Supplementary Material

1 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 observations, and provider 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,

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

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

$$\begin{aligned} &\left\|\hat{V}_t^{1/4}\left(x_{t+1}-x^*\right)\right\|^2 \le \left\|\hat{V}_t^{1/4}\left(x_t - \alpha_t \hat{V}_t^{-1/2}\left(\left(1-\mu_t\right)m_t + \mu_t g_t\right) - x^*\right)\right\|^2 \\ =& \left\|\hat{V}_t^{1/4}\left(x_t - x^*\right)\right\|^2 + \alpha_t^2\left\|\hat{V}_t^{-1/4}\left(\left(1-\mu_t\right)m_t + \mu_t g_t\right)\right\|^2 - 2\alpha_t\left\langle\left(1-\mu_t\right)m_t + \mu_t g_t, x_t - x^*\right\rangle \\ =& \left\|\hat{V}_t^{1/4}\left(x_t - x^*\right)\right\|^2 + \alpha_t^2\left\|\hat{V}_t^{-1/4}\left(\left(1-\mu_t\right)m_t + \mu_t g_t\right)\right\|^2 \\ &- 2\alpha_t\left\langle\left(1-\mu_t\right)\beta_1 m_{t-1} + \left(\mu_t + \left(1-\mu_t\right)\left(1-\beta_{1t}\right)\right)g_t, x_t - x^*\right\rangle \\ \le& \left\|\hat{V}_t^{1/4}\left(x_t - x^*\right)\right\|^2 + 2\alpha_t^2\left(\left(1-\mu_t\right)^2\left\|\hat{V}_t^{-1/4}m_t\right\|^2 + \mu_t^2\left\|\hat{V}_t^{-1/4}g_t\right\|^2\right) \\ &- 2\alpha_t\left\langle\left(1-\mu_t\right)\beta_{1t} m_{t-1} + \left(1-\beta_{1t}+\beta_{1t}\mu_t\right)g_t, x_t - x^*\right\rangle, \end{aligned} \tag{A2}$$

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

Since $0 \le \beta_{1t} < 1$, $\mu_t = \eta_t\left(1-\beta_{1t}\right)/\beta_{1t}$, and $\beta_{1t} \le \eta_t \le \beta_{1t}/\left(1-\beta_{1t}\right)$, we obtain

$$0 < 1 - \beta_{1t} \le \mu_t \le 1. \tag{A3}$$

Rearrange the inequity (A2), we obtain

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

where the second inequality also follows from Cauchy-Schwarz and Young's inequality, the last equity is due to (A3).

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

$$
\begin{aligned}
& R_T = \sum_{i=1}^{T} \left(f_t\left(x_t\right) - f_t\left(x^*\right)\right) \leq \sum_{t=1}^{T} \langle g_t, x_t - x^* \rangle \\
\leq & \sum_{t=1}^{T} \left(\frac{1}{2\alpha_t \left(1 - \beta_{1t} \left(1 - \mu_t\right)\right)} \left(\left\| \hat{V}_t^{1/4} \left(x_t - x^*\right) \right\|^2 - \left\| \hat{V}_t^{1/4} \left(x_{t+1} - x^*\right) \right\|^2 \right) \right. \qquad (A5) \\
& + \frac{\alpha_t \beta_{1t}^2}{1 - \beta_{1t}^2} \left\| \hat{V}_t^{-1/4} m_t \right\|^2 + \frac{\alpha_t}{1 - \beta_{1t}^2} \left\| \hat{V}_t^{-1/4} g_t \right\|^2 + \frac{\beta_{1t}^2 \alpha_t}{2\left(1 - \beta_{1t}^2\right)} \left\| \hat{V}_t^{-1/4} m_{t-1} \right\|^2 \\
& \left. + \frac{\beta_{1t}^2}{2\alpha_t \left(1 - \beta_{1t}^2\right)} \left\| \hat{V}_t^{1/4} \left(x_t - x^*\right) \right\|^2 \right).
\end{aligned}
$$

The first inequity follows from the convexity of function f_t. The second inequality is due to (A4).

