Divergence Results and Convergence of a Variance Reduced Version of ADAM: Supplementary Materials

1 Proofs

1.1 Technical Lemmas

Lemma 1 *In a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let there be an $\mathcal{F}$-measurable random variable $X(w)$ and an event $A \in \mathcal{F}$ such that $\mathbb{P}\{A\} > 0$. For a convex function $\phi(x)$, we have*

$$\frac{\mathbb{E}\left[\phi(X)\mathbf{1}\{A\}\right]}{\mathbb{P}\{A\}} \geq \phi\left(\frac{\mathbb{E}\left[X\mathbf{1}\{A\}\right]}{\mathbb{P}\{A\}}\right).$$

Proof: Let

$$x_0 = \frac{\mathbb{E}\left[X\mathbf{1}\{A\}\right]}{\mathbb{P}\{A\}} = \frac{1}{\mathbb{P}\{A\}}\int_A X(w)d\mathbb{P}(w).$$

Since ϕ is convex, there exists a sub-gradient of ϕ at x_0, i.e., there exists an a such that

$$\phi(x) \geq \phi(x_0) + a(x - x_0)$$

for any $x \in \mathbb{R}$. Then we have

$$\begin{aligned}
\frac{\mathbb{E}\left[\phi(X)\mathbf{1}\{A\}\right]}{\mathbb{P}\{A\}} &= \frac{1}{\mathbb{P}\{A\}}\int_A \phi(X(w))d\mathbb{P}(w) \\
&\geq \frac{1}{\mathbb{P}\{A\}}\int_A a(X(w) - x_0) + \phi(x_0)d\mathbb{P}(w) \\
&= \frac{a}{\mathbb{P}\{A\}}\int_A X(w)d\mathbb{P}(w) + \frac{\phi(x_0) - ax_0}{\mathbb{P}\{A\}}\int_A 1d\mathbb{P}(w) \\
&= ax_0 + \phi(x_0) - ax_0 \\
&= \phi(x_0),
\end{aligned}$$

which finishes the proof.

Lemma 2 *For any $x, y > 0$,*

$$(x + y)^3 \leq 4(x^3 + y^3).$$

Proof: For $t \geq 0$, let

$$h(t) := \frac{(1+t)^3}{1+t^3} = 1 + 3\frac{t + t^2}{1 + t^3}.$$

The derivative of $h(t)$ reads

$$h'(t) = 3\frac{(1+2t)(1+t^3) - 3t^2(t+t^2)}{(1+t^3)^2} = -3\frac{(t-1)(t+1)^3}{(t^3+1)^2}.$$

Apparently $h(t)$ achieves the maximum at $t = 1$, thus $h(x/y) \leq h(1) = 4$, and we have

$$\frac{(x+y)^3}{x^3 + y^3} \leq 4.$$

Lemma 3 *Given $\alpha_t = \alpha/t$ and $\beta_1 \in [0, 1)$, there exists a constant $\bar{C} > 0$, such that for any $t \geq 2$ we have*

$$\sum_{j=1}^{t-1} \alpha_j \beta_1^{t-j} \leq \bar{C}\alpha_t.$$

Proof: Letting $t^* = \lfloor \frac{t-1}{2} \rfloor$, we have

$$\begin{aligned}
\sum_{j=1}^{t-1} \alpha_j \beta_1^{t-j} &= \sum_{j=1}^{t^*} \alpha_j \beta_1^{t-j} + \sum_{j=t^*+1}^{t-1} \alpha_j \beta_1^{t-j} \\
&\leq \alpha \sum_{j=1}^{t^*} \beta_1^{t-j} + \frac{\alpha}{t^*+1} \sum_{j=t^*+1}^{t-1} \beta_1^{t-j} \\
&\leq \frac{\alpha \beta_1^{t-t^*}}{1-\beta_1} + \frac{\alpha}{t^*+1} \frac{\beta_1}{1-\beta_1} \\
&\leq \frac{\alpha \beta_1^{(t+1)/2}}{1-\beta_1} + \frac{2\alpha\beta_1}{(1-\beta_1)} \frac{1}{t-1} = \mathcal{O}(t^{-1}).
\end{aligned}$$

Thus there exists a positive constant $\bar{C}$ such that for any $t \geq 2$,

$$\sum_{j=1}^{t-1} \alpha_j \beta_1^{t-j} \leq \bar{C}\alpha_t.$$

Lemma 4 *Consider $0 < A < 1$ and $T \geq 2$, and let*

$$\begin{aligned}
\lambda_{T-1} &= \prod_{t=1}^{T-1} \left(1 - \frac{A}{t}\right), \\
\nu_{T-1} &= \sum_{t=1}^{T-1} \frac{1}{t^2} \prod_{j=t+1}^{T-1} \left(1 - \frac{A}{t}\right).
\end{aligned}$$

Then we have

$$\lambda_{T-1} \leq \mathcal{O}(T^{-A})$$

and

$$\nu_{T-1} \leq \mathcal{O}(T^{-A}).$$

Proof: Notice that

$$\log \lambda_{T-1} = \sum_{t=1}^{T-1} \log \left(1 - \frac{A}{t}\right) \leq -A \sum_{t=1}^{T-1} \frac{1}{t} \leq -A \log T,$$

where the first inequality comes from $\log(1-x) \leq -x$ for $x \geq 0$ and the second inequality uses the integral approximation

$$\sum_{t=1}^{T-1} \frac{1}{t} \geq \sum_{t=1}^{T-1} \int_t^{t+1} \frac{1}{s}\, ds = \int_1^T \frac{1}{s}\, ds = \log T.$$

Thus $\lambda_{T-1} \leq T^{-A}$. Similarly, we have

$$\log \frac{\lambda_{T-1}}{\lambda_t} = \sum_{k=t+1}^{T-1} \log \left(1 - \frac{A}{t}\right) \leq -A \log \frac{T}{t+1},$$

and then,

$$\nu_{T-1} \leq \sum_{t=1}^{T-1} \frac{1}{t^2} \left(\frac{T}{t+1} \right)^{-A} \leq T^{-A} \sum_{t=1}^{T-1} \frac{(t+1)^A}{t^2} \leq 2^A T^{-A} \sum_{t=1}^{T-1} t^{-2+A}.$$

Again, applying the integral approximation yields

$$\sum_{t=1}^{T-1} t^{-2+A} \leq 1 + \sum_{t=2}^{T-1} \int_{t-1}^{t} s^{-2+A} \, ds = 1 + \int_{1}^{T-1} s^{-2+A} \, ds \leq 1 + \frac{1}{1-A} = \frac{2-A}{1-A} < \infty.$$

Then we have $\nu_{T-1} \leq \mathcal{O}(T^{-A})$.

1.2 Proof of Theorem 1

Let $p = \mathbb{P}(\xi = 1)$. In each step, the update value is

$$\begin{aligned} \Delta_t &= -\frac{\alpha g_t}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2) g_t^2}} \\ &= \begin{cases} -\frac{\alpha (w_t/\delta + \delta^4)}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)(w_t/\delta + \delta^4)^2}} & \text{with probability } \ p \\ \frac{\alpha (1 - w_t/\delta)}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)(1 - w_t/\delta)^2}} & \text{with probability } \ 1-p. \end{cases} \end{aligned}$$

Apparently,

$$F(w) = \frac{w^2}{2\delta} + \delta w,$$

and

$$w^* = -\delta^2.$$

We use contradiction to prove the theorem. Assume that $\mathbb{E}[F(w_t) - F(w^*)] \to 0$. Notice that

$$F(w_t) - F(w^*) = \frac{1}{2\delta}(w_t - w^*)^2,$$

which means that $\mathbb{E}[F(w_t) - F(w^*)] \to 0$ is equivalent to $\mathbb{E}\left[(w_t - w^*)^2\right] \to 0$. Let us select $0 < \epsilon < 1/2$, and we choose T_ϵ such that $t > T_\epsilon$ implies $\mathbb{E}\left[(w_t - w^*)^2\right] < \epsilon$. The following discussion is based on w_t such that $t > T_\epsilon$.

