

Supplementary Materials

Proof of Theorem 1

The proof presented below is along the lines of the Theorem 4 in [Reddi et al. 2018]. We further consider the terms modified by remote gradient observation, and provide a proof of convergence of NAMSG in the convex settings.

Proof.

In this proof, we use y_i to denote the i^{th} coordinate of a vector y .

From Algorithm 1,

$$\begin{aligned} x_{t+1} &= \prod_{\mathcal{F}, \sqrt{\hat{V}_t}} \left(x_t - \alpha_t \hat{V}_t^{-1/2} \left((1 - \mu_t) m_t + \mu_t g_t \right) \right) \\ &= \min_{x \in \mathcal{F}} \left\| \hat{V}_t^{1/4} \left(x - \left(x_t - \alpha_t \hat{V}_t^{-1/2} \left((1 - \mu_t) m_t + \mu_t g_t \right) \right) \right) \right\| \end{aligned} \quad (\text{A1})$$

Furthermore, $\prod_{\mathcal{F}, \sqrt{\hat{V}_t}}(x^*) = x^*$ for all $x^* \in \mathcal{F}$. Using Lemma A1 with $\hat{u}_1 = x_{t+1}$ and $\hat{u}_2 = x^*$, we

have

$$\begin{aligned} \left\| \hat{V}_t^{1/4} (x_{t+1} - x^*) \right\|^2 &\leq \left\| \hat{V}_t^{1/4} \left(x_t - \alpha_t \hat{V}_t^{-1/2} \left((1 - \mu_t) m_t + \mu_t g_t \right) - x^* \right) \right\|^2 \\ &= \left\| \hat{V}_t^{1/4} (x_t - x^*) \right\|^2 + \alpha_t^2 \left\| \hat{V}_t^{-1/4} \left((1 - \mu_t) m_t + \mu_t g_t \right) \right\|^2 - 2\alpha_t \left\langle (1 - \mu_t) m_t + \mu_t g_t, x_t - x^* \right\rangle \\ &= \left\| \hat{V}_t^{1/4} (x_t - x^*) \right\|^2 + \alpha_t^2 \left\| \hat{V}_t^{-1/4} \left((1 - \mu_t) m_t + \mu_t g_t \right) \right\|^2 - 2\alpha_t \left\langle (1 - \mu_t) \beta_{lt} m_{t-1} + (\mu_t + (1 - \mu_t)(1 - \beta_{lt})) g_t, x_t - x^* \right\rangle \\ &\leq \left\| \hat{V}_t^{1/4} (x_t - x^*) \right\|^2 + 2\alpha_t^2 \left((1 - \mu_t)^2 \left\| \hat{V}_t^{-1/4} m_t \right\|^2 + \mu_t^2 \left\| \hat{V}_t^{-1/4} g_t \right\|^2 \right) - 2\alpha_t \left\langle (1 - \mu_t) \beta_{lt} m_{t-1} + (1 - \beta_{lt} + \beta_{lt} \mu_t) g_t, x_t - x^* \right\rangle, \end{aligned} \quad (\text{A2})$$

where the second inequality follows from Cauchy-Schwarz and Young's inequality.

