Theorem 2 proof

Anonymous Author(s)
Affiliation
Address
email

Abstract

H1. *For any $t > 0$, the estimated parameter w_t stays within a $\ell_\infty-$ball. There exists a constant $W > 0$ such that $\|w_t\|_\infty \leq W$ almost surely.*

H2. *The function f is L-smooth (has L-Lipschitz gradients) w.r.t. the parameter w. There exists some constant $L > 0$ such that for $(w, \vartheta) \in \Theta^2$, $f(w) - f(\vartheta) - \nabla f(\vartheta)^\top (w - \vartheta) \leq \frac{L}{2} \|w - \vartheta\|^2$.*

We assume that the optimistic guess m_t at iteration t and the true gradient g_t are correlated:

H3. *For any $t > 0$, $0 < \langle m_t \,|\, g_t \rangle = a_t \|g_t\|^2$ with some $0 < a_t \leq 1$, where $\langle \, | \, \rangle$ denotes the inner product*

We make a classical assumption in nonconvex optimization [**?**] on the magnitude of the gradient:

H4. *There exists a constant $\mathsf{M} > 0$ such that for any w and ξ, it holds $\|\nabla f(w, \xi)\| < \mathsf{M}$.*

Lemma 1. *Assume H4, then the quantities defined in Algorithm **??** satisfy for any $w \in \Theta$ and $t > 0$, $\|\nabla f(w_t)\| < \mathsf{M}$, $\quad \|\theta_t\| < \mathsf{M}$ and $\|\hat{v}_t\| < \mathsf{M}^2$.*

Lemma 2. *Assume H4, a strictly positive and a sequence of constant stepsizes $\{\eta_t\}_{t>0}$, $(\beta_1, \beta_2) \in [0, 1]$, then the following holds:*

$$\sum_{t=1}^{T_{\mathsf{M}}} \eta_t^2 \mathbb{E}\left[\left\|\hat{v}_t^{-1/2}\theta_t\right\|_2^2\right] \leq \frac{\eta^2 d T_{\mathsf{M}} (1-\beta_1)}{(1-\beta_2)(1-\gamma)} \, . \tag{1}$$

Lemma 3. *Assume a strictly positive and non increasing sequence of stepsizes $\{\eta_t\}_{t>0}$, $\beta_1 < \beta_2 \in [0, 1)$, then the following holds:*

$$\overline{w}_{t+1} - \overline{w}_t \leq \frac{\beta_1}{1-\beta_1} \tilde{\theta}_{t-1} \left[\eta_{t-1}\hat{v}_{t-1}^{-1/2} - \eta_t \hat{v}_t^{-1/2}\right] - \eta_t \hat{v}_t^{-1/2} \tilde{g}_t \, ,$$

where $\tilde{\theta}_t = \theta_t + \beta_1 \theta_{t-1}$ and $\tilde{g}_t = g_t - \beta_1 m_t + \beta_1 g_{t-1} + m_{t+1}$.

1 Proof of Theorem ??

Proof Using H2 and the iterate $\overline{w}_t$ we have:

$$\begin{aligned} f(\overline{w}_{t+1}) \leq & f(\overline{w}_t) + \nabla f(\overline{w}_t)^\top (\overline{w}_{t+1} - \overline{w}_t) + \frac{L}{2}\|\overline{w}_{t+1} - \overline{w}_t\|^2 \\ \leq & f(\overline{w}_t) + \underbrace{\nabla f(w_t)^\top (\overline{w}_{t+1} - \overline{w}_t)}_{A} \\ & + \underbrace{\left(\nabla f(\overline{w}_t) - \nabla f(w_t)\right)^\top (\overline{w}_{t+1} - \overline{w}_t)}_{B} + \frac{L}{2}\|\overline{w}_{t+1} - \overline{w}_t\| \, . \end{aligned} \tag{2}$$

Term A. Using Lemma 3, we have that:

$$\begin{aligned}\nabla f(w_t)^\top(\overline{w}_{t+1}-\overline{w}_t) &\leq \nabla f(w_t)^\top\left[\frac{\beta_1}{1-\beta_1}\tilde{\theta}_{t-1}\left[\eta_{t-1}\hat{v}_{t-1}^{-1/2}-\eta_t\hat{v}_t^{-1/2}\right]-\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\right]\\ &\leq \frac{\beta_1}{1-\beta_1}\|\nabla f(w_t)\|\|\eta_{t-1}\hat{v}_{t-1}^{-1/2}-\eta_t\hat{v}_t^{-1/2}\|\|\tilde{\theta}_{t-1}\|-\nabla f(w_t)^\top\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\ ,\end{aligned}$$