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

$$
\begin{aligned}
& \sum_{t=1}^{T} \alpha_t \left\| \hat{V}_t^{-1/4} g_t \right\|^2 = \sum_{t=1}^{T-1} \alpha_t \left\| \hat{V}_t^{-1/4} g_t \right\|^2 + \alpha_T \sum_{i=1}^{d} \frac{g_{T,i}^2}{\sqrt{\hat{v}_{T,i}}} \\
\leq & \sum_{t=1}^{T-1} \alpha_t \left\| \hat{V}_t^{-1/4} g_t \right\|^2 + \alpha_T \sum_{i=1}^{d} \frac{g_{T,i}^2}{\sqrt{v_{T,i}}} \leq \sum_{t=1}^{T-1} \alpha_t \left\| \hat{V}_t^{-1/4} g_t \right\|^2 + \frac{\alpha}{\sqrt{T}} \sum_{i=1}^{d} \frac{g_{T,i}^2}{\sqrt{\left(1 - \beta_2\right) \sum_{j=1}^{T} \beta_2^{T-j} g_{j,i}^2}} \\
\leq & \sum_{t=1}^{T-1} \alpha_t \left\| \hat{V}_t^{-1/4} g_t \right\|^2 + \frac{\alpha}{\sqrt{T\left(1 - \beta_2\right)}} \sum_{i=1}^{d} \left|g_{T,i}\right| \leq \frac{\alpha}{\sqrt{1 - \beta_2}} \sum_{t=1}^{T} \left(\frac{1}{\sqrt{t}} \sum_{i=1}^{d} \left|g_{t,i}\right| \right) \\
\leq & \frac{\alpha}{\sqrt{1 - \beta_2}} \sum_{i=1}^{d} \left\| g_{1:T,i} \right\|_2 \sqrt{\sum_{t=1}^{T} \frac{1}{t}} \leq \frac{\alpha \sqrt{1 + \log(T)}}{\sqrt{1 - \beta_2}} \sum_{i=1}^{d} \left\| g_{1:T,i} \right\|_2 .
\end{aligned}
\qquad (A6)
$$

In (A6), 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 $1 - \beta_{1t}\left(1 - \mu_t\right) = \left(1 - \beta_{1t}\right)\left(1 + \eta_t\right)$. From (A5), (A6), and Lemma A2, which bound $\sum_{t=1}^{T} \alpha_t \left\|\hat{V}_t^{-1/4} m_t\right\|^2$, we further bound the regret as

