Supplementary Materials

1 Proof of Theorem 1

The proof presented below is along the lines of Theorem 4 in [Reddi et al., 2018]. We further consider the terms modified by remote gradient observations, 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,

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

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

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

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

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

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

Rearrange the inequity (A2), we obtain

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

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

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

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

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

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 $1 - \beta_{1t}(1 - \mu_t) = (1 - \beta_{1t})(1 + \eta_t)$. 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 (1-\beta_{1t})(1+\eta_t)} \left(\left\| \hat{V}_t^{1/4} (x_t - x^*) \right\|^2 - \left\| \hat{V}_t^{1/4} (x_{t+1} - x^*) \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(1-\beta_{1t}^2)} \left\| \hat{V}_t^{-1/4} m_{t-1} \right\|^2 \\
& \left. + \frac{\beta_{1t}^2}{2\alpha_t (1-\beta_{1t}^2)} \left\| \hat{V}_t^{1/4} (x_t - x^*) \right\|^2 \right) + \frac{\alpha \sqrt{1+\log(T)}}{(1-\beta_1^2)\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 (1-\beta_{1t})(1+\eta_t)} \left(\left\| \hat{V}_t^{1/4} (x_t - x^*) \right\|^2 - \left\| \hat{V}_t^{1/4} (x_{t+1} - x^*) \right\|^2 \right) \right. \\
& \left. + \frac{\beta_{1t}^2}{2\alpha_t (1-\beta_{1t}^2)} \left\| \hat{V}_t^{1/4} (x_t - x^*) \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(1-\beta_{1t}^2)} \left\| \hat{V}_t^{-1/4} m_t \right\|^2 + \frac{\alpha \sqrt{1+\log(T)}}{(1-\beta_1^2)\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 (1-\beta_{1t})(1+\eta_t)} \left(\left\| \hat{V}_t^{1/4} (x_t - x^*) \right\|^2 - \left\| \hat{V}_t^{1/4} (x_{t+1} - x^*) \right\|^2 \right) \right. \\
& \left. + \frac{\beta_{1t}^2}{2\alpha_t (1-\beta_{1t}^2)} \left\| \hat{V}_t^{1/4} (x_t - x^*) \right\|^2 \right) + \frac{3\beta_1^2}{2(1-\beta_1^2)} \sum_{t=1}^{T} \alpha_t \left\| \hat{V}_t^{-1/4} m_t \right\|^2 \\
& + \frac{\alpha \sqrt{1+\log(T)}}{(1-\beta_1^2)\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 (1-\beta_{1t})(1+\eta_t)} \left(\left\| \hat{V}_t^{1/4} (x_t - x^*) \right\|^2 - \left\| \hat{V}_t^{1/4} (x_{t+1} - x^*) \right\|^2 \right) \right. \\
& \left. + \frac{\beta_{1t}^2}{2\alpha_t (1-\beta_{1t}^2)} \left\| \hat{V}_t^{1/4} (x_t - x^*) \right\|^2 \right) \\
& + \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} \left\| g_{1:T,i} \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

$$
\begin{aligned}
& \sum_{t=1}^{T} \left(\frac{1}{2\alpha_t (1-\beta_{1t})(1+\eta_t)} \left(\left\| \hat{V}_t^{1/4} (x_t - x^*) \right\|^2 - \left\| \hat{V}_t^{1/4} (x_{t+1} - x^*) \right\|^2 \right) \right. \\
& \left. + \frac{\beta_{1t}^2}{2\alpha_t (1-\beta_{1t}^2)} \left\| \hat{V}_t^{1/4} (x_t - x^*) \right\|^2 \right)
\end{aligned}
$$

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

In (A8), the second inequity follows from the assumption $\eta_t \geq \beta_{1t}$ and $(1-\beta_{1t})(1+\eta_t) \geq (1-\beta_{1t-1})(1+\eta_{t-1})$, 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

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

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

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].

2 More details on experiments

We use constant hyper-parameters in the experiments. For ADAM, NADAM, AMSGRAD, and NAMSG, the hyper-parameters $(\alpha, \beta_1, \beta_2)$ are selected from $\{0.0005, 0.001, 0.002, 0.005, 0.01\} \times \{0.9\} \times \{0.99, 0, 999\}$ by grid search. For SGD, the hyper-parameters (α, β) are selected from $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0\} \times \{0.9\}$ by grid search. In the experiments of logistic regression and CNN on MNIST, we run grid search for 10 and 6 epochs respectively, since the train losses are already quite low. For training Resnet-20 on CIFAR-10, we run grid search for 30 epochs since it is time consuming. Table A1 shows the hyper-parameters selected.