where the inequality is due to trivial inequality for positive diagonal matrix. Using Lemma 1 and assumption H3 we obtain:

$$\nabla f(w_t)^\top(\overline{w}_{t+1}-\overline{w}_t) \leq \frac{\beta_1(1+\beta_1)}{1-\beta_1}\mathsf{M}^2[\|\eta_{t-1}\hat{v}_{t-1}^{-1/2}\|-\|\eta_t\hat{v}_t^{-1/2}\|]-\nabla f(w_t)^\top\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\ , \tag{3}$$

where we have used the fact that $\eta_t\hat{v}_t^{-1/2}$ is a diagonal matrix such that $\eta_{t-1}\hat{v}_{t-1}^{-1/2} \succcurlyeq \eta_t\hat{v}_t^{-1/2} \succcurlyeq 0$ (decreasing stepsize and $\max$ operator). Also note that:

$$\begin{aligned}-\nabla f(w_t)^\top\eta_t\hat{v}_t^{-1/2}\tilde{g}_t &= -\nabla f(w_t)^\top\eta_{t-1}\hat{v}_{t-1}^{-1/2}\bar{g}_t-\nabla f(w_t)^\top\left[\eta_t\hat{v}_t^{-1/2}-\eta_t\hat{v}_t^{-1/2}\right]\bar{g}_t\\ &\quad -\nabla f(w_t)^\top\eta_{t-1}\hat{v}_{t-1}^{-1/2}(\beta_1 g_{t-1}+m_{t+1})\\ &\leq -\nabla f(w_t)^\top\eta_{t-1}\hat{v}_{t-1}^{-1/2}\bar{g}_t+(1-a_t\beta_1)\mathsf{M}^2[\|\eta_{t-1}\hat{v}_{t-1}^{-1/2}\|-\|\eta_t\hat{v}_t^{-1/2}\|]\\ &\quad -\nabla f(w_t)^\top\eta_t\hat{v}_t^{-1/2}(\beta_1 g_{t-1}+m_{t+1})\ ,\end{aligned} \tag{4}$$

where we have used Lemma 1 on $\|g_t\|$ and where that $\tilde{g}_t = \bar{g}_t+\beta_1 g_{t-1}+m_{t+1} = g_t-\beta_1 m_t+\beta_1 g_{t-1}+m_{t+1}$. Plugging (4) into (3) yields:

$$\begin{aligned}&\nabla f(w_t)^\top(\overline{w}_{t+1}-\overline{w}_t)\\ &\leq -\nabla f(w_t)^\top\eta_{t-1}\hat{v}_{t-1}^{-1/2}\bar{g}_t+\frac{1}{1-\beta_1}(a_t\beta_1^2-2a_t\beta_1+\beta1)\mathsf{M}^2[\|\eta_{t-1}\hat{v}_{t-1}^{-1/2}\|-\|\eta_t\hat{v}_t^{-1/2}\|]\\ &-\nabla f(w_t)^\top\eta_t\hat{v}_t^{-1/2}(\beta_1 g_{t-1}+m_{t+1})\ .\end{aligned} \tag{5}$$

Term B. By Cauchy-Schwarz (CS) inequality we have:

$$(\nabla f(\overline{w}_t)-\nabla f(w_t))^\top(\overline{w}_{t+1}-\overline{w}_t) \leq \|\nabla f(\overline{w}_t)-\nabla f(w_t)\|\|\overline{w}_{t+1}-\overline{w}_t\|\ . \tag{6}$$

Using smoothness assumption H2:

$$\begin{aligned}\|\nabla f(\overline{w}_t)-\nabla f(w_t)\| &\leq L\|\overline{w}_t-w_t\|\\ &\leq L\frac{\beta_1}{1-\beta_1}\|w_t-\tilde{w}_{t-1}\|\ .\end{aligned} \tag{7}$$

