A. Proof of Theorem 4

Theorem 4. *Suppose that f and $\mathcal{D}$ satisfies (A1), (A2), (A3), and (A4). Further, suppose $\|\nabla f(\vec{w},\xi)\| \le \mathfrak{g}$ for all $\vec{w}$ with probability 1, and that $F(\vec{w}) \le M$ for all $\vec{w}$ for some M. Then Algorithm 2 guarantees:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(\vec{w}_t)\|\right] \le \tilde{O}\left(\frac{1}{\sqrt{T}}+\frac{\sigma^{4/7}}{T^{2/7}}\right)$$

where the $\tilde{O}$ notation hides constants that depend on R, G, and M and a factor of $\log(T)$.

Proof. Notice from the definition of α that we always have

$$\begin{aligned}\alpha_t &= \frac{1}{C^2 t^{1/7} G_{t-1}^{3/7}}\\ &\le \frac{1}{C^2 D^{3/7}}\\ &\le 1\end{aligned}$$

where we defined $G_0 = D$. Thus $\alpha_t \le 1$ always.

We begin with the now-familiar definitions:

$$\begin{aligned}\epsilon_t &= \nabla f(\vec{x}_t,\xi_t) - \nabla F(\vec{x}_t)\\ \epsilon'_t &= \nabla f(\vec{x}_t,\xi'_t) - \nabla F(\vec{x}_t)\\ \hat{\epsilon}_t &= \vec{m}_t - \nabla F(\vec{w}_t)\end{aligned}$$

Notice that $\mathbb{E}[\langle \epsilon_i, \epsilon_j\rangle] = \sigma^2\delta_{i,j}$. Now we write the recursive formulation for $\hat{\epsilon}_{t+1}$:

$$\begin{aligned}\vec{m}_t &= (1-\alpha_t)\vec{m}_{t-1} + \alpha_t\nabla f(\vec{x}_t,\xi_t)\\ &= (1-\alpha_t)(\nabla F(\vec{w}_{t-1}) + \hat{\epsilon}_{t-1}) + \alpha_t(\nabla F(\vec{w}_t) + \epsilon_t)\\ &= \nabla F(\vec{w}_t) + (1-\alpha_t)Z(\vec{w}_{t-1},\vec{w}_t) + \alpha_t Z(\vec{x}_t,\vec{w}_t) + (1-\alpha_t)\hat{\epsilon}_{t-1} + \alpha\epsilon_t\\ \hat{\epsilon}_t &= (1-\alpha_t)Z(\vec{w}_{t-1},\vec{w}_t) + \alpha_t Z(\vec{x}_t,\vec{w}_t) + (1-\alpha_t)\hat{\epsilon}_{t-1} + \alpha_t\epsilon_t\end{aligned}$$

Unfortunately, it is no longer clear how to unroll this recurrence to solve for $\hat{\epsilon}_t$ in a tractable manner. Instead, we will take a different path, inspired by the potential function analysis of (Cutkosky & Orabona, 2019). Start with the observations:

$$\begin{aligned}\alpha_t\|Z(\vec{x}_t,\vec{w}_t)\| &\le \rho\frac{(1-\alpha_t)^2\eta_{t-1}^2}{\alpha_t} \le \rho\frac{(1-\alpha_t)\eta_{t-1}^2}{\alpha_t}\\ (1-\alpha_t)\|Z(\vec{w}_{t-1},\vec{w}_t)\| &\le (1-\alpha_t)\rho\eta_{t-1}^2 \le \rho\frac{(1-\alpha_t)\eta_{t-1}^2}{\alpha_t}\end{aligned}$$