Table A1: The hyper-parameters in the experiments

	MNIST		CIFAR-10
Methods	Logistic regression	CNN	Resnet-20
SGD	$(1.0, 0.9)$	$(1.0, 0.9)$	$(0.2, 0.9)$
ADAM	$(0.005, 0.9, 0.99)$	$(0.001, 0.9, 0.999)$	$(0.001, 0.9, 0.99)$
NADAM	$(0.005, 0.9, 0.99)$	$(0.001, 0.9, 0.999)$	$(0.001, 0.9, 0.999)$
AMSGRAD	$(0.005, 0.9, 0.999)$	$(0.002, 0.9, 0.99)$	$(0.001, 0.9, 0.99)$
NAMSG	$(0.005, 0.9, 0.999)$	$(0.002, 0.9, 0.99)$	$(0.001, 0.9, 0.99)$

Table A2 shows the train loss at the 30^{th} epoch in grid search for training Resnet-20 on CIFAR-10, that are the average of 2 runs. It is observed that NAMSG is faster than AMSGRAD for almost all the hyper-parameter settings. It is also faster than ADAM and NADAM for most of the settings in the experiments.

Table A2: Train loss in grid search for Resnet-20 on CIFAR-10

Hyper-parameters	ADAM	NADAM	AMSGRAD	NAMSG
$(0.0005, 0.9, 0.99)$	0.3502	**0.3437**	0.3675	0.3576
$(0.001, 0.9, 0.99)$	0.3394	0.3220	0.3306	**0.3152**
$(0.002, 0.9, 0.99)$	0.4048	0.3836	0.3770	**0.3429**
$(0.005, 0.9, 0.99)$	0.5791	0.5438	0.5156	**0.4831**
$(0.01, 0.9, 0.99)$	0.7181	0.6735	0.6513	**0.6137**
$(0.0005, 0.9, 0.999)$	0.3538	0.3507	0.3363	**0.3170**
$(0.001, 0.9, 0.999)$	0.3415	0.3214	0.3310	**0.3162**
$(0.002, 0.9, 0.999)$	0.4125	0.3814	0.4031	**0.3763**
$(0.005, 0.9, 0.999)$	0.5757	**0.5293**	0.5889	0.5570
$(0.01, 0.9, 0.999)$	0.7455	0.6854	0.8575	**0.7102**

In order to promote the generalization performance, we introduce 2 strategies to switch from NAMSG to SGD. The first one is named as SWNTS, which switches from NAMSG to SGD by disabling the dividing of $\sqrt{\hat{v}_t}$ in Algorithm 1. At the switching point, it sets the learning rate of SGD as $\hat{\alpha}_t = \|\hat{u}_t\|^2 / < \hat{m}_t, \hat{u}_t >$, insuring $\mathrm{proj}_{\hat{\alpha}_t \hat{m}_t} \hat{u}_t = \hat{u}_t$, where $\hat{u}_t = \alpha_t \hat{m}_t / \sqrt{\hat{v}_t}$, and $\mathrm{proj}_a b$ denotes the orthogonal projection of a onto b. The second one is NAMSB, which shares the concept of AMSBOUND [Luo et al., 2019] by employing dynamic bounds on learning rates to achieve a gradual and smooth transformation to SGD. NAMSB is described in Algorithm A1. The bound functions are $\xi_l(t) = \alpha_{\mathrm{SGD}}(1 - 1/(\hat{\gamma}t + 1))$ and $\xi_u(t) = \alpha_{\mathrm{SGD}}(1 + 1/\hat{\gamma}t)$, where α_{SGD} is the step size of SGD which the method gradually switching to, and $\hat{\gamma}$ is the switching rate. $\mathrm{Clip}(x, a, b)$ is defined as a clipping operation constraining the output vector elementwisely in $[a, b]$, where x is a vector, and a, b are scalars.

We first compare the performance of NAMSB and AMSBOUND. The hyper-parameters are assigned as follows: The minibatch size is 256. The $(\alpha_t, \beta_1, \beta_{2t})$ are selected as $(0.001, 0.9, 0.99)$, which are optimal constant values for NAMSG and AMSGRAD. The SGD step size and the switching

Algorithm A1 NAMSB Algorithm

Input: initial parameter vector x_1, step size $\{\alpha_t\}_{t=1}^T$, lower bound function $\xi_l(t)$, upper bound function $\xi_u(t)$, coefficients $\{\beta_{1t}\}_{t=1}^T$, β_2, ϵ, iteration number T
Output: parameter vector x_T

Set $m_0 = 0$, $v_0 = 0$, and $\hat{v}_0 = \epsilon$.
for $t = 1$ to $T - 1$ **do**
  $g_t = \nabla f_t(x_t)$.
  $m_t = \beta_{1t} m_{t-1} + (1 - \beta_{1t}) g_t$.
  $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$.
  $\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$, $\hat{V}_t = \mathrm{diag}(v_t)$.
  $\xi = \mathrm{Clip}(\alpha_1/\sqrt{\hat{V}_t}, \xi_l(t), \xi_u(t))$ and $\xi_t = \alpha_t/\alpha_1 \xi$
  $\hat{m}_t = (1 - \mu_t) m_t + \mu_t g_t$, where $\mu_t = \eta_t(1 - \beta_{1t})/\beta_{1t}$.
  $x_{t+1} = \Pi_{\mathcal{F}, \mathrm{diag}(\xi_t^{-1})}(x_t - \alpha_t \hat{m}_t/\sqrt{\hat{v}_t})$.
end for