We have

$$|\Delta_t| = \frac{\alpha |g_t|}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2) g_t^2}} \leq \frac{\alpha |g_t|}{\sqrt{(1-\beta_2) g_t^2}} = \frac{\alpha}{\sqrt{1-\beta_2}}, \tag{1}$$

where the inequality is due to v_k being non-negative for any k.

Let $\mathcal{F}_t$ be the filtration including all the information obtained until the update of w_t, including w_t. We define the following event

$$E := \left\{ |w_t - w^*| < \delta^2 \right\}$$

which is known given $\mathcal{F}_t$. We have

$$\mathbb{P}\{E^c\} = \mathbb{P}\left\{ |w_t - w^*| \geq \delta^2 \right\} \leq \frac{\mathbb{E}\left[(w_t - w^*)^2\right]}{\delta^4} < \frac{\epsilon}{\delta^4}.$$

Given E^c, we simply bound the step size with the lower bound

$$\mathbb{E}\left[\Delta_t \mathbf{1}\{E^c\}\right] \geq -\frac{\alpha}{\sqrt{1-\beta_2}} \mathbb{E}\left[\mathbf{1}\{E^c\}\right] \geq -\frac{\epsilon}{\delta^4} \frac{\alpha}{\sqrt{1-\beta_2}}.$$

For the samples in E, we have

$$\begin{aligned}
\mathbb{E}\left[\Delta_t \mathbf{1}\{E\}\right] &= \mathbb{E}\left[\mathbb{E}\left[\Delta_t \mathbf{1}\{E\}\,|\mathcal{F}_t\,]\right] = \mathbb{E}\left[\mathbb{E}\left[\Delta_t\,|\mathcal{F}_t\,\right]\mathbf{1}\{E\}\right] \\
&= \mathbb{E}\left[\mathbf{1}\{E\}\left\{(1-p)\frac{\alpha(1-w_t/\delta)}{\sqrt{\beta_2 v_{t-1}+(1-\beta_2)(1-w_t/\delta)^2}}\right\}\right] \\
&\quad -\mathbb{E}\left[\left\{p\frac{\alpha(w_t/\delta+\delta^4)}{\sqrt{\beta_2 v_{t-1}+(1-\beta_2)(w_t/\delta+\delta^4)^2}}\right\}\right] \\
&\geq \mathbb{E}\left[\mathbf{1}\{E\}(1-p)\frac{\alpha}{\sqrt{\beta_2 v_{t-1}+(1-\beta_2)(1+2\delta)^2}}\right] \\
&\quad -p\frac{\alpha}{\sqrt{1-\beta_2}}\mathbb{P}\{E\}. \qquad (2)
\end{aligned}$$

In the inequality, the first term is bounded because $-2\delta < w_t/\delta < 0$ by the definition of E and the second term is bounded by the bound of the step length in (1).

By applying Lemma 1 to (2), we have

$$\mathbb{E}[\Delta_t \mathbf{1}\{E\}] \geq (1-p)\mathbb{P}(E)\frac{\alpha}{\sqrt{\beta_2 \mathbb{E}[v_{t-1}\mathbf{1}\{E\}]/\mathbb{P}\{E\}+(1-\beta_2)(1+2\delta)^2}} - p\mathbb{P}(E)\frac{\alpha}{\sqrt{1-\beta_2}}.$$

We next focus on the conditional expectation

$$\mathbb{E}\left[v_{t-1}\mathbf{1}\{E\}\right] = (1-\beta_2)\sum_{k=1}^{t-1}\beta_2^{t-1-k}\mathbb{E}\left[\mathbf{1}\{E\}g_k^2\right].$$

We claim that for any trajectory in E and for any $k < t$, we have

$$|w_t - w_k| = \left|\sum_{j=k}^{t-1}\Delta_j\right| \leq \sum_{j=k}^{t-1}|\Delta_j| \leq \frac{\alpha(t-k)}{\sqrt{1-\beta_2}}.$$

The last inequality comes from the bound of the step length in (1). Then we have

$$w_t - \frac{\alpha(t-k)}{\sqrt{1-\beta_2}} \leq w_k \leq w_t + \frac{\alpha(t-k)}{\sqrt{1-\beta_2}}.$$

Let us recall that given E, we have $-2\delta^2 < w_k < 0$, and hence

$$-2\delta^2 - \frac{\alpha(t-k)}{\sqrt{1-\beta_2}} \leq w_k \leq \frac{\alpha(t-k)}{\sqrt{1-\beta_2}}. \qquad (3)$$

Then for each $k = 1, \ldots, t-1$, we obtain

$$\begin{aligned}
\mathbb{E}\left[g_k^2\mathbf{1}\{E\}\right] &= \mathbb{E}\left[\mathbb{E}[g_k^2\mathbf{1}\{E\}\,|\mathcal{F}_k\,]\right] \\
&\leq \mathbb{E}\left[\left(\mathbb{E}\left[g_k^{2(1+\mu)}|\mathcal{F}_k\right]\right)^{1/(1+\mu)}\left(\mathbb{E}\left[\mathbf{1}\{E\}\right]\right)^{\mu/(1+\mu)}\right] \\
&\leq \mathbb{E}\left[\left(\mathbb{E}\left[g_k^{2(1+\mu)}|\mathcal{F}_k\right]\right)^{1/(1+\mu)}\right]
\end{aligned}$$

where the inequality holds for any μ according to the Holder inequality. Let $0 < \mu < 1/2$. According to the bound given previously in (3) and $\delta \geq 2$, we have

$$\begin{aligned}
\left(\frac{w_k}{\delta}+\delta^4\right)^2 &\leq \left(\frac{\alpha(t-k)}{\delta\sqrt{1-\beta_2}}+2\delta+\delta^4\right)^2 \\
\left(1-\frac{w_k}{\delta}\right)^2 &\leq \left(\frac{\alpha(t-k)}{\delta\sqrt{1-\beta_2}}+2\delta+1\right)^2.
\end{aligned}$$

Then we derive,

$$\begin{aligned}
\mathbb{E}\left[g_k^{2(1+\mu)}|\mathcal{F}_k\right] &= p\left(\frac{w_k}{\delta}+\delta^4\right)^{2(1+\mu)} + (1-p)\left(\frac{w_k}{\delta}-1\right)^{2(1+\mu)} \\
&\leq \frac{1+\delta}{\delta^4}\left(\frac{\alpha(t-k)}{\delta\sqrt{1-\beta_2}}+2\delta+\delta^4\right)^{2(1+\mu)} + \left(\frac{\alpha(t-k)}{\delta\sqrt{1-\beta_2}}+2\delta+1\right)^{2(1+\mu)} \\
&= (1+\delta)\left(\frac{\alpha(t-k)}{\delta^5\sqrt{1-\beta_2}}+\frac{2}{\delta^3}+1\right)^{2(1+\mu)}\delta^{4+8\mu} + \left(\frac{\alpha(t-k)}{\delta\sqrt{1-\beta_2}}+2\delta+1\right)^{2(1+\mu)} \\
&\leq 2\delta\cdot\left(\frac{\alpha(t-k)}{\delta^5\sqrt{1-\beta_2}}+\frac{2}{\delta^3}+1\right)^{3}\cdot\delta^{4+8\mu} + \left(\frac{\alpha(t-k)}{\delta\sqrt{1-\beta_2}}+2\delta+1\right)^{3} \\
&\leq 2\left(\frac{\alpha(t-k)}{\sqrt{1-\beta_2}}+2\right)^{3}\delta^{5+8\mu} + \left(\frac{\alpha(t-k)}{\sqrt{1-\beta_2}}+3\delta\right)^{3} \\
&\leq \left(\frac{8\alpha^3(t-k)^3}{(1-\beta_2)^{3/2}}+64\right)\delta^{5+8\mu} + \frac{4\alpha^3(t-k)^3}{(1-\beta_2)^{3/2}}+108\delta^3.
\end{aligned}$$