Rearrange the above equity, we obtain

$$\begin{aligned}
& \langle g_t, x_t - x^* \rangle \\
& \leq \frac{1}{2\alpha_t(1-\beta_{1t}(1-\mu_t))} \left(\|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \|\hat{V}_t^{1/4}(x_{t+1} - x^*)\|^2 \right) + \frac{\alpha_t}{1-\beta_{1t}(1-\mu_t)} \left((1-\mu_t)^2 \|\hat{V}_t^{-1/4} m_t\|^2 + \mu_t^2 \|\hat{V}_t^{-1/4} g_t\|^2 \right) \\
& \quad - \frac{(1-\mu_t)\beta_{1t}}{1-\beta_{1t}(1-\mu_t)} \langle m_{t-1}, x_t - x^* \rangle \\
& \leq \frac{1}{2\alpha_t(1-\beta_{1t}(1-\mu_t))} \left(\|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \|\hat{V}_t^{1/4}(x_{t+1} - x^*)\|^2 \right) + \frac{\alpha_t}{1-\beta_{1t}(1-\mu_t)} \left((1-\mu_t)^2 \|\hat{V}_t^{-1/4} m_t\|^2 + \mu_t^2 \|\hat{V}_t^{-1/4} g_t\|^2 \right) \\
& \quad + \frac{|1-\mu_t|\beta_{1t}\alpha_t}{2(1-\beta_{1t}(1-\mu_t))} \|\hat{V}_t^{-1/4} m_{t-1}\|^2 + \frac{|1-\mu_t|\beta_{1t}}{2(1-\beta_{1t}(1-\mu_t))\alpha_t} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 \\
& \leq \frac{1}{2\alpha_t(1-\beta_{1t}(1-\mu_t))} \left(\|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \|\hat{V}_t^{1/4}(x_{t+1} - x^*)\|^2 \right) + \frac{\alpha_t}{1-\beta_{1t}^2} \left(\beta_{1t}^2 \|\hat{V}_t^{-1/4} m_t\|^2 + \|\hat{V}_t^{-1/4} g_t\|^2 \right) \\
& \quad + \frac{\beta_{1t}^2 \alpha_t}{2(1-\beta_{1t}^2)} \|\hat{V}_t^{-1/4} m_{t-1}\|^2 + \frac{\beta_{1t}^2}{2(1-\beta_{1t}^2)\alpha_t} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2,
\end{aligned} \tag{A3}$$

where the second inequality also follows from Cauchy-Schwarz and Young's inequality, the last equality is due to $1-\mu_t \leq \beta_{1t}$.

Because of the convexity of the objective function, the regret satisfies

$$\begin{aligned}
R_T &= \sum_{t=1}^T (f_t(x_t) - f_t(x^*)) \leq \sum_{t=1}^T \langle g_t, x_t - x^* \rangle \\
&\leq \sum_{t=1}^T \left(\frac{1}{2\alpha_t(1-\beta_{1t}(1-\mu_t))} \left(\|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \|\hat{V}_t^{1/4}(x_{t+1} - x^*)\|^2 \right) + \frac{\alpha_t \beta_{1t}^2}{1-\beta_{1t}^2} \|\hat{V}_t^{-1/4} m_t\|^2 \right. \\
&\quad \left. + \frac{\alpha_t}{1-\beta_{1t}^2} \|\hat{V}_t^{-1/4} g_t\|^2 + \frac{\beta_{1t}^2 \alpha_t}{2(1-\beta_{1t}^2)} \|\hat{V}_t^{-1/4} m_{t-1}\|^2 + \frac{\beta_{1t}^2}{2\alpha_t(1-\beta_{1t}^2)} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 \right).
\end{aligned} \tag{A4}$$

The first inequity follows from the convexity of function f_t . The second inequality is due to (A3).

We now bound the term $\sum_{t=1}^T \alpha_t \|\hat{V}_t^{-1/4} g_t\|^2$. We have