$$
\begin{aligned}
R_T \leq & \sum_{t=1}^{T} \left(\frac{1}{2\alpha_t\left(1-\beta_{1t}\right)\left(1+\eta_t\right)} \left(\left\|\hat{V}_t^{1/4}\left(x_t - x^*\right)\right\|^2 - \left\|\hat{V}_t^{1/4}\left(x_{t+1} - x^*\right)\right\|^2 \right) \right. \\
& + \frac{\alpha_t \beta_{1t}^2}{1-\beta_{1t}^2} \left\|\hat{V}_t^{-1/4} m_t\right\|^2 + \frac{\alpha_t \beta_{1t}^2}{2\left(1-\beta_{1t}^2\right)} \left\|\hat{V}_t^{-1/4} m_{t-1}\right\|^2 \\
& \left. + \frac{\beta_{1t}^2}{2\alpha_t\left(1-\beta_{1t}^2\right)} \left\|\hat{V}_t^{1/4}\left(x_t - x^*\right)\right\|^2 \right) + \frac{\alpha\sqrt{1+\log(T)}}{\left(1-\beta_1^2\right)\sqrt{1-\beta_2}} \sum_{i=1}^{d} \left\|g_{1:T,i}\right\|_2 \\
\leq & \sum_{t=1}^{T} \left(\frac{1}{2\alpha_t\left(1-\beta_{1t}\right)\left(1+\eta_t\right)} \left(\left\|\hat{V}_t^{1/4}\left(x_t - x^*\right)\right\|^2 - \left\|\hat{V}_t^{1/4}\left(x_{t+1} - x^*\right)\right\|^2 \right) \right. \\
& \left. + \frac{\beta_{1t}^2}{2\alpha_t\left(1-\beta_{1t}^2\right)} \left\|\hat{V}_t^{1/4}\left(x_t - x^*\right)\right\|^2 \right) + \sum_{t=1}^{T} \frac{\alpha_t \beta_{1t}^2}{1-\beta_{1t}^2} \left\|\hat{V}_t^{-1/4} m_t\right\|^2 \\
& + \sum_{t=1}^{T-1} \frac{\alpha_t \beta_{1t}^2}{2\left(1-\beta_{1t}^2\right)} \left\|\hat{V}_t^{-1/4} m_t\right\|^2 + \frac{\alpha\sqrt{1+\log(T)}}{\left(1-\beta_1^2\right)\sqrt{1-\beta_2}} \sum_{i=1}^{d} \left\|g_{1:T,i}\right\|_2 \\
\leq & \sum_{t=1}^{T} \left(\frac{1}{2\alpha_t\left(1-\beta_{1t}\right)\left(1+\eta_t\right)} \left(\left\|\hat{V}_t^{1/4}\left(x_t - x^*\right)\right\|^2 - \left\|\hat{V}_t^{1/4}\left(x_{t+1} - x^*\right)\right\|^2 \right) \right. \\
& \left. + \frac{\beta_{1t}^2}{2\alpha_t\left(1-\beta_{1t}^2\right)} \left\|\hat{V}_t^{1/4}\left(x_t - x^*\right)\right\|^2 \right) + \frac{3\beta_1^2}{2\left(1-\beta_1^2\right)} \sum_{t=1}^{T} \alpha_t \left\|\hat{V}_t^{-1/4} m_t\right\|^2 \\
& + \frac{\alpha\sqrt{1+\log(T)}}{\left(1-\beta_1^2\right)\sqrt{1-\beta_2}} \sum_{i=1}^{d} \left\|g_{1:T,i}\right\|_2 \\
\leq & \sum_{t=1}^{T} \left(\frac{1}{2\alpha_t\left(1-\beta_{1t}\right)\left(1+\eta_t\right)} \left(\left\|\hat{V}_t^{1/4}\left(x_t - x^*\right)\right\|^2 - \left\|\hat{V}_t^{1/4}\left(x_{t+1} - x^*\right)\right\|^2 \right) \right. \\
& \left. + \frac{\beta_{1t}^2}{2\alpha_t\left(1-\beta_{1t}^2\right)} \left\|\hat{V}_t^{1/4}\left(x_t - x^*\right)\right\|^2 \right) \\
& + \left(\frac{3\beta_1^2}{2\left(1-\beta_1\right)\left(1-\gamma\right)} + 1 \right) \frac{\alpha\sqrt{1+\log(T)}}{\left(1-\beta_1^2\right)\sqrt{1-\beta_2}} \sum_{i=1}^{d} \left\|g_{1:T,l}\right\|_2 .
\end{aligned}
\tag{A7}
$$