We have used $\delta > 2$ and $0 < \mu < 1/2$. The last inequality holds because of Lemma 2. Then we obtain

$$\begin{aligned}
\left(\mathbb{E}\left[g_k^{2(1+\mu)}|\mathcal{F}_k\right]\right)^{1/(1+\mu)} &\leq \left[\left(\frac{8\alpha^3(t-k)^3}{(1-\beta_2)^{3/2}}+64\right)\delta^{5+8\mu} + \frac{4\alpha^3(t-k)^3}{(1-\beta_2)^{3/2}}+108\delta^3\right]^{1/(1+\mu)} \\
&= \left(\frac{8\alpha^3(t-k)^3}{(1-\beta_2)^{3/2}}+64+\frac{4\alpha^3(t-k)^3}{(1-\beta_2)^{3/2}}\delta^{-5-8\mu}+108\delta^{-2-8\mu}\right)^{1/(1+\mu)}\delta^{(5+8\mu)/(1+\mu)} \\
&\leq \left(\frac{12\alpha^3(t-k)^3}{(1-\beta_2)^{3/2}}+172\right)^{1/(1+\mu)}\delta^{(5+8\mu)/(1+\mu)} \\
&\leq \left(\frac{12\alpha^3(t-k)^3}{(1-\beta_2)^{3/2}}+172\right)\delta^{(5+8\mu)/(1+\mu)}.
\end{aligned}$$

The third inequality uses $\delta > 1$ and thus

$$\begin{aligned}
\mathbb{E}[v_{t-1}\mathbf{1}\{E\}] &\leq (1-\beta_2)\sum_{k=1}^{t-1}\beta_2^{t-1-k}\left(\frac{12\alpha^3(t-k)^3}{(1-\beta_2)^{3/2}}+172\right)\delta^{(5+8\mu)/(1+\mu)} \\
&\leq \delta^{(5+8\mu)/(1+\mu)}\left\{\frac{12\alpha^3}{\sqrt{1-\beta_2}}\sum_{k=1}^{\infty}\beta_2^{k-1}k^3+172\right\} \\
&\leq \delta^{(5+8\mu)/(1+\mu)}\left\{\frac{72\alpha^3}{(1-\beta_2)^{9/2}}+172\right\} := M_1\delta^{(5+8\mu)/(1+\mu)}
\end{aligned}$$

where the last inequality is because $\sum_{k=1}^{\infty}\beta_2^{k-1}k^3 = (1+4\beta_2+\beta_2^2)/(1-\beta_2)^4 < 6/(1-\beta_2)^4$. Thus, we have

$$\begin{aligned}
\mathbb{E}\left[\Delta_t\mathbf{1}\{E\}\right] &\geq \mathbb{P}\{E\}\left\{\left(1-\frac{1+\delta}{1+\delta^4}\right)\frac{\alpha}{\sqrt{\beta_2 M_1\delta^{(5+8\mu)/(1+\mu)}/\mathbb{P}\{E\}+(1-\beta_2)(1+2\delta)^2}}\right. \\
&\quad \left.-\frac{1+\delta}{1+\delta^4}\frac{\alpha}{\sqrt{1-\beta_2}}\right\} \\
&\geq \frac{1}{2}\left\{\left(1-\frac{1+\delta}{1+\delta^4}\right)\frac{\alpha}{\sqrt{2\beta_2 M_1\delta^{(5+8\mu)/(1+\mu)}+(1-\beta_2)(1+2\delta)^2}}\right. \\
&\quad \left.-\frac{1+\delta}{1+\delta^4}\frac{\alpha}{\sqrt{1-\beta_2}}\right\},
\end{aligned}$$

where the second inequality follows from

$$\mathbb{P}\{E\} > 1-\frac{\epsilon}{\delta^4} > 1-\epsilon > \frac{1}{2}.$$

Then the full expectation of Δ_t is

$$\mathbb{E}\left[\Delta_t\right] \geq \frac{1}{2}\left\{\left(1-\frac{1+\delta}{1+\delta^4}\right)\underbrace{\frac{\alpha}{\sqrt{2\beta_2 M_1\delta^{(5+8\mu)/(1+\mu)}+(1-\beta_2)(1+2\delta)^2}}}_{T_1} - \underbrace{\frac{1+\delta}{1+\delta^4}\frac{\alpha}{\sqrt{1-\beta_2}}}_{T_2}\right\} - \underbrace{\frac{1}{2\delta^4}\frac{\alpha}{\sqrt{1-\beta_2}}}_{T_3}.$$

Notice that $T_1 = \Omega(\delta^{-(5+8\mu)/(2+2\mu)})$, $T_2 = \mathcal{O}(\delta^{-3})$ and $T_3 = \mathcal{O}(\delta^{-4})$. As long as we set $\mu < 1/2$, we have $(5+8\mu)/(2+2\mu) < 3 < 4$, thus the right hand side can be positive for sufficiently large δ, only dependent on α and β_2. We conclude that we can assume $\mathbb{E}[\Delta_t] > c_0 > 0$. This means w_t keeps drifting in the positive direction. Then for any $k \geq 1$ and $t > T_\epsilon$, we have

$$\begin{aligned}\mathbb{E}\left[(w_{t+k}-w^*)^2\right] &= \mathbb{E}\left[(w_{t+k}-w_t)^2\right] + 2\mathbb{E}\left[(w_{t+k}-w_t)(w_t-w^*)\right] + \mathbb{E}\left[(w_t-w^*)^2\right] \\ &\geq \mathbb{E}\left[(w_{t+k}-w_t)^2\right] - 2\sqrt{\mathbb{E}\left[(w_{t+k}-w_t)^2\right]\mathbb{E}\left[(w_t-w^*)^2\right]} + \mathbb{E}\left[(w_t-w^*)^2\right] \\ &= \left(\sqrt{\mathbb{E}\left[(w_{t+k}-w_t)^2\right]} - \sqrt{\mathbb{E}\left[(w_t-w^*)^2\right]}\right)^2, \end{aligned} \tag{4}$$

where the inequality is the Cauchy-Schwartz inequality for random variables. If we select k large enough such that $k > 3\sqrt{\epsilon}/c_0$, which implies $kc_0 - \sqrt{\epsilon} > 2\sqrt{\epsilon}$, then $\mathbb{E}[(w_{t+k}-w_t)^2] \geq (\mathbb{E}[w_{t+k}-w_t])^2 \geq k^2c_0^2$, and thus from (4) we have

$$\mathbb{E}\left[(w_{t+k}-w^*)^2\right] \geq (kc_0-\sqrt{\epsilon})^2 \geq 4\epsilon > \epsilon,$$

which contradicts the convergence assumption. Thus, ADAM diverges for this unconstrained stochastic optimization problem. This completes the proof of Theorem 1.