Define $K_t = \frac{1}{\alpha_t^2\eta_{t-1}G_t}$. Then we use $(a+b)^2 \le (1+1/x)a^2 + (1+x)b^2$ for all x and the fact that ϵ_t is uncorrelated with anything that does not depend on ξ_t to obtain:

$$\begin{aligned}\mathbb{E}[K_t\|\hat{\epsilon}_t\|^2] &\le \mathbb{E}\left[K_t(1+1/x)4\rho^2\frac{(1-\alpha_t)^2\eta_{t-1}^4}{\alpha_t^2} + K_t(1+x)(1-\alpha_t)^2\|\epsilon_{t-1}\|^2 + K_t\alpha_t^2\|\epsilon_t\|^2\right]\\ &\le \mathbb{E}\left[K_t\rho^2\frac{8(1-\alpha_t)^2\eta_{t-1}^4}{\alpha_t^3} + K_t(1-\alpha_t)\|\epsilon_{t-1}\|^2 + K_t\alpha_t^2\|\epsilon_t\|^2\right]\end{aligned}$$

where in the last inequality we have set $x = \alpha_t$. This implies:

$$\begin{aligned}&\mathbb{E}[K_t\|\hat{\epsilon}_t\|^2 - K_{t-1}\|\hat{\epsilon}_{t-1}\|^2] \le \mathbb{E}\left[K_t\rho^2\frac{8(1-\alpha_t)^2\eta_{t-1}^4}{\alpha_t^3} + (K_t(1-\alpha_t) - K_{t-1})\|\epsilon_{t-1}\|^2 + K_t\alpha_t^2\|\epsilon_t\|^2\right]\\ &\le \mathbb{E}\left[\rho^2\frac{8(1-\alpha_t)^2\eta_{t-1}^3}{\alpha_t^5 G_t} + \frac{\|\epsilon_t\|^2}{G_t\eta_{t-1}} - \left(\frac{1}{\alpha_{t-1}^2 G_{t-1}\eta_{t-2}} - \frac{1}{\alpha_t^2 G_t\eta_{t-1}} + \frac{1}{\alpha_t G_t\eta_{t-1}}\right)\|\epsilon_{t-1}\|^2\right]\end{aligned}$$

Let $\delta_t = G_t - G_{t-1}$. Then we have $\mathbb{E}[\epsilon_t/\sqrt{G_t\eta_{t-1}}] = 0 = \mathbb{E}[\epsilon'_t/\sqrt{G_t\eta_{t-1}}]$. Therefore we have

$$\mathbb{E}\left[\frac{\|\epsilon_t\|^2}{G_t}\right] \le \mathbb{E}\left[\frac{\|\nabla f(\vec{x}_t,\xi_t)\|^2}{G_t}\right]$$
$$\mathbb{E}\left[\frac{\|\epsilon_t\|^2}{G_t}\right] \le \mathbb{E}\left[\frac{\|\nabla f(\vec{x}_t,\xi_t) - \nabla f(\vec{x}_t,\xi'_t)\|^2}{G_t}\right]$$

so that we have

$$\mathbb{E}\left[\frac{\|\epsilon_t\|^2}{G_t}\right] \le \mathbb{E}\left[\frac{\delta_{t+1}}{G_t}\right]$$

Now, observe that $\delta_{t+1} \le 2\mathfrak{g}^2$, so that we have

$$\begin{aligned}
\mathbb{E}\left[\frac{\delta_{t+1}}{G_t\eta_{t-1}}\right] &= \mathbb{E}\left[\frac{\delta_{t+1}/\eta_{t-1}}{D + 3\mathfrak{g}^2 + \sum_{\tau=1}^t \delta_t}\right]\\
&\le \mathbb{E}\left[\frac{1}{\eta_T}\frac{\delta_{t+1}}{D + 2\mathfrak{g}^2 + \sum_{\tau=1}^{t+1} \delta_t}\right]\\
&\le \mathbb{E}\left[\frac{1}{\eta_T}\frac{\delta_{t+1}}{D + \sum_{\tau=1}^{t+1} \delta_t}\right]\\
\sum_{t=2}^{T+1}\mathbb{E}\left[\frac{\|\epsilon_t\|^2}{G_t\eta_{t-1}}\right] &\le \mathbb{E}\left[\frac{1}{\eta_T}\log\left(\frac{G_{T+1}}{D}\right)\right]
\end{aligned}$$

where we have used the fact that η_t is non-increasing.