rate are set as $\alpha_{\mathrm{SGD}} = 0.1$ and $\hat{\gamma} = 0.001$ according to the default value of AMSBOUND. We also try a larger SGD step size $\alpha_{\mathrm{SGD}} = 0.5$ to achieve better generalization. The SGD step size α_{SGD} is divided by 10 at the 12000^{th} iteration (in the 62^{th} epoch). In NAMSB, the observation distance η increases from 0.9 to 3.0 linearly in the first epoch. The results are shown in Figure A1, that are the average of 5 runs. It is observed that the small SGD step size $\alpha_{\mathrm{SGD}} = 0.1$ is beneficial to fast convergence, but may lead to poor generalization. In some sense, switching to SGD with a low step size performs similar to a step size decreasing scheme that accelerates convergence while causing insufficient exploration of the parameter space. On the contrary, switching to SGD with a relatively large step size can achieve good generalization at the cost of more iterations. With the same hyper-parameters settings, NAMSB converges faster than AMSBOUND benefiting from remote gradients observations. NAMSB also achieves slightly better generalization compared with AMSBOUND in the experiments.

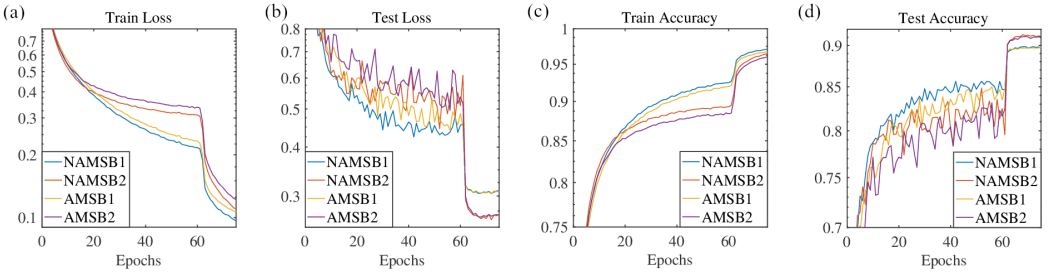

Figure A1: Comparison of NAMSB and AMSBOUND for Resnet-20 on CIFAR-10. NAMSB1 and NAMSB2 denote NAMSB with $\alpha_{\mathrm{SGD}} = 0.1$ and 0.5; AMSB1 and AMSB2 denote AMSB with $\alpha_{\mathrm{SGD}} = 0.1$ and 0.5, respectively.

In the experiments of the strategies for NAMSG to promote generalization on CIFAR-10, the hyper-parameters are assigned without grid search. The relatively large step size for NAMSG1 and SWNTS1 is $\alpha = 0.0015$, β_1 and β_2 are the same as NAMSG. SWNTS and SWNTS1 switch to SGD at the 3000^{th} iteration (in the 16^{th} epoch). NAMSB uses the same hyper-parameters as NAMSB2 in Figure A1, but runs for 5 more times.

We also present the test accuracy for Resnet-20 on CIFAR-10 with mean and standard deviation in Table A3. The results show that training by NAMSG with a relatively large step size or switching to SGD during the process can achieve good generalization.

The experiments are carried out on a workstation with an Intel Xeon E5-2680 v3 CPU and a NVIDIA K40 GPU. The source code of NAMSG can be downloaded at `https://github.com/rationalspark/NAMSG/blob/master/Namsg.py`. The simulation environment is MXNET, which can be downloaded at `http://mxnet.incubator.apache.org`. The MNIST dataset can be downloaded at `http://yann.lecun.com/exdb/mnist`; the CIFAR-10 dataset can be downloaded at `http://www.cs.toronto.edu/~kriz/cifar.html`.

Table A3: Test accuracy in the experiments for Resnet-20 on CIFAR-10

	Mean of the last 5 epoches		The best accuracy in each run	
Methods	Mean	Standard deviation	Mean	Standard deviation
SGD	0.9068	0.0024	0.9088	0.0022
ADAM	0.9052	0.0026	0.9073	0.0033
NADAM	0.9053	0.0020	0.9070	0.0022
AMSGRAD	0.9026	0.0027	0.9043	0.0026
NAMSG	0.9033	0.0020	0.9048	0.0020
SWNTS	0.9092	0.0024	0.9109	0.0025
NAMSG1	0.9082	0.0015	0.9102	0.0010
SWNTS1	**0.9121**	0.0016	**0.9145**	0.0013
AMSB1	0.8969	0.0017	0.8986	0.0078
AMSB2	0.9101	0.0025	0.9122	0.0024
NAMSB1	0.8976	0.0037	0.8997	0.0035
NAMSB2	0.9113	0.0019	0.9145	0.0014

References

Liangchen Luo, Yuanhao Xiong, and Yan Liu. Adaptive gradient methods with dynamic bound of learning rate. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg3g2R9FX`.

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`.