1.3 Proof of Theorem 2

Consider function $\pi(\delta) = (1+\delta)/(1+\delta^4)$. We notice that $\pi(1) = 1$, $\pi(\delta) \leq 1$ for $\delta \geq 1$, and $\pi(+\infty) = 0$. Since π has only a finite number of stationary points, there exists a $\bar{\delta}$ such that π is decreasing on $[\bar{\delta}, \infty)$. Thus for any b, there exists an N_b^* such that for any $N \geq N_b^*, N > b$ there exists a $\delta_{N,b} > \max(\delta^*, \bar{\delta}) > \delta^*$ with

$$\frac{b}{N} = \pi(\delta_{N,b}).$$

Let us consider the following mini-batch problem with sample size $N > N_b^*$.

$$\begin{aligned} f_n(w) &= \frac{w^2}{2\delta_{N,b}} - w \quad \text{for } n = 1,\ldots,N-1, \\ f_N(w) &= \frac{w^2}{2\delta_{N,b}} + (b\delta_{N,b}^4 + b - 1)w. \end{aligned}$$

Apparently, the selection of mini-batch $\mathcal{B}_t$ satisfies

$$\mathbb{P}\{N \in \mathcal{B}_t\} = \frac{\binom{N-1}{b-1}}{\binom{N}{b}} = \frac{b}{N} = \frac{1+\delta_{N,b}}{1+\delta_{N,b}^4} := p.$$

If $M \in \mathcal{B}_t$, we have

$$F^{\mathcal{B}_t}(w) = \frac{1}{b}\left(f_N(w) + (b-1)f_1(w)\right) = \frac{w^2}{2\delta_{N,b}} + \delta_{N,b}^b w.$$

Otherwise, it is clear that

$$F^{\mathcal{B}_t}(w) = f_1(w) = \frac{w^2}{2\delta_{N,b}} - w.$$

To summarize, the mini-batch loss reads

$$F^{\mathcal{B}_t}(w) = \begin{cases} \frac{w^2}{2\delta_{N,b}} + \delta_{N,b}^4 w & \text{with probability } \ p \\ \frac{w^2}{2\delta_{N,b}} - w & \text{with probability } \ 1-p \end{cases}$$

which is an $\mathrm{OP}(\delta_{N,b})$, since $\delta_{N,b} > \delta^*$. By Theorem 1, ADAM diverges on this problem.

1.4 Proof of Theorem 3

Similarly to the proof of Theorem 2, there exists N^* such that for each $N > N^*$, there exists a $\delta_N > \delta^*$ such that

$$\frac{1}{N} = \frac{1+\delta_N}{1+\delta_N^4}.$$

We let

$$\begin{aligned} f_n(w) &= \frac{w^2}{2\delta_N} + \delta_N^4 w \ \text{ for } n = 1, \ldots, N-1 \\ f_N(w) &= \frac{w^2}{2\delta_N} - ((N-1) + (N-2)\delta_N^4) w. \end{aligned}$$

The selection of mini-batch $\mathcal{B}_t$ satisfies

$$\mathbb{P}\{N \notin \mathcal{B}_t\} = \frac{1}{N}.$$

If $N \notin \mathcal{B}_t$, we have

$$F^{\mathcal{B}_t}(w) = f_1(w) = \frac{w^2}{2\delta_N} + \delta_N^4 w,$$

and otherwise

$$F^{\mathcal{B}_t}(w) = \frac{N-2}{N-1} f_1(w) + \frac{1}{N-1} f_N(w) = \frac{w^2}{2\delta_N} - w.$$

This is an $\mathrm{OP}(\delta_N)$, which is divergent according to Theorem 1.

1.5 Proof of Theorem 4

We first introduce the following lemma.

Lemma 5 *Given Assumption 1 is satisfied, there exist positive constants Q_1 and Q_2 such that for any t,*

$$\mathbb{E}[F(w_{t+1})] - \mathbb{E}[F(w_t)] \leq -\frac{\alpha_t}{4\sqrt{G^2+\epsilon}} \mathbb{E}\left[\|\nabla F(w_t)\|_2^2\right] + Q_1 \alpha_t \lambda_t + Q_2 \alpha_t \sum_{k=1}^{t-1} \beta_1^{t-k} \lambda_k + Q_3 \alpha_t^2.$$

Proof: Let us start from the application of L-smoothness of gradient of $F(w)$ as follows.

$$\begin{aligned}
F(w_{t+1}) &\leq F(w_t) + \nabla F(w_t)^\top (w_{t+1} - w_t) + \frac{L}{2}\|w_{t+1} - w_t\|_2^2 \\
&= F(w_t) - \frac{\alpha_t}{1-\beta_1^t}\nabla F(w_t)^\top V_t^{-1/2} m_t + \frac{\alpha_t^2 L}{2(1-\beta_1^t)^2}\left\|V_t^{-1/2} m_t\right\|_2^2 \\
&= F(w_t) - \frac{\alpha_t(1-\beta_1)}{1-\beta_1^t}\nabla F(w_t)^\top V_t^{-1/2}\sum_{k=1}^{t}\beta_1^{t-k} g_k + \frac{\alpha_t^2 L}{2(1-\beta_1^t)^2}\left\|V_t^{-1/2} m_t\right\|_2^2 \\
&= F(w_t) - \alpha_t \nabla F(w_t)^\top V_t^{-1/2} g_t - \frac{\alpha_t(1-\beta_1)}{1-\beta_1^t}\sum_{k=1}^{t-1}\beta_1^{t-k}\nabla F(w_t)^\top V_t^{-1/2}(g_k - g_t) \\
&\quad + \frac{\alpha_t^2 L}{2(1-\beta_1^t)^2}\left\|V_t^{-1/2} m_t\right\|_2^2 \\
&= F(w_t) - \alpha_t \underbrace{\nabla F(w_t)^\top V_t^{-1/2}\mathcal{G}(w_t;\xi_t)}_{T_1} \\
&\quad \underbrace{- \frac{\alpha_t(1-\beta_1)}{1-\beta_1^t}\sum_{k=1}^{t-1}\beta_1^{t-k}\nabla F(w_t)^\top V_t^{-1/2}\left(\mathcal{G}(w_k;\xi_k) - \mathcal{G}(w_k;\xi_t)\right)}_{T_2} \\
&\quad \underbrace{- \frac{\alpha_t(1-\beta_1)}{1-\beta_1^t}\sum_{k=1}^{t-1}\beta_1^{t-k}\nabla F(w_t)^\top V_t^{-1/2}\left(\mathcal{G}(w_k;\xi_t) - \mathcal{G}(w_t;\xi_t)\right)}_{T_3} \\
&\quad + \frac{\alpha_t^2 L}{2(1-\beta_1^t)^2}\underbrace{\left\|V_t^{-1/2} m_t\right\|_2^2}_{T_4}.
\end{aligned}$$

Bounding T_1***:*** We start from

$$\begin{aligned}
\mathbb{E}[T_1] &= \mathbb{E}\left[\left\|V_t^{-1/4}\nabla F(w_t)\right\|_2^2\right] + \mathbb{E}\left[\nabla F(w_t)^\top V_t^{-1/2}(\mathcal{G}(w_t;\xi_t) - \nabla F(w_t))\right] \\
&\geq \frac{1}{2}\mathbb{E}\left[\left\|V_t^{-1/4}\nabla F(w_t)\right\|_2^2\right] - \frac{1}{2}\mathbb{E}\left[\left\|V_t^{-1/4}(\mathcal{G}(w_t;\xi_t) - \nabla F(w_t))\right\|_2^2\right] \\
&\geq \frac{1}{2\sqrt{G^2+\epsilon}}\mathbb{E}\left[\|\nabla F(w_t)\|_2^2\right] - \frac{1}{2\sqrt{\epsilon}}\mathbb{E}\left[\|\mathcal{G}(w_t;\xi_t) - \nabla F(w_t)\|_2^2\right],
\end{aligned}$$

where the first inequality applies the Cauchy-Schwartz inequality and the second inequality is due to

$$\left\|V_t^{-1/4}\nabla F(w_t)\right\|_2^2 = \sum_{i=1}^{d}\frac{\left(\nabla_i F(w_t)\right)^2}{\sqrt{\tilde{v}_{t,i}+\epsilon}} \geq \frac{1}{\sqrt{G^2+\epsilon}}\sum_{i=1}^{d}\left(\nabla_i F(w_t)\right)^2 = \frac{1}{\sqrt{G^2+\epsilon}}\|\nabla F(w_t)\|_2^2$$

and

$$\begin{aligned}
\left\|V_t^{-1/4}(\mathcal{G}(w_t;\xi_t) - \nabla F(w_t))\right\|_2^2 &= \sum_{i=1}^{d}\frac{\left(\nabla_i F(w_t) - \mathcal{G}_i(w_t;\xi_t)\right)^2}{\sqrt{\tilde{v}_{t,i}+\epsilon}} \\
&\leq \frac{1}{\sqrt{\epsilon}}\sum_{i=1}^{d}\left(\nabla_i F(w_t) - \mathcal{G}_i(w_t;\xi_t)\right)^2 \\
&= \frac{1}{\sqrt{\epsilon}}\|\mathcal{G}(w_t;\xi_t) - \nabla F(w_t)\|_2^2.
\end{aligned}$$