Next, we tackle $\rho^2\frac{8(1-\alpha_t)^2\eta_{t-1}^3}{\alpha_t^5 G_t}$. We have

$$\begin{aligned}
\sum_{t=2}^{T+1}\mathbb{E}\left[\frac{\eta_{t-1}^3}{\alpha_t^5 G_t}\right] &\le \sum_{t=2}^{T+1}\mathbb{E}\left[\frac{\eta_{t-1}^4}{\eta_T\alpha_t^5 G_{t-1}}\right]\\
&\le \sum_{t=2}^{T+1}\mathbb{E}\left[\frac{C^{14}}{t\eta_T}\right]\\
&\le C^{14}\log(T+2)\,\mathbb{E}[\eta_T^{-1}]
\end{aligned}$$

Now, finally we turn to bounding $-\left(\frac{1}{\alpha_{t-1}^2 G_{t-1}\eta_{t-2}} - \frac{1}{\alpha_t^2 G_t\eta_{t-1}} + \frac{1}{\alpha_t G_t\eta_{t-1}}\right)\|\epsilon_{t-1}\|^2$. To do this, we first upper-bound $\frac{1}{\alpha_t^2 G_t\eta_{t-1}} - \frac{1}{\alpha_{t-1}^2 G_{t-1}\eta_{t-2}}$. Note that:

$$\frac{1}{\alpha_t^2 G_t\eta_{t-1}} - \frac{1}{\alpha_{t-1}^2 G_{t-1}\eta_{t-2}} \le \frac{1}{\alpha_t^2 G_t\eta_{t-1}} - \frac{1}{\alpha_{t-1}^2 G_t\eta_{t-2}}$$

So now we can upper bound $\frac{1}{\alpha_t^2\eta_{t-1}} - \frac{1}{\alpha_{t-1}^2\eta_{t-2}}$ and divide the bound by G_t.

$$\begin{aligned}
\frac{1}{\alpha_t^2\eta_{t-1}} - \frac{1}{\alpha_{t-1}^2\eta_{t-2}} &= t^2\eta_{t-1}^3 G_{t-1}^2 - (t-1)^2\eta_{t-2}^3 G_{t-2}^2\\
&= C^3(t^{5/7}G_{t-1}^{8/7} - (t-1)^{5/7}G_{t-2}^{8/7})\\
&\le C^3 t^{5/7}(G_{t-1}^{8/7} - G_{t-2}^{8/7}) + C^3(t^{5/7} - (t-1)^{5/7})G_{t-1}^{8/7}
\end{aligned}$$

Next, we analyze $G_{t-1}^{8/7} - G_{t-2}^{8/7}$. Recall our definition $\delta_t = G_t - G_{t-1}$, and we have $0 \le \delta_t \le 2\mathfrak{g}^2$ for all t. Then by convexity of the function $x \mapsto x^{8/7}$, we have

$$G_{t-1}^{8/7} - G_{t-2}^{8/7} \le \frac{8\delta_{t-1}}{7}G_{t-1}^{1/7} \le \frac{16\mathfrak{g}^2}{7}G_{t-1}^{1/7}$$

Therefore we have

$$\frac{1}{\alpha_t^2\eta_{t-1}} - \frac{1}{\alpha_{t-1}^2\eta_{t-2}} \le \frac{16C^3\mathfrak{g}^2}{7}t^{5/7}G_{t-1}^{1/7} + C^3(t^{5/7}-(t-1)^{5/7})G_{t-1}^{8/7}$$
$$= \frac{16C^3\mathfrak{g}^2 t^{1/7}}{7G_{t-1}^{4/7}}t^{4/7}G_{t-1}^{5/7} + C^3(t^{5/7}-(t-1)^{5/7})G_{t-1}^{8/7}$$

