)\|^2\right] \leq \delta^2 \mathbb{E}\left[\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2\right].$$

Putting the pieces together yields the claim. □

A.3 PROOF OF LEMMA 3

Proof. The first-order optimality condition for the problem in the definition of $\mathbf{w}_{k+1}$ writes

$$\nabla \phi_k(\mathbf{w}_{k+1}) = 0 = \nabla f_m(\mathbf{w}_{k+1}) + \mathbf{v}_k - \nabla f_m(\mathbf{w}_k) + \frac{1}{\eta}(\mathbf{w}_{k+1} - \mathbf{w}_k),$$

where $m = m_k$. From this equation, we obtain

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

By Cauchy-Schwarz inequality, it holds $\|a+b+c\|^2 \leq 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$ for any vectors $a, b, c \in \mathbb{R}^d$. Rearranging, it also implies $\|a\|^2 \geq \frac{1}{3}\|a+b+c\|^2 - \|b\|^2 - \|c\|^2$, which in our case gives

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2\right] &\geq \frac{\eta^2}{3}\mathbb{E}\left[\|\nabla f(\mathbf{w}_{k+1})\|^2\right] - \eta^2 \mathbb{E}\left[\|\mathbf{v}_k - \nabla f(\mathbf{w}_k)\|^2\right] \\
&\quad - \eta^2 \mathbb{E}\left[\|\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}_{k+1}) + \nabla f_m(\mathbf{w}_{k+1}) - \nabla f_m(\mathbf{w}_k)\|^2\right] \\
&\overset{(2)}{\geq} \frac{\eta^2}{3}\mathbb{E}\left[\|\nabla f(\mathbf{w}_{k+1})\|^2 - \eta^2 \|\mathbf{v}_k - \nabla f(\mathbf{w}_k)\|^2 - \eta^2\delta^2 \|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2\right].
\end{aligned}$$

Notice that $\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2$ appears in both sides, so we can rearrange and divide by $1 + \eta^2\delta^2$:

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2\right] &\geq \frac{1}{1+\eta^2\delta^2}\mathbb{E}\left[\frac{\eta^2}{3}\|\nabla f(\mathbf{w}_{k+1})\|^2 - \eta^2 \|\mathbf{v}_k - \nabla f(\mathbf{w}_k)\|^2\right] \\
&\overset{\eta \leq \frac{1}{3\delta}}{\geq} \frac{9}{10}\mathbb{E}\left[\frac{\eta^2}{3}\|\nabla f(\mathbf{w}_{k+1})\|^2 - \eta^2 \|\mathbf{v}_k - \nabla f(\mathbf{w}_k)\|^2\right] \\
&\geq \mathbb{E}\left[\frac{\eta^2}{4}\|\nabla f(\mathbf{w}_{k+1})\|^2 - \eta^2 \|\mathbf{v}_k - \nabla f(\mathbf{w}_k)\|^2\right].
\end{aligned}$$

□

A.4 PROOF OF THEOREM 1

Proof. Recall that we define a Lyapunov function

$$\mathcal{L}_k \stackrel{\text{def}}{=} f(\mathbf{w}_k) + c\|\mathbf{v}_k - \nabla f(\mathbf{w}_k)\|^2,$$

where we will choose $c > 0$ later in the proof. Lemmas 1 and 2 already bound the first and the second terms in $\mathcal{L}_{k+1}$ correspondingly, giving us the following recursion:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_{k+1}\right] &\leq f(\mathbf{w}_k) + \mathbb{E}\left[-\frac{1}{2\eta}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 + 2\eta\|\nabla f(\mathbf{w}_k) - \mathbf{v}_k\|^2\right] \\
&\quad + c(1-p)\|\nabla f(\mathbf{w}_k) - \mathbf{v}_k\|^2 + c\delta^2\mathbb{E}\left[\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2\right] \\
&= \mathcal{L}_k - c\|\nabla f(\mathbf{w}_k) - \mathbf{v}_k\|^2 + \left(c\delta^2 - \frac{1}{2\eta}\right)\mathbb{E}\left[\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2\right] \\
&\quad + (2\eta + c(1-p))\|\nabla f(\mathbf{w}_k) - \mathbf{v}_k\|^2 \\
&= \mathcal{L}_k + \left(c\delta^2 - \frac{1}{2\eta}\right)\mathbb{E}\left[\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2\right] + (2\eta - cp)\|\nabla f(\mathbf{w}_k) - \mathbf{v}_k\|^2.
\end{aligned}$$

Let us set $c = \frac{3\eta}{p}$ to make the last term negative. Then, we obtain

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_{k+1}\right] &\leq \mathcal{L}_k + \left(\frac{3\eta\delta^2}{p} - \frac{1}{2\eta}\right)\mathbb{E}\left[\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2\right] - \eta\|\nabla f(\mathbf{w}_k) - \mathbf{v}_k\|^2 \\
&\stackrel{\eta \leq \frac{\sqrt{p}}{4\delta}}{\leq} \mathcal{L}_k - \frac{1}{4\eta}\mathbb{E}\left[\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2\right] - \eta\|\mathbf{v}_k - \nabla f(\mathbf{w}_k)\|^2 \\
&\stackrel{(4)}{\leq} \mathcal{L}_k - \frac{\eta}{16}\mathbb{E}\left[\|\nabla f(\mathbf{w}_{k+1})\|^2\right] + \frac{\eta}{4}\|\mathbf{v}_k - \nabla f(\mathbf{w}_k)\|^2 - \eta\|\mathbf{v}_k - \nabla f(\mathbf{w}_k)\|^2 \\
&\leq \mathcal{L}_k - \frac{\eta}{16}\mathbb{E}\left[\|\nabla f(\mathbf{w}_{k+1})\|^2\right].
\end{aligned}$$

Recurring this to $\mathcal{L}_0 = f(\mathbf{w}_0) + c\|\mathbf{v}_0 - \nabla f(\mathbf{w}_0)\|^2 = f(\mathbf{w}_0)$, we get

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{w}_k)\|^2\right] \leq \frac{16}{\eta K}\left(\mathcal{L}_0 - \mathbb{E}\left[\mathcal{L}_K\right]\right) \leq \frac{16(f(\mathbf{w}_0) - f_*)}{\eta K},$$

where we used the fact that $\mathcal{L}_K = f(\mathbf{w}_K) + c\|\mathbf{v}_K - \nabla f(\mathbf{w}_K)\|^2 \geq f(\mathbf{w}_k) \geq f_*$. □