According to the unbiased assumption, we have

$$\mathbb{E}\left[\|\mathcal{G}(w_t;\xi_t) - \nabla F(w_t)\|_2^2\right] = \sum_{i=1}^{d}\mathbb{E}\left[\left(\mathcal{G}_i(w_t;\xi_t) - \nabla_i F(w_t)\right)^2\right] = \sum_{i=1}^{d}\mathrm{Var}(\mathcal{G}_i(w_t;\xi_t)) \leq d\lambda_t. \tag{5}$$

Then we can lower bound the expectation of T_1 as

$$\mathbb{E}[T_1] \geq \frac{1}{2\sqrt{G^2+\epsilon}}\mathbb{E}\left[\|\nabla F(w_t)\|_2^2\right] - \frac{d}{2\sqrt{\epsilon}}\lambda_t. \tag{6}$$

Bounding T_2***:*** Notice that for two random vectors X and Y, and a constant a we have

$$\left\|aX + \frac{1}{a}Y\right\|_2^2 = a^2\|X\|_2^2 + \frac{1}{a^2}\|Y\|_2^2 + 2Y^\top X,$$

and thus

$$\mathbb{E}\left[Y^\top X\right] \geq -\frac{a^2}{2}\mathbb{E}[\|X\|_2^2] - \frac{1}{2a^2}\mathbb{E}[\|Y\|_2^2].$$

If $\mathcal{F}_k = \{\xi_1, \ldots, \xi_{k-1}\}$, then w_k is known given $\mathcal{F}_k$. We then have

$$\begin{aligned}
\mathbb{E}\left[T_2\right] &= \sum_{k=1}^{t-1}\beta_1^{t-k}\mathbb{E}\left[\nabla F(w_t)^\top V_t^{-1/2}\left(\mathcal{G}(w_k;\xi_k) - \mathcal{G}(w_k;\xi_t)\right)\right] \\
&= \sum_{k=1}^{t-1}\beta_1^{t-k}\mathbb{E}\left[\mathbb{E}\left[\nabla F(w_t)^\top V_t^{-1/2}\left(\mathcal{G}(w_k;\xi_k) - \mathcal{G}(w_k;\xi_t)\right)|\mathcal{F}_k\right]\right] \\
&\geq -\frac{1}{2}\sum_{k=1}^{t-1}\beta_1^{t-k}\mathbb{E}\left[a^2\mathbb{E}\left[\|V_t^{-1/2}\nabla F(w_t)\|_2^2|\mathcal{F}_k\right] + \frac{1}{a^2}\mathbb{E}\left[\|\mathcal{G}(w_k;\xi_k) - \mathcal{G}(w_k;\xi_t)\|_2^2|\mathcal{F}_k\right]\right] \\
&\geq -\frac{1}{2}\sum_{k=1}^{t-1}\beta_1^{t-k}\left\{\frac{a^2}{\epsilon}\mathbb{E}\left[\|\nabla F(w_t)\|_2^2\right] + \frac{1}{a^2}\mathbb{E}\left[\mathbb{E}\left[\|\mathcal{G}(w_k;\xi_k) - \mathcal{G}(w_k;\xi_t)\|_2^2|\mathcal{F}_k\right]\right]\right\} \\
&= -\frac{1}{2}\sum_{k=1}^{t-1}\beta_1^{t-k}\left\{\frac{a^2}{\epsilon}\mathbb{E}\left[\|\nabla F(w_t)\|_2^2\right] + \frac{2}{a^2}\mathbb{E}\left[\mathbb{E}\left[\|\nabla\mathcal{G}(w_k;\xi_k) - \nabla F(w_k)\|_2^2|\mathcal{F}_k\right]\right]\right\} \\
&\geq -\frac{a^2}{2\epsilon}\frac{1}{1-\beta_1}\mathbb{E}\left[\|\nabla F(w_t)\|_2^2\right] - \frac{1}{a^2}\sum_{k=1}^{t-1}\beta_1^{t-k}\mathbb{E}\left[\|\nabla\mathcal{G}(w_k;\xi_k) - \nabla F(w_k)\|_2^2\right] \\
&\geq -\frac{a^2}{2\epsilon}\frac{1}{1-\beta_1}\mathbb{E}\left[\|\nabla F(w_t)\|_2^2\right] - \frac{d}{a^2}\sum_{k=1}^{t-1}\beta_1^{t-k}\lambda_k
\end{aligned}$$

for any positive constant a, where the third equality holds because $\mathcal{G}(w_k;\xi_k)$ and $\mathcal{G}(w_k;\xi_t)$ are i.i.d. given $\mathcal{F}_k$, and thus

$$\begin{aligned}
\mathbb{E}\left[\|\mathcal{G}(w_k;\xi_k) - \mathcal{G}(w_k;\xi_t)\|_2^2|\mathcal{F}_k\right] &= \mathbb{E}\left[\|\mathcal{G}(w_k;\xi_k) - \nabla F(w_k)\|_2^2|\mathcal{F}_k\right] + \mathbb{E}\left[\|\mathcal{G}(w_k;\xi_t) - \nabla F(w_k)\|_2^2|\mathcal{F}_k\right] \\
&\quad -2\mathbb{E}\left[\left(\mathcal{G}(w_k;\xi_k) - \nabla F(w_k)\right)|\mathcal{F}_k\right]^\top\mathbb{E}\left[\left(\mathcal{G}(w_k;\xi_t) - \nabla F(w_k)\right)|\mathcal{F}_k\right] \\
&= 2\mathbb{E}\left[\|\mathcal{G}(w_k;\xi_k) - \nabla F(w_k)\|_2^2|\mathcal{F}_k\right].
\end{aligned}$$

The last inequality applies (5).

If $a = \sqrt{\epsilon(1-\beta_1)/2\sqrt{G^2+\epsilon}}$, then we have

$$\mathbb{E}[T_2] \geq -\frac{1}{4\sqrt{G^2+\epsilon}}\mathbb{E}\left[\|\nabla F(w_t)\|_2^2\right] - \frac{2d\sqrt{G^2+\epsilon}}{\epsilon(1-\beta_1)}\sum_{k=1}^{t-1}\beta_1^{t-k}\lambda_k. \tag{7}$$

Bounding T_3***:*** We derive

$$\begin{aligned}
T_3 &= \sum_{k=1}^{t-1} \beta_1^{t-k} \nabla F(w_t) V_t^{-1/2} \left(\mathcal{G}(w_k;\xi_t) - \mathcal{G}(w_t;\xi_t)\right) \\
&\geq -\sum_{k=1}^{t-1} \beta_1^{t-k} \left\| \nabla F(w_t) V_t^{-1/2} \right\|_2 \left\| \mathcal{G}(w_k;\xi_t) - \mathcal{G}(w_t;\xi_t) \right\|_2 \\
&\geq -\frac{LG}{\sqrt{\epsilon}} \sum_{k=1}^{t-1} \beta_1^{t-k} \|w_t - w_k\| \\
&= -\frac{LG}{\sqrt{\epsilon}} \sum_{k=1}^{t-1} \beta_1^{t-k} \left\| \sum_{j=k}^{t-1} \alpha_j V_j^{-1/2} \tilde{m}_j \right\|_2 \\
&\geq -\frac{LG}{\sqrt{\epsilon}} \sum_{k=1}^{t-1} \beta_1^{t-k} \sum_{j=k}^{t-1} \alpha_j \left\| V_j^{-1/2} \tilde{m}_j \right\|_2 \\
&\geq -\frac{LG^2}{\epsilon\sqrt{1-\beta_1}} \sum_{k=1}^{t-1} \sum_{j=k}^{t-1} \beta_1^{t-k} \alpha_j \\
&= -\frac{LG^2}{\epsilon\sqrt{1-\beta_1}} \sum_{j=1}^{t-1} \alpha_j \sum_{k=1}^{j} \beta_1^{t-k} \\
&\geq -\frac{LG^2}{\epsilon(1-\beta_1)^{3/2}} \sum_{j=1}^{t-1} \alpha_j \beta_1^{t-j} \\
&\geq -\frac{LG^2 \bar{C}}{\epsilon(1-\beta_1)^{3/2}} \alpha_t, \qquad (8)
\end{aligned}$$