Now use $G_{t-1} \ge \mathfrak{g}^2 t^{1/4}$,

$$\le \frac{16C^3\mathfrak{g}^{6/7}}{7}t^{4/7}G_{t-1}^{5/7} + C^3(t^{5/7}-(t-1)^{5/7})G_{t-1}^{8/7}$$
$$\le \frac{16C^3\mathfrak{g}^{6/7}}{7}t^{4/7}G_{t-1}^{5/7} + \frac{5C^3}{7(t-1)^{2/7}}G_{t-1}^{8/7}$$

Use $G_{t-1} \le D + 3\mathfrak{g}^2(t-1)$,

$$\le \frac{16C^3\mathfrak{g}^{6/7}}{7}t^{4/7}G_{t-1}^{5/7} + \frac{5C^3(D+3\mathfrak{g}^2(t-1))^{3/7}}{7(t-1)^{2/7}}G_{t-1}^{5/7}$$
$$\le \frac{21C^3\mathfrak{g}^{6/7}}{7}t^{4/7}G_{t-1}^{5/7} + \frac{5C^3D^{3/7}}{7(t-1)^{2/7}}G_{t-1}^{5/7}$$

Use the definition of D,

$$\le \frac{21C^3\mathfrak{g}^{6/7}}{7}t^{4/7}G_{t-1}^{5/7} + \frac{5C}{7(t-1)^{2/7}}G_{t-1}^{5/7}$$

Use $C \ge 1/\mathfrak{g}^{3/7}$,

$$\le \frac{26C^3\mathfrak{g}^{6/7}}{7}t^{4/7}G_{t-1}^{5/7}$$

Now observe that

$$\frac{26C^3\mathfrak{g}^{6/7}}{7}t^{4/7}G_{t-1}^{5/7} \le \frac{26C^2\mathfrak{g}^{6/7}}{7\alpha_t\eta_{t-1}}$$

So putting all this together, we have

$$-\left(\frac{1}{\alpha_{t-1}^2G_{t-1}\eta_{t-2}} - \frac{1}{\alpha_t^2G_t\eta_{t-1}} + \frac{1}{\alpha_tG_t\eta_{t-1}}\right) \le -\left(\frac{1}{\alpha_tG_t\eta_{t-1}} - \frac{26C^2\mathfrak{g}^{6/7}}{7\alpha_tG_t\eta_{t-1}}\right)$$

Then since we set C so that $\frac{26C^2\mathfrak{g}^{6/7}}{7} = 1/2$, we obtain:

$$-\left(\frac{1}{\alpha_{t-1}^2G_{t-1}\eta_{t-2}} - \frac{1}{\alpha_t^2G_t\eta_{t-1}} + \frac{1}{\alpha_tG_t\eta_{t-1}}\right) \le -\frac{1}{2\alpha_tG_t\eta_{t-1}}$$

Putting all this together, we have shown:

$$\sum_{t=1}^T \mathbb{E}[K_{t+1}\|\hat\epsilon_{t+1}\|^2 - K_t\|\hat\epsilon_t\|^2] \le \mathbb{E}\left[\frac{\log(T+2)}{\eta_T} + \frac{1}{\eta_T}\log\left(\frac{G_{T+1}}{D}\right) - \sum_{t=1}^T \frac{\|\hat\epsilon_t\|^2}{2\alpha_{t+1}G_{t+1}\eta_t}\right]$$

Now, define the potential $\Phi_t = \frac{3F(\vec w_{t+1})}{\eta_t} + K_{t+1}\|\hat\epsilon_{t+1}\|^2$. Then, by Lemma 2, we obtain:

$$\Phi_t - \Phi_{t-1} \le -\|\nabla F(\vec w_t)\| + 8\|\hat\epsilon_t\| + \frac{3L\eta_t}{2}$$
$$+ 3F(\vec w_t)\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + K_{t+1}\|\hat\epsilon_{t+1}\|^2 - K_t\|\hat\epsilon_t\|^2$$