By Lemma 3 we also have:

$$\begin{aligned}\overline{w}_{t+1}-\overline{w}_t &= \frac{\beta_1}{1-\beta_1}\tilde{\theta}_{t-1}\left[\eta_{t-1}\hat{v}_{t-1}^{-1/2}-\eta_t\hat{v}_t^{-1/2}\right]-\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\\ &= \frac{\beta_1}{1-\beta_1}\tilde{\theta}_{t-1}\eta_{t-1}\hat{v}_{t-1}^{-1/2}\left[I-(\eta_t\hat{v}_t^{-1/2})(\eta_{t-1}\hat{v}_{t-1}^{-1/2})^{-1}\right]-\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\\ &= \frac{\beta_1}{1-\beta_1}\left[I-(\eta_t\hat{v}_t^{-1/2})(\eta_{t-1}\hat{v}_{t-1}^{-1/2})^{-1}\right](\tilde{w}_{t-1}-w_t)-\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\ ,\end{aligned} \tag{8}$$

where the last equality is due to $\tilde{\theta}_{t-1}\eta_{t-1}\hat{v}_{t-1}^{-1/2} = \tilde{w}_{t-1}-w_t$ by construction of $\tilde{\theta}_t$. Taking the norms on both sides, observing $\|I-(\eta_t\hat{v}_t^{-1/2})(\eta_{t-1}\hat{v}_{t-1}^{-1/2})^{-1}\| \leq 1$ due to the decreasing stepsize and the construction of $\hat{v}_t$ and using CS inequality yield:

$$\|\overline{w}_{t+1}-\overline{w}_t\| \leq \frac{\beta_1}{1-\beta_1}\|\tilde{w}_{t-1}-w_t\|+\|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|\ . \tag{9}$$

We recall Young's inequality with a constant $\delta \in (0,1)$ as follows:

$$\langle X\,|\,Y\rangle \leq \frac{1}{\delta}\|X\|^2+\delta\|Y\|^2\ .$$

Plugging (7) and (9) into (6) returns:

$$
\begin{aligned}
\left(\nabla f(\overline{w}_t)-\nabla f(w_t)\right)^\top\left(\overline{w}_{t+1}-\overline{w}_t\right) \leq & L\frac{\beta_1}{1-\beta_1}\|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|\|w_t-\tilde{w}_{t-1}\| \\
& + L\left(\frac{\beta_1}{1-\beta_1}\right)^2\|\tilde{w}_{t-1}-w_t\|^2 \,.
\end{aligned}
$$

Applying Young's inequality with $\delta \to \frac{\beta_1}{1-\beta_1}$ on the product $\|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|\|w_t-\tilde{w}_{t-1}\|$ yields:

$$
\left(\nabla f(\overline{w}_t)-\nabla f(w_t)\right)^\top\left(\overline{w}_{t+1}-\overline{w}_t\right) \leq L\|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|^2 + 2L\left(\frac{\beta_1}{1-\beta_1}\right)^2\|\tilde{w}_{t-1}-w_t\|^2 \,. \tag{10}
$$

The last term $\frac{L}{2}\|\overline{w}_{t+1}-\overline{w}_t\|$ can be upper bounded using (9):

$$
\begin{aligned}
\frac{L}{2}\|\overline{w}_{t+1}-\overline{w}_t\|^2 &\leq \frac{L}{2}\left[\frac{\beta_1}{1-\beta_1}\|\tilde{w}_{t-1}-w_t\| + \|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|\right] \\
&\leq L\|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|^2 + 2L\left(\frac{\beta_1}{1-\beta_1}\right)^2\|\tilde{w}_{t-1}-w_t\|^2 \,.
\end{aligned} \tag{11}
$$

Plugging (5), (10) and (11) into (2) and taking the expectations on both sides give:

$$
\begin{aligned}
&\mathbb{E}\left[f(\overline{w}_{t+1}) + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}_t^2\|\eta_t\hat{v}_t^{-1/2}\| - \left(f(\overline{w}_t) + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}_t^2\|\eta_{t-1}\hat{v}_{t-1}^{-1/2}\|\right)\right] \\
&\leq \mathbb{E}\left[-\nabla f(w_t)^\top\eta_{t-1}\hat{v}_{t-1}^{-1/2}\overline{g}_t - \nabla f(w_t)^\top\eta_t\hat{v}_t^{-1/2}(\beta_1 g_{t-1}+m_{t+1})\right] \\
&+ \mathbb{E}\left[2L\|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|^2 + 4L\left(\frac{\beta_1}{1-\beta_1}\right)^2\|\tilde{w}_{t-1}-w_t\|^2\right] \,,
\end{aligned}
$$

where $\tilde{\mathsf{M}}_t^2 = (a_t\beta_1^2+\beta_1)\mathsf{M}^2$. Note that the expectation of $\tilde{g}_t$ conditioned on the filtration $\mathcal{F}_t$ reads as follows

$$
\mathbb{E}\left[\nabla f(w_t)^\top\overline{g}_t\right] = \mathbb{E}\left[\nabla f(w_t)^\top(g_t-\beta_1 m_t)\right] = (1-a_t\beta_1)\|\nabla f(w_t)\|^2 \,. \tag{12}
$$