where the first inequality is the Cauchy-Schwartz inequality, the second inequality applies L smoothness of $G(\cdot;\xi)$ for any ξ, the forth inequality holds because

$$\left\| V_j^{-1/2} \tilde{m}_j \right\|_2 = \sqrt{\frac{1}{1-\beta_1^j} \sum_{i=1}^{d} \frac{m_{j,i}^2}{v_{j,i}+\epsilon}} \leq \frac{G}{\sqrt{\epsilon(1-\beta_1)}}$$

and the last inequality comes from Lemma 3.

Bounding T_4***:*** It is easy to show that

$$T_4 = \sum_{i=1}^{d} \frac{m_{t,i}^2}{v_{t,i}+\epsilon} \leq \frac{G^2}{\epsilon}. \qquad (9)$$

According to the bounds in (6), (7), (8) and (9),we get

$$\begin{aligned}
\mathbb{E}[F(w_{t+1})] - \mathbb{E}[F(w_t)] &\leq -\alpha_t \left\{ \frac{1}{2\sqrt{G^2+\epsilon}} \mathbb{E}\left[\|\nabla F(w_t)\|_2^2 \right] - \frac{d}{2\sqrt{\epsilon}} \lambda_t \right\} \\
&\quad - \frac{\alpha_t(1-\beta_1)}{1-\beta_1^t} \left\{ -\frac{1}{4\sqrt{G^2+\epsilon}} \mathbb{E}\left[\|\nabla F(w_t)\|_2^2 \right] - \frac{2d\sqrt{G^2+\epsilon}}{\epsilon(1-\beta_1)} \sum_{k=1}^{t-1} \beta_1^{t-k} \lambda_k \right\} \\
&\quad + \frac{LG^2\bar{C}}{\epsilon\sqrt{1-\beta_1}(1-\beta_1^t)} \alpha_t^2 + \frac{LG^2}{2\epsilon(1-\beta_1^t)^2} \alpha_t^2 \\
&\leq -\alpha_t \frac{1}{4\sqrt{G^2+\epsilon}} \mathbb{E}\left[\|\nabla F(w_t)\|_2^2 \right] + \frac{d}{2\sqrt{\epsilon}} \alpha_t \lambda_t + \frac{2d\sqrt{G^2+\epsilon}}{\epsilon(1-\beta_1)} \alpha_t \sum_{k=1}^{t-1} \beta_1^{t-k} \lambda_k \\
&\quad + \left\{ \frac{LG^2\bar{C}}{\epsilon(1-\beta_1)^{3/2}} + \frac{LG^2}{2\epsilon(1-\beta_1)^2} \right\} \alpha_t^2.
\end{aligned}$$

Letting

$$\begin{aligned} Q_1 &= \frac{d}{2\sqrt{\epsilon}} \\ Q_2 &= \frac{2d\sqrt{G^2+\epsilon}}{\epsilon(1-\beta_1)} \\ Q_3 &= \frac{LG^2\bar{C}}{\epsilon(1-\beta_1)^{3/2}} + \frac{LG^2}{2\epsilon(1-\beta_1)^2} \end{aligned}$$

completes the proof.

Proof of Theorem 4: According to Lemma 5, we have

$$\begin{aligned} F_{\inf} - F(w_1) &\leq \mathbb{E}[F(w_{T+1})] - F(w_1) \\ &= \sum_{t=1}^{T} \mathbb{E}[F(w_{t+1})] - \mathbb{E}[F(w_t)] \\ &\leq -\frac{1}{4\sqrt{G^2+\epsilon}} \sum_{i=1}^{T} \alpha_t \mathbb{E}\left[\|\nabla F(w_t)\|_2^2\right] + Q_1 \sum_{i=1}^{T} \alpha_t \lambda_t + Q_2 \sum_{i=1}^{T} \alpha_t \sum_{k=1}^{t-1} \beta_1^{t-k} \lambda_k \\ &\quad + Q_3 \sum_{i=1}^{T} \alpha_t^2. \end{aligned}$$

Then we obtain

$$\begin{aligned} \frac{1}{4\sqrt{G^2+\epsilon}} \sum_{t=1}^{T} \alpha_t \mathbb{E}\left[\|\nabla F(w_t)\|_2^2\right] &\leq F(w_1) - F_{\inf} + Q_1 \sum_{t=1}^{T} \alpha_t \lambda_t \\ &\quad + Q_2 \sum_{t=1}^{T} \alpha_t \sum_{k=1}^{t-1} \beta_1^{t-k} \lambda_k + Q_3 \sum_{t=1}^{T} \alpha_t^2 \\ &\leq F(w_1) - F_{\inf} + Q_1 \sum_{t=1}^{T} \alpha_t \lambda_t \\ &\quad + Q_2 \sum_{k=1}^{T} \lambda_k \sum_{t=k}^{T} \beta_1^{t-k} \alpha_t + Q_3 \sum_{t=1}^{T} \alpha_t^2 \\ &\leq F(w_1) - F_{\inf} + Q_1 \sum_{t=1}^{T} \alpha_t \lambda_t \\ &\quad + \frac{Q_2}{1-\beta_1} \sum_{k=1}^{T} \lambda_k \alpha_k + Q_3 \sum_{t=1}^{T} \alpha_t^2. \end{aligned}$$

Noticing that the left-hand side can be bounded as

$$\sum_{t=1}^{T} \alpha_t \mathbb{E}\left[\|\nabla F(w_t)\|_2^2\right] \geq \sum_{t=1}^{T} \alpha_t \min_{1 \leq t \leq T} \mathbb{E}\left[\|\nabla F(w_t)\|_2^2\right],$$

we obtain

$$\begin{aligned} \min_{1 \leq t \leq T} \mathbb{E}\left[\|\nabla F(w_t)\|_2^2\right] &\leq \frac{4\sqrt{G^2+\epsilon}}{\sum_{t=1}^{T} \alpha_t} + 4\sqrt{G^2+\epsilon}\left(Q_1 + \frac{Q_2}{1-\beta_2}\right) \frac{\sum_{t=1}^{T} \lambda_t \alpha_t}{\sum_{t=1}^{T} \alpha_t} \\ &\quad + 4\sqrt{G^2+\epsilon} Q_3 \frac{\sum_{t=1}^{T} \alpha_t^2}{\sum_{t=1}^{T} \alpha_t}. \end{aligned}$$

1.6 Proof of Theorem 5

Applying L-smoothness of gradients of F and strong convexity of F, we have

$$\begin{aligned} F\left(\hat{w}_2^{(2)}\right) &\geq F\left(w_1^{(2)}\right) + F'\left(w_1^{(2)}\right)\left(\hat{w}_2^{(2)} - w_1^{(2)}\right) + \frac{c}{2}\left(\hat{w}_2^{(2)} - w_1^{(2)}\right)^2 \\ F\left(w_2^{(2)}\right) &\leq F\left(w_1^{(2)}\right) + F'\left(w_1^{(2)}\right)\left(w_2^{(2)} - w_1^{(2)}\right) + \frac{L}{2}\left(w_2^{(2)} - w_1^{(2)}\right)^2. \end{aligned}$$