So summing over t and taking expectations yields:

$$\mathbb{E}[\Phi_T - \Phi_0] \le \mathbb{E}\left[\sum_{t=1}^{T} \frac{3L\eta_t}{2} - \|\nabla F(\vec{w}_t)\| + \frac{3M}{\eta_T} + \sum_{t=1}^{T} 8\|\hat{\epsilon}_t\| - \frac{\|\hat{\epsilon}_t\|^2}{2\alpha_{t+1}G_{t+1}\eta_t} + \frac{1}{\eta_T}\log\left(\frac{G_{T+1}}{D}\right) + \frac{\log(T+2)}{\eta_T}\right]$$

Now, we examine the term $\sum_{t=1}^{T} 8\|\hat{\epsilon}_t\| - \frac{\|\hat{\epsilon}_t\|^2}{2\alpha_{t+1}G_{t+1}\eta_t}$. By Cauchy-Schwarz we have:

$$\sum_{t=1}^{T} 8\|\hat{\epsilon}_t\| \le 8\sqrt{\sum_{t=1}^{T} \frac{\|\hat{\epsilon}_t\|^2}{2\alpha_{t+1}G_{t+1}\eta_t} \sum_{t=1}^{T} 2\alpha_{t+1}G_{t+1}\eta_t}$$

Therefore

$$\begin{aligned}
\sum_{t=1}^{T} 8\|\hat{\epsilon}_t\| - \frac{\|\hat{\epsilon}_t\|^2}{2\alpha_{t+1}G_{t+1}\eta_t} &\le \sup_M 8\sqrt{M\sum_{t=1}^{T} 2\alpha_{t+1}G_{t+1}\eta_t} - M \\
&\le 32\sum_{t=1}^{T} \alpha_{t+1}G_{t+1}\eta_t \\
&= 32\sum_{t=1}^{T} \frac{1}{(t+1)\eta_t} \\
&\le 32\sum_{t=1}^{T} \frac{1}{(t+1)\eta_T} \\
&\le \frac{32(\log(T+1))}{\eta_T}
\end{aligned}$$

Finally, observe that since $G_t \ge \mathfrak{g}^2 t^{1/4}$, we have $\eta_t \le \frac{C}{\sqrt{T}}$. Therefore $\sum_{t=1}^{T} \eta_t \le 2C\sqrt{T}$. Putting all this together again, we have

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla F(\vec{w}_t)\|] \le \Phi_0 + 3LC\sqrt{T} + \mathbb{E}[\eta_T^{-1}]\left[3M + \log(2\mathfrak{g}^2(T+1)/D) + \log(T+2) + 32(\log(T+1))\right]$$

Observe that we have $\Phi_0 \le \frac{3M}{\eta_0} + K_1\mathfrak{g}^2$.

Let us define $Z = 3M + \log(2\mathfrak{g}^2(T+1)/D) + \log(T+2) + 32(\log(T)+1)$. Then we have

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla F(\vec{w}_t)\|] \le \Phi_1 + 3LC\sqrt{T} + \mathbb{E}[\eta_T^{-1}]Z$$

Now we look carefully at the definition of G_t and η_t. By Jensen inequality, we have

$$\begin{aligned}
\mathbb{E}[\eta_T^{-1}] &= \frac{1}{C}(T+1)^{3/7}\,\mathbb{E}\left[\left(D + 2\mathfrak{g}^2 + \mathfrak{g}T^{1/4} + \sum_{t=1}^{T}\|\nabla f(\vec{x}_t,\xi_t) - \nabla f(\vec{x}_t,\xi_t')\|^2\right)^{2/7}\right] \\
&\le \frac{(T+1)^{3/7}(D + 2\mathfrak{g}^2 + \mathfrak{g}T^{1/4} + 4T\sigma^2)^{2/7}}{C} \\
&= O\left(\sqrt{T} + \sigma^{4/7}T^{5/7}\right)
\end{aligned}$$

The Theorem statement now follows. □