$$\begin{aligned}
& \sum_{t=1}^T \alpha_t \|\hat{V}_t^{-1/4} g_t\|^2 \\
&= \sum_{t=1}^{T-1} \alpha_t \|\hat{V}_t^{-1/4} g_t\|^2 + \alpha_T \sum_{i=1}^d \frac{g_{T,i}^2}{\sqrt{\hat{V}_{T,i}}} \\
&\leq \sum_{t=1}^{T-1} \alpha_t \|\hat{V}_t^{-1/4} g_t\|^2 + \alpha_T \sum_{i=1}^d \frac{g_{T,i}^2}{\sqrt{V_{T,i}}} \\
&\leq \sum_{t=1}^{T-1} \alpha_t \|\hat{V}_t^{-1/4} g_t\|^2 + \frac{\alpha}{\sqrt{T}} \sum_{i=1}^d \frac{g_{T,i}^2}{\sqrt{(1-\beta_2) \sum_{j=1}^T \beta_2^{T-j} g_{j,i}^2}} \\
&\leq \sum_{t=1}^{T-1} \alpha_t \|\hat{V}_t^{-1/4} g_t\|^2 + \frac{\alpha}{\sqrt{T(1-\beta_2)}} \sum_{i=1}^d |g_{T,i}| \\
&\leq \frac{\alpha}{\sqrt{1-\beta_2}} \sum_{t=1}^T \left(\frac{1}{\sqrt{t}} \sum_{i=1}^d |g_{t,i}| \right) \\
&\leq \frac{\alpha}{\sqrt{1-\beta_2}} \sum_{i=1}^d \|g_{1:T,i}\|_2 \sqrt{\sum_{t=1}^T \frac{1}{t}} \\
&\leq \frac{\alpha \sqrt{1+\log(T)}}{\sqrt{1-\beta_2}} \sum_{i=1}^d \|g_{1:T,i}\|_2.
\end{aligned} \tag{A5}$$

In (A5), the third inequity is follows from the definition of v_t , the fifth inequality is due to Cauchy-Schwarz inequality. The final inequality is due to the following bound on harmonic sum:

$$\sum_{t=1}^T 1/t \leq 1 + \log(T).$$

By definition, we have $1 - \beta_{t+1} = (1 - \beta_t)(1 + \eta_t)$. From (A4), (A5) and Lemma A2, which

bounded $\sum_{t=1}^T \alpha_t \|\hat{V}_t^{-1/4} m_t\|^2$, we further bound the regret as

$$\begin{aligned} R_T &\leq \sum_{t=1}^T \left(\frac{1}{2\alpha_t(1-\beta_t)(1+\eta_t)} \left(\|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \|\hat{V}_t^{1/4}(x_{t+1} - x^*)\|^2 \right) + \frac{\alpha_t \beta_{t+1}^2}{1-\beta_{t+1}^2} \|\hat{V}_t^{-1/4} m_t\|^2 \right. \\ &\quad \left. + \frac{\alpha_t \beta_{t+1}^2}{2(1-\beta_{t+1}^2)} \|\hat{V}_t^{-1/4} m_{t-1}\|^2 + \frac{\beta_{t+1}^2}{2\alpha_t(1-\beta_{t+1}^2)} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 \right) + \frac{\alpha \sqrt{1+\log(T)}}{(1-\beta_1^2)\sqrt{1-\beta_2}} \sum_{i=1}^d \|g_{1:T,i}\|_2 \\ &\leq \sum_{t=1}^T \left(\frac{1}{2\alpha_t(1-\beta_t)(1+\eta_t)} \left(\|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \|\hat{V}_t^{1/4}(x_{t+1} - x^*)\|^2 \right) + \frac{\beta_{t+1}^2}{2\alpha_t(1-\beta_{t+1}^2)} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 \right) \\ &\quad + \sum_{t=1}^T \frac{\alpha_t \beta_{t+1}^2}{1-\beta_{t+1}^2} \|\hat{V}_t^{-1/4} m_t\|^2 + \sum_{t=1}^{T-1} \frac{\alpha_t \beta_{t+1}^2}{2(1-\beta_{t+1}^2)} \|\hat{V}_t^{-1/4} m_t\|^2 + \frac{\alpha \sqrt{1+\log(T)}}{(1-\beta_1^2)\sqrt{1-\beta_2}} \sum_{i=1}^d \|g_{1:T,i}\|_2 \\ &\leq \sum_{t=1}^T \left(\frac{1}{2\alpha_t(1-\beta_t)(1+\eta_t)} \left(\|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \|\hat{V}_t^{1/4}(x_{t+1} - x^*)\|^2 \right) + \frac{\beta_{t+1}^2}{2\alpha_t(1-\beta_{t+1}^2)} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 \right) \\ &\quad + \frac{3\beta_1^2}{2(1-\beta_1^2)} \sum_{t=1}^T \alpha_t \|\hat{V}_t^{-1/4} m_t\|^2 + \frac{\alpha \sqrt{1+\log(T)}}{(1-\beta_1^2)\sqrt{1-\beta_2}} \sum_{i=1}^d \|g_{1:T,i}\|_2 \\ &\leq \sum_{t=1}^T \left(\frac{1}{2\alpha_t(1-\beta_t)(1+\eta_t)} \left(\|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \|\hat{V}_t^{1/4}(x_{t+1} - x^*)\|^2 \right) + \frac{\beta_{t+1}^2}{2\alpha_t(1-\beta_{t+1}^2)} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 \right) \\ &\quad + \left(\frac{3\beta_1^2}{2(1-\beta_1)(1-\gamma)} + 1 \right) \frac{\alpha \sqrt{1+\log(T)}}{(1-\beta_1^2)\sqrt{1-\beta_2}} \sum_{i=1}^d \|g_{1:T,i}\|_2. \end{aligned} \tag{A6}$$

The second inequity is due to $\beta_{t+1} \geq \beta_{t+1}$ and $\hat{v}_{t,i}^{1/2} / \alpha_t \geq \hat{v}_{t-1,i}^{1/2} / \alpha_{t-1}$ by definition.

We also have

$$\begin{aligned}
& \sum_{t=1}^T \left(\frac{1}{2\alpha_t(1-\beta_{l_t})(1+\eta_t)} \left(\|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \|\hat{V}_{t+1}^{1/4}(x_{t+1} - x^*)\|^2 \right) + \frac{\beta_{l_t}^2}{2\alpha_t(1-\beta_{l_t}^2)} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 \right) \\
& \leq \frac{1}{2\alpha_1(1-\beta_{l_1})(1+\eta_1)} \|\hat{V}_1^{1/4}(x_1 - x^*)\|^2 + \sum_{t=2}^T \left(\frac{1}{2(1-\beta_{l_t})(1+\eta_t)\alpha_t} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \frac{1}{2(1-\beta_{l_{t-1}})(1+\eta_{t-1})\alpha_{t-1}} \|\hat{V}_{t-1}^{1/4}(x_t - x^*)\|^2 \right) \\
& \quad + \sum_{t=1}^T \frac{\beta_{l_t}^2}{2\alpha_t(1-\beta_{l_t}^2)} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 \\
& \leq \frac{1}{2(1-\beta_1^2)\alpha_1} \|\hat{V}_1^{1/4}(x_1 - x^*)\|^2 + \sum_{t=2}^T \frac{1}{2(1-\beta_{l_t})(1+\eta_t)} \left(\frac{1}{\alpha_t} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \frac{1}{\alpha_{t-1}} \|\hat{V}_{t-1}^{1/4}(x_t - x^*)\|^2 \right) \\
& \quad + \sum_{t=1}^T \frac{\beta_{l_t}^2}{2\alpha_t(1-\beta_{l_t}^2)} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 \\
& \leq \frac{1}{2(1-\beta_1^2)} \left(\frac{1}{\alpha_1} \|\hat{V}_1^{1/4}(x_1 - x^*)\|^2 + \sum_{t=2}^T \left(\frac{1}{\alpha_t} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \frac{1}{\alpha_{t-1}} \|\hat{V}_{t-1}^{1/4}(x_t - x^*)\|^2 \right) + \sum_{t=1}^T \frac{\beta_{l_t}^2}{\alpha_t} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 \right) \\
& = \frac{1}{2(1-\beta_1^2)} \left(\frac{1}{\alpha_1} \sum_{i=1}^d \hat{v}_{1,i}^{1/2} (x_{1,i} - x_i^*)^2 + \sum_{t=2}^T \left(\sum_{i=1}^d (x_{t,i} - x_i^*)^2 \left(\frac{\hat{v}_{t,i}^{1/2}}{\alpha_t} - \frac{\hat{v}_{t-1,i}^{1/2}}{\alpha_{t-1}} \right) \right) + \sum_{t=1}^T \sum_{i=1}^d \frac{\beta_{l_t}^2 (x_{t,i} - x_i^*)^2 \hat{v}_{t,i}^{1/2}}{\alpha_t} \right) \\
& \leq \frac{D_\infty^2}{2(1-\beta_1^2)} \left(\frac{1}{\alpha_1} \sum_{i=1}^d \hat{v}_{1,i}^{1/2} + \sum_{t=2}^T \left(\sum_{i=1}^d \left(\frac{\hat{v}_{t,i}^{1/2}}{\alpha_t} - \frac{\hat{v}_{t-1,i}^{1/2}}{\alpha_{t-1}} \right) \right) + \sum_{t=1}^T \sum_{i=1}^d \frac{\beta_{l_t}^2 \hat{v}_{t,i}^{1/2}}{\alpha_t} \right) \\
& = \frac{D_\infty^2}{2(1-\beta_1^2)\alpha_T} \sum_{i=1}^d \hat{v}_{T,i}^{1/2} + \frac{D_\infty^2}{2(1-\beta_1^2)} \sum_{t=1}^T \sum_{i=1}^d \frac{\beta_{l_t}^2 \hat{v}_{t,i}^{1/2}}{\alpha_t}.
\end{aligned} \tag{A7}$$

In (A7), the second inequality follows from the assumption $(1-\beta_{l_t})(1+\eta_t) \geq (1-\beta_{l_{t-1}})(1+\eta_{t-1})$ and

$\eta_t \geq \beta_{l_t}$, the third and the last inequality is due to $\hat{v}_{t,i}^{1/2} / \alpha_t \geq \hat{v}_{t-1,i}^{1/2} / \alpha_{t-1}$ by definition.

Combining (A6) and (A7), we obtain

$$R_T \leq \frac{D_\infty^2 \sqrt{T}}{2\alpha(1-\beta_1^2)} \sum_{i=1}^d \hat{v}_{T,i}^{1/2} + \frac{D_\infty^2}{2(1-\beta_1^2)} \sum_{t=1}^T \sum_{i=1}^d \frac{\beta_{l_t}^2 \hat{v}_{t,i}^{1/2}}{\alpha_t} + \left(\frac{3\beta_1^2}{2(1-\beta_1)(1-\gamma)} + 1 \right) \frac{\alpha \sqrt{1+\log(T)}}{(1-\beta_1^2)\sqrt{1-\beta_2}} \sum_{i=1}^d \|g_{\mathbf{i};T,i}\|_2. \tag{A8}$$

The proof is complete.

The Lemmas used in the proof are as follows:

Lemma A1. [McMahan & Streeter, 2010]

For any $Q \in \mathcal{S}_+^d$ and convex feasible set $\mathcal{F} \in \mathbb{R}^d$, suppose $\hat{u}_1 = \min_{x \in \mathcal{F}} \|Q^{1/2}(x - z_1)\|$ and

$\hat{u}_2 = \min_{x \in \mathcal{F}} \|Q^{1/2}(x - z_2)\|$ then we have $\|Q^{1/2}(\hat{u}_1 - \hat{u}_2)\| \leq \|Q^{1/2}(z_1 - z_2)\|$.

Lemma A2. [Reddi et al. 2018]

For the parameter settings and conditions assumed in Theorem 1, which is the same as Theorem 4 in [Reddi et al. 2018], we have

$$\sum_{t=1}^T \alpha_t \left\| \hat{V}_t^{-1/4} m_t \right\|^2 \leq \frac{\alpha \sqrt{1 + \log T}}{(1 - \beta_1)(1 - \gamma) \sqrt{1 - \beta_2}} \sum_{i=1}^d \|g_{1:T,i}\|_2.$$

The proofs of Lemma A1 and A2 are described in Reddi et al. [2018].

References:

- [McMahan & Streeter, 2010] H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In Proceedings of the 23rd Annual Conference On Learning Theory, pp. 244-256, 2010.
- [Reddi et al., 2018] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In International Conference on Learning Representations, 2018.