Summing from $t=1$ to $t=T$ leads to

$$
\begin{aligned}
&\frac{1}{\mathsf{M}}\sum_{t=1}^{T_{\mathsf{M}}}\left((1-a_t\beta_1)\eta_{t-1}+(\beta_1+a_t)\eta_t\right)\|\nabla f(w_t)\|^2 \leq \\
&\mathbb{E}\left[f(\overline{w}_1) + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}_t^2\|\eta_0\hat{v}_0^{-1/2}\| - \left(f(\overline{w}_{T_{\mathsf{M}}+1}) + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}_t^2\|\eta_{T_{\mathsf{M}}}\hat{v}_{T_{\mathsf{M}}}^{-1/2}\|\right)\right] \\
&+ 2L\sum_{t=1}^{T_{\mathsf{M}}}\mathbb{E}\left[\|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|^2\right] + 4L\left(\frac{\beta_1}{1-\beta_1}\right)^2\sum_{t=1}^{T_{\mathsf{M}}}\mathbb{E}\left[\|\tilde{w}_{t-1}-w_t\|^2\right] \\
&\leq \mathbb{E}\left[\Delta f + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}_t^2\|\eta_0\hat{v}_0^{-1/2}\|\right] + 2L\sum_{t=1}^{T_{\mathsf{M}}}\mathbb{E}\left[\|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|^2\right] \\
&+ 4L\left(\frac{\beta_1}{1-\beta_1}\right)^2\sum_{t=1}^{T_{\mathsf{M}}}\mathbb{E}\left[\|\tilde{w}_{t-1}-w_t\|^2\right] \,,
\end{aligned} \tag{13}
$$

where we denote $\Delta f := f(\overline{w}_1) - f(\overline{w}_{T_{\mathsf{M}}+1})$. We note that by definition of $\hat{v}_t$, and a constant learning rate η_t, we have

$$
\begin{aligned}
\|\tilde{w}_{t-1}-w_t\|^2 &= \|\eta_{t-1}\hat{v}_{t-1}^{-1/2}(\theta_{t-1}+h_t)\|^2 \\
&= \|\eta_{t-1}\hat{v}_{t-1}^{-1/2}(\theta_{t-1}+\beta_1\theta_{t-2}+(1-\beta_1)m_t)\|^2 \\
&\leq \|\eta_{t-1}\hat{v}_{t-1}^{-1/2}\theta_{t-1}\|^2 + \|\eta_{t-2}\hat{v}_{t-2}^{-1/2}\beta_1\theta_{t-2}\|^2 + (1-\beta_1)^2\|\eta_{t-1}\hat{v}_{t-1}^{-1/2}m_t\|^2 \,.
\end{aligned}
$$

Using Lemma 2 we have

$$\sum_{t=1}^{T_{\mathsf{M}}} \mathbb{E}\left[\|\tilde{w}_{t-1} - w_t\|^2\right]$$
$$\leq (1+\beta_1^2)\frac{\eta^2 d T_{\mathsf{M}}(1-\beta_1)}{(1-\beta_2)(1-\gamma)} + (1-\beta_1)^2 \sum_{t=1}^{T_{\mathsf{M}}} \mathbb{E}[\|\eta_{t-1}\hat{v}_{t-1}^{-1/2} m_t\|] \ .$$

And thus, setting the learning rate to a constant value η, noting that $\frac{1}{(1-a_t\beta_1)+(\beta_1+a_t)}$ is a decreasing function for all $t > 0$ and is upper bounded by 1, injecting in (13) yields:

$$\mathbb{E}[\|\nabla f(w_T)\|^2] = \frac{1}{\sum_{j=1}^{T_{\mathsf{M}}} \eta_j} \sum_{t=1}^{T_{\mathsf{M}}} \eta_t \|\nabla f(w_t)\|^2$$
$$\leq \sum_{t=1}^{T_{\mathsf{M}}} \frac{\mathsf{M}}{(1-a_t\beta_1)+(\beta_1+a_t)} \frac{1}{\sum_{j=1}^{T_{\mathsf{M}}} \eta_j} \mathbb{E}\left[\Delta f + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}_t^2 \|\eta_0 \hat{v}_0^{-1/2}\|\right]$$
$$+ \frac{4L\left(\frac{\beta_1}{1-\beta_1}\right)^2 \mathsf{M}}{\sum_{j=1}^{T_{\mathsf{M}}} \eta_j}(1+\beta_1^2)\frac{\eta^2 d T_{\mathsf{M}}(1-\beta_1)}{(1-\beta_2)(1-\gamma)} \sum_{t=1}^{T_{\mathsf{M}}} \frac{1}{(1-a_t\beta_1)+(\beta_1+a_t)}$$
$$+ \frac{\mathsf{M}}{\sum_{j=1}^{T_{\mathsf{M}}} \eta_j}(1-\beta_1)^2 \sum_{t=1}^{T_{\mathsf{M}}} \mathbb{E}[\|\eta_{t-1}\hat{v}_{t-1}^{-1/2} m_t\|] \sum_{t=1}^{T_{\mathsf{M}}} \frac{1}{(1-a_t\beta_1)+(\beta_1+a_t)}$$
$$+ \frac{2L\mathsf{M}}{\sum_{j=1}^{T_{\mathsf{M}}} \eta_j} \sum_{t=1}^{T_{\mathsf{M}}} \mathbb{E}[\|\eta_t \hat{v}_t^{-1/2} \tilde{g}_t\|^2] \sum_{t=1}^{T_{\mathsf{M}}} \frac{1}{(1-a_t\beta_1)+(\beta_1+a_t)} \ ,$$

where T is a random termination number distributed according (**??**). Setting the stepsize to $\eta = \frac{1}{\sqrt{dT_{\mathsf{M}}}}$ yields :

$$\mathbb{E}[\|\nabla f(w_T)\|^2] \leq \sum_{t=1}^{T_{\mathsf{M}}} C_{1,t}\sqrt{\frac{d}{T_{\mathsf{M}}}} + \sum_{t=1}^{T_{\mathsf{M}}} C_{2,t}\frac{1}{T_{\mathsf{M}}} + \frac{\eta}{T_{\mathsf{M}}} \sum_{t=1}^{T_{\mathsf{M}}} D_{1,t}\mathbb{E}[\|\hat{v}_{t-1}^{-1/2} m_t\|] + \frac{\eta}{T_{\mathsf{M}}} \sum_{t=1}^{T_{\mathsf{M}}} D_{2,t}\mathbb{E}[\|\hat{v}_{t-1}^{-1/2} \tilde{g}_t\|] \ ,$$

where

$$C_{1,t} = \frac{\mathsf{M}}{(1-a_t\beta_1)+(\beta_1+a_t)}\Delta f + \frac{4L\left(\frac{\beta_1}{1-\beta_1}\right)^2 \mathsf{M}}{(1-a_t\beta_1)+(\beta_1+a_t)} \frac{(1+\beta_1^2)(1-\beta_1)}{(1-\beta_2)(1-\gamma)} \ ,$$
$$C_{2,t} = \frac{\mathsf{M}}{(1-\beta_1)\left((1-a_t\beta_1)+(\beta_1+a_t)\right)}(a_t\beta_1^2+\beta_1)\mathsf{M}^2\mathbb{E}[\|\hat{v}_0^{-1/2}\|] \ .$$

Simple case as in [?]: if $\beta_1 = 0$ then $\tilde{g}_t = g_t + m_{t+1}$ and $g_t = \theta_t$. Also using Lemma 2 we have that:

$$\sum_{t=1}^{T_{\mathsf{M}}} \eta_t^2 \mathbb{E}\left[\left\|\hat{v}_t^{-1/2} g_t\right\|_2^2\right] \leq \frac{\eta^2 d T_{\mathsf{M}}}{(1-\beta_2)} \ ;$$

which leads to the final bound:

$$\mathbb{E}[\|\nabla f(w_T)\|^2] \leq \sqrt{\frac{d}{T_{\mathsf{M}}}} \sum_{t=1}^{T_{\mathsf{M}}} \tilde{C}_{1,t} + \frac{1}{T_{\mathsf{M}}} \sum_{t=1}^{T_{\mathsf{M}}} \tilde{C}_{2,t} \ ,$$

where

$$\tilde{C}_{1,t} = C_{1,t} + \frac{\mathsf{M}}{(1-a_t\beta_1)+(\beta_1+a_t)}\left[\frac{a_t(1-\beta_1)^2}{1-\beta_2} + 2L\frac{1}{1-\beta_2}\right] \ ,$$
$$\tilde{C}_{2,t} = C_{2,t} = \frac{\mathsf{M}}{(1-\beta_1)\left((1-a_t\beta_1)+(\beta_1+a_t)\right)}\tilde{\mathsf{M}}^2\mathbb{E}[\|\hat{v}_0^{-1/2}\|] \ .$$

$\square$