By definition, we have

$$\begin{aligned} F\left(\hat{w}_2^{(2)}\right) - F\left(w_2^{(2)}\right) &\geq F'\left(w_1^{(2)}\right)\left(\hat{w}_2^{(2)} - w_2^{(2)}\right) + \frac{c}{2}\left(\hat{w}_2^{(2)} - w_1^{(2)}\right)^2 - \frac{L}{2}\left(w_2^{(2)} - w_1^{(2)}\right)^2 \\ &= F'\left(w_1^{(2)}\right)\left(\alpha_2 \frac{\widetilde{m}_1^{(2)}}{\sqrt{\widetilde{v}_1^{(2)} + \epsilon}} - \alpha_2 \frac{\hat{\widetilde{m}}_1^{(2)}}{\sqrt{\hat{\widetilde{v}}_1^{(2)} + \epsilon}}\right) + \frac{c\alpha_2^2}{2}\frac{\hat{\widetilde{m}}_1^{(2)}}{\hat{\widetilde{v}}_1^{(2)} + \epsilon} - \frac{L\alpha_2^2}{2}\frac{\widetilde{m}_1^{(2)}}{\widetilde{v}_1^{(2)} + \epsilon} \\ &= \alpha_2 F'\left(w_1^{(2)}\right)\left(\frac{F'\left(w_1^{(2)}\right)}{\sqrt{Q_3}} - \gamma\frac{(1-\beta_1)F'\left(w_1^{(2)}\right) + \beta_1 m_{m+1}^{(1)}}{\sqrt{Q_4}}\right) \\ &\quad + \frac{c\alpha_2^2\gamma^2}{2}\frac{\left((1-\beta_1)F'\left(w_1^{(2)}\right) + \beta_1 m_2^{(1)}\right)^2}{Q_4} - \frac{L\alpha_2^2}{2}\frac{\left(F'\left(w_1^{(2)}\right)\right)^2}{Q_3} \end{aligned}$$

where

$$\begin{aligned} Q_3 &= \widetilde{v}_1^{(2)} + \epsilon \\ Q_4 &= \hat{\widetilde{v}}_1^{(2)} + \epsilon \\ \gamma &= \frac{1}{1-\beta_1^{m+1}}. \end{aligned}$$

Thus we have

$$F\left(\hat{w}_2^{(2)}\right) - F\left(w_2^{(2)}\right) \geq \left(F'\left(w_1^{(2)}\right)\right)^2 q\left(\frac{m_{m+1}^{(1)}}{F'\left(w_1^{(2)}\right)}\right)$$

where $q(x) = Q_5 x^2 + Q_6 x + Q_7$ is a function with parameters

$$\begin{aligned} Q_5 &= \frac{c\alpha_2^2\gamma^2\beta_1^2}{2Q_4} \\ Q_6 &= \frac{c\alpha_2^2\gamma^2\beta_1(1-\beta_1)}{Q_4} - \frac{\alpha_2\gamma\beta_1}{\sqrt{Q_2}} \\ Q_7 &= \frac{\alpha_2}{\sqrt{Q_3}} - \frac{\alpha_2\gamma(1-\beta_1)}{\sqrt{Q_4}} + \frac{c\alpha_2^2\gamma^2(1-\beta_1)^2}{2Q_4} - \frac{L\alpha_2^2}{2Q_3}. \end{aligned}$$

Apparently, from

$$\begin{aligned} \widetilde{v}_1^{(2)} &= \left(g_1^{(2)}\right)^2 \leq G^2 \\ \hat{\widetilde{v}}_1^{(2)} &= \frac{1-\beta_1}{1-\beta_2^{m+1}}\left(\sum_{k=1}^{m}\beta_2^{m+1-k}\left(g_k^{(1)}\right)^2 + \left(g_1^{(2)}\right)^2\right) \leq \frac{1-\beta_1}{1-\beta_2^{m+1}}\left(\sum_{k=1}^{m}\beta_2^{m+1-k} + 1\right)G^2 = G^2, \end{aligned}$$

we have

$$\begin{aligned} \epsilon \leq \quad & Q_3 \quad \leq \epsilon + G^2 \\ \epsilon \leq \quad & Q_4 \quad \leq \epsilon + G^2. \end{aligned}$$

Noticing that

$$\begin{aligned}\Delta &= Q_6^2 - 4Q_5Q_7 \\ &= \frac{\alpha_2^2\gamma^2\beta_1^2}{Q_4}\left(1 - \frac{2c\alpha_2}{\sqrt{Q_3}} + \frac{cL\alpha_2^2}{Q_3}\right) \\ &> \frac{\alpha_2^2\gamma^2\beta_1^2}{Q_4}\left(1 - \frac{2c\alpha_2}{\sqrt{Q_3}} + \frac{c^2\alpha_2^2}{Q_3}\right) \\ &= \frac{\alpha_2^2\gamma^2\beta_1^2}{Q_4}\left(1 - \frac{c\alpha_2}{\sqrt{Q_3}}\right)^2 \geq 0,\end{aligned}$$

where the first inequality uses the property of the strong convexity parameter and the L-smoothness gradient parameter $c < L$, we have that there exists

$$\begin{aligned}x_1 &= \frac{-Q_6 + \sqrt{\Delta}}{2Q_5} \\ x_2 &= \frac{-Q_6 - \sqrt{\Delta}}{2Q_5}\end{aligned}$$

such that $q(x_1) = q(x_2) = 0$. We claim that $|x_1| \leq 1$ and $|x_2| \leq 1$, which is implied by

$$\sqrt{\Delta} \leq \min\{2Q_5 + Q_6, 2Q_5 - Q_6\}. \tag{10}$$

We notice that

$$\begin{aligned}2Q_5 + Q_6 &= \frac{\alpha_2\gamma\beta_1}{\sqrt{Q_4}}\left(\frac{c\alpha_2\gamma}{\sqrt{Q_4}} - 1\right) \\ 2Q_5 - Q_6 &= \frac{\alpha_2\gamma\beta_1}{\sqrt{Q_4}}\left(\frac{c\alpha_2\gamma(2\beta_1 - 1)}{\sqrt{Q_4}} + 1\right)\end{aligned}$$

and

$$\begin{aligned}\Delta &\leq \frac{\alpha_2^2\gamma^2\beta_1^2}{Q_4}\left(1 - \frac{2L\alpha_2}{\sqrt{Q_3}} + \frac{L^2\alpha_2^2}{Q_3}\right) \\ &= \frac{\alpha_2^2\gamma^2\beta_1^2}{Q_4}\left(1 - \frac{L\alpha_2}{\sqrt{Q_3}}\right)^2\end{aligned}$$

where the inequality holds according to Assumption 2 $L\alpha_2 \geq 2\sqrt{G^2 + \epsilon} \geq 2\sqrt{Q_3}$. Thus we have

$$\sqrt{\Delta} \leq \frac{\alpha_2\gamma\beta_1}{\sqrt{Q_4}}\left(\frac{L\alpha_2}{\sqrt{Q_3}} - 1\right) \leq \frac{\alpha_2\gamma\beta_1}{\sqrt{Q_4}}\left(\frac{c\alpha_2\gamma(2\beta_1 - 1)}{\sqrt{Q_4}} - 1\right) \leq \min\{2Q_5 + Q_6, 2Q_5 - Q_6\},$$

where the second inequality holds according to Assumption 2, $\frac{L}{c} \leq \frac{2\beta_1 - 1}{1 - \beta_1^{m+1}}\sqrt{\frac{\epsilon}{G+\epsilon}} \leq \frac{2\beta_1 - 1}{1 - \beta_1^{m+1}}\sqrt{\frac{Q_3}{Q_4}}$.

Hence we obtain (10), which implies that $q(x) \geq 0$ where $|x| \geq 1$. As we assume

$$\left|m_{m+1}^{(1)}\right| \geq \left|F'\left(w_1^{(2)}\right)\right|,$$

we have

$$F\left(\hat{w}_2^{(2)}\right) - F\left(w_2^{(2)}\right) \geq 0,$$

which finishes the proof.

1.7 Proof of Theorem 6 and Theorem 7

We start proving the following lemmas.

Lemma 6 *Given Assumption 3, we have that for any* $1 \leq k \leq m$ *and* $1 \leq t \leq T$*, the ADAM states in Algorithm 2 with option A satisfy*

$$\begin{aligned} \left\| m_k^{(t)} \right\|_2 &\leq 3G, \\ \left\| v_k^{(t)} \right\|_2 &\leq 9G^2. \end{aligned}$$

Proof: By definition, we have

$$\begin{aligned} m_k^{(t)} &= (1-\beta_1) \sum_{j=1}^{k} \beta_1^{k-j} g_j^{(t)} \\ v_k^{(t)} &= (1-\beta_2) \sum_{j=1}^{k} \beta_2^{k-j} g_j^{(t)} \odot g_j^{(t)}. \end{aligned}$$

Applying the Cauchy-Schwartz inequality, we obtain

$$\begin{aligned} \left\| m_k^{(t)} \right\|_2 &\leq (1-\beta_1) \sum_{j=1}^{k} \beta_1^{k-j} \left\| g_j^{(t)} \right\|_2 \\ &\leq (1-\beta_1) \sum_{j=1}^{k} \beta_1^{k-j} \left(\left\| \nabla F^{\mathcal{B}_j^{(t)}} \left(w_j^{(t)} \right) \right\|_2 + \left\| \nabla F^{\mathcal{B}_j^{(t)}} \left(\widetilde{w}_t \right) \right\|_2 + \left\| \nabla F \left(\widetilde{w}_t \right) \right\|_2 \right) \\ &\leq (1-\beta_1) \sum_{j=1}^{k} \beta_1^{k-j} 3G \leq 3G \end{aligned}$$

and

$$\begin{aligned} \left\| v_k^{(t)} \right\|_2 &\leq (1-\beta_2) \sum_{j=1}^{k} \beta_2^{k-j} \left\| g_j^{(t)} \odot g_j^{(t)} \right\|_2 \\ &= (1-\beta_2) \sum_{j=1}^{k} \beta_2^{k-j} \left\| g_j^{(t)} \right\|_2^2 \\ &\leq (1-\beta_2) \sum_{j=1}^{k} \beta_2^{k-j} \left(\left\| \nabla F^{\mathcal{B}_j^{(t)}} \left(w_j^{(t)} \right) \right\|_2 + \left\| \nabla F^{\mathcal{B}_j^{(t)}} \left(\widetilde{w}_t \right) \right\|_2 + \left\| \nabla F \left(\widetilde{w}_t \right) \right\|_2 \right)^2 \\ &\leq (1-\beta_2) \sum_{j=1}^{k} \beta_2^{k-j} 9G^2 \leq 9G^2. \end{aligned}$$

Lemma 7 *Given Assumption 3, there exist positive constants* Q_8 *and* Q_9 *such that Algorithm 2 with option A satisfies that for any* t*,*

$$F(\widetilde{w}_{t+1}) - F(\widetilde{w}_t) \leq -Q_8 \alpha_t m \left\| \nabla F(\widetilde{w}_t) \right\|_2^2 + Q_9 \alpha_t^2$$

holds almost surely.

Proof: We start from the application of L-smoothness of gradient of $F(w)$ as follows

$$\begin{aligned} F(\widetilde{w}_{t+1}) &\leq F(\widetilde{w}_t) + \nabla F(\widetilde{w}_t)^\top (\widetilde{w}_{t+1} - \widetilde{w}_t) + \frac{L}{2} \left\| \widetilde{w}_{t+1} - \widetilde{w}_t \right\|_2^2 \\ &= F(\widetilde{w}_t) + \nabla F(\widetilde{w}_t)^\top \sum_{k=1}^{m} \left(w_{k+1}^{(t)} - w_k^{(t)} \right) + \frac{L}{2} \left\| \sum_{k=1}^{m} \left(w_{k+1}^{(t)} - w_k^{(t)} \right) \right\|_2^2 \\ &= F(\widetilde{w}_t) - \alpha_t \nabla F(\widetilde{w}_t)^\top \sum_{k=1}^{m} \left(V_k^{(t)} \right)^{-1/2} \widetilde{m}_k^{(t)} + \frac{\alpha_t^2 L}{2} \left\| \sum_{k=1}^{m} \left(V_k^{(t)} \right)^{-1/2} \widetilde{m}_k^{(t)} \right\|_2^2. \end{aligned}$$

By definition and the resetting option, we have

$$\widetilde{m}_k^{(t)} = \frac{1-\beta_1}{1-\beta_1^k}\left(V_k^{(t)}\right)^{-1/2}\sum_{j=1}^{k}\beta_1^{k-j}g_j^{(t)},$$

and thus

$$\begin{aligned}
F(\widetilde{w}_{t+1}) &\leq F(\widetilde{w}_t) - \alpha_t(1-\beta_1)\nabla F(\widetilde{w}_t)^\top \sum_{k=1}^{m}\frac{1}{1-\beta_1^k}\left(V_k^{(t)}\right)^{-1/2}\sum_{j=1}^{k}\beta_1^{k-j}g_j^{(t)} \\
&\quad + \frac{\alpha_t^2 L}{2}\left\|\sum_{k=1}^{m}\left(V_k^{(t)}\right)^{-1/2}\widetilde{m}_k^{(t)}\right\|_2^2 \\
&= F(\widetilde{w}_t) - \alpha_t(1-\beta_1)\nabla F(\widetilde{w}_t)^\top \sum_{j=1}^{m}\left(\sum_{k=j}^{m}\frac{\beta_1^{k-j}}{1-\beta_1^k}\left(V_k^{(t)}\right)^{-1/2}\right)g_j^{(t)} \\
&\quad + \frac{\alpha_t^2 L}{2}\left\|\sum_{k=1}^{m}\left(V_k^{(t)}\right)^{-1/2}\widetilde{m}_k^{(t)}\right\|_2^2 \\
&= F(\widetilde{w}_t) - \alpha_t(1-\beta_1) \\
&\quad \times \underbrace{\nabla F(\widetilde{w}_t)^\top \sum_{j=1}^{m}\left(\sum_{k=j}^{m}\frac{\beta_1^{k-j}}{1-\beta_1^k}\left(V_k^{(t)}\right)^{-1/2}\right)\left(\nabla F^{\mathcal{B}_j^{(t)}}\left(w_j^{(t)}\right) - \nabla F^{\mathcal{B}_j^{(t)}}(\widetilde{w}_t)\right)}_{T_1} \\
&\quad - \alpha_t(1-\beta_1)\underbrace{\nabla F(\widetilde{w}_t)^\top\left(\sum_{j=1}^{m}\sum_{k=j}^{m}\frac{\beta_1^{k-j}}{1-\beta_1^k}\left(V_k^{(t)}\right)^{-1/2}\right)\nabla F(\widetilde{w}_t)}_{T_2} \\
&\quad + \frac{\alpha_t^2 L}{2}\underbrace{\left\|\sum_{k=1}^{m}\left(V_k^{(t)}\right)^{-1/2}\widetilde{m}_k^{(t)}\right\|_2^2}_{T_3}.
\end{aligned}$$

Bounding T_1: We have

$$T_1 \geq -\sum_{j=1}^{m}\left\|\sum_{k=j}^{m}\frac{\beta_1^{k-j}}{1-\beta_1^k}\left(V_k^{(t)}\right)^{-1/2}\nabla F(\widetilde{w}_t)\right\|_2\left\|\nabla F^{\mathcal{B}_j^{(t)}}\left(w_j^{(t)}\right) - F^{\mathcal{B}_j^{(t)}}(\widetilde{w}_t)\right\|_2. \tag{11}$$

The first thing to notice is that

$$\begin{aligned}
\left\