The second inequity is due to $\beta_{1t} \geq \beta_{1t+1}$ 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\left(1-\beta_{1t}\right)\left(1+\eta_t\right)}\left(\left\|\hat{V}_t^{1/4}\left(x_t-x^*\right)\right\|^2-\left\|\hat{V}_t^{1/4}\left(x_{t+1}-x^*\right)\right\|^2\right)\right.\\
&\left.+\frac{\beta_{1t}^2}{2\alpha_t\left(1-\beta_{1t}^2\right)}\left\|\hat{V}_t^{1/4}\left(x_t-x^*\right)\right\|^2\right)\\
\leq&\frac{1}{2\alpha_1\left(1-\beta_1\right)\left(1+\eta_1\right)}\left\|\hat{V}_1^{1/4}\left(x_1-x^*\right)\right\|^2+\sum_{t=2}^{T}\left(\frac{1}{2\left(1-\beta_{1t}\right)\left(1+\eta_t\right)\alpha_t}\left\|\hat{V}_t^{1/4}\left(x_t-x^*\right)\right\|^2\right.\\
&\left.-\frac{1}{2\left(1-\beta_{1t-1}\right)\left(1+\eta_{t-1}\right)\alpha_{t-1}}\left\|\hat{V}_{t-1}^{1/4}\left(x_t-x^*\right)\right\|^2\right)+\sum_{t=1}^{T}\frac{\beta_{1t}^2}{2\alpha_t\left(1-\beta_{1t}^2\right)}\left\|\hat{V}_t^{1/4}\left(x_t-x^*\right)\right\|^2\\
\leq&\frac{1}{2\left(1-\beta_1^2\right)\alpha_1}\left\|\hat{V}_1^{1/4}\left(x_1-x^*\right)\right\|^2+\sum_{t=1}^{T}\frac{\beta_{1t}^2}{2\alpha_t\left(1-\beta_{1t}^2\right)}\left\|\hat{V}_t^{1/4}\left(x_t-x^*\right)\right\|^2\\
&+\sum_{t=2}^{T}\frac{1}{2\left(1-\beta_{1t}\right)\left(1+\eta_t\right)}\left(\frac{1}{\alpha_t}\left\|\hat{V}_t^{1/4}\left(x_t-x^*\right)\right\|^2-\frac{1}{\alpha_{t-1}}\left\|\hat{V}_{t-1}^{1/4}\left(x_t-x^*\right)\right\|^2\right)\\
\leq&\frac{1}{2\left(1-\beta_1^2\right)}\left(\frac{1}{\alpha_1}\left\|\hat{V}_1^{1/4}\left(x_1-x^*\right)\right\|^2+\sum_{t=1}^{T}\frac{\beta_{1t}^2}{\alpha_t}\left\|\hat{V}_t^{1/4}\left(x_t-x^*\right)\right\|^2\right.\\
&\left.+\sum_{t=2}^{T}\left(\frac{1}{\alpha_t}\left\|\hat{V}_t^{1/4}\left(x_t-x^*\right)\right\|^2-\frac{1}{\alpha_{t-1}}\left\|\hat{V}_{t-1}^{1/4}\left(x_t-x^*\right)\right\|^2\right)\right)\\
=&\frac{1}{2\left(1-\beta_1^2\right)}\left(\frac{1}{\alpha_1}\sum_{i=1}^{d}\hat{v}_{1,i}^{1/2}\left(x_{1,i}-x_i^*\right)^2+\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1t}^2\left(x_{t,i}-x_i^*\right)^2\hat{v}_{t,i}^{1/2}}{\alpha_t}\right.\\
&\left.+\sum_{t=2}^{T}\left(\sum_{i=1}^{d}\left(x_{t,i}-x_i^*\right)^2\left(\frac{\hat{v}_{t,i}^{1/2}}{\alpha_t}-\frac{\hat{v}_{t-1,i}^{1/2}}{\alpha_{t-1}}\right)\right)\right)\\
\leq&\frac{D_\infty^2}{2\left(1-\beta_1^2\right)}\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_{1t}^2\hat{v}_{t,i}^{1/2}}{\alpha_t}\right)\\
=&\frac{D_\infty^2}{2\left(1-\beta_1^2\right)\alpha_T}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2}+\frac{D_\infty^2}{2\left(1-\beta_1^2\right)}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1i}^2\hat{v}_{t,i}^{1/2}}{\alpha_t}.
\end{aligned}
\tag{A8}
$$

In (A8), the second inequity follows from the assumption $\eta_t \geq \beta_{lt}$ and $\left(1-\beta_{1t}\right)\left(1+\eta_t\right) \geq \left(1-\beta_{1t-1}\right)\left(1+\eta_{t-1}\right)$, 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 (A7), (A8), and the assumption $\alpha_t = \alpha/\sqrt{t}$, we obtain

$$
\begin{aligned}
R_T \leq&\frac{D_\infty^2\sqrt{T}}{2\alpha\left(1-\beta_1^2\right)}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2}+\frac{D_\infty^2}{2\left(1-\beta_1^2\right)}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{11}^2\hat{v}_{t,i}^{1/2}}{\alpha_t}\\
&+\left(\frac{3\beta_1^2}{2\left(1-\beta_1\right)\left(1-\gamma\right)}+1\right)\frac{\alpha\sqrt{1+\log(T)}}{\left(1-\beta_1^2\right)\sqrt{1-\beta_2}}\sum_{i=1}^{d}\left\|g_{1:T,t}\right\|_2.
\end{aligned}
\tag{A9}
$$

The proof is complete.

The Lemmas used in the proof are as follows:

Lemma A1. [McMahan and Streeter, 2010]
For any $Q \in \mathcal{S}_+^d$ and convex feasible set $\mathcal{F} \in R^d$, suppose $\hat{u}_1 = \min_{x\in\mathcal{F}}\left\|Q^{1/2}\left(x-z_1\right)\right\|$ and $\hat{u}_2 = \min_{x\in\mathcal{F}}\left\|Q^{1/2}\left(x-z_2\right)\right\|$ then we have $\left\|Q^{1/2}\left(\hat{u}_1-\hat{u}_2\right)\right\| \leq \left\|Q^{1/2}\left(z_1-z_2\right)\right\|$.

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

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

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

References

H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. *CoRR*, abs/1002.4908, 2010. URL `http://arxiv.org/abs/1002.4908`.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=ryQu7f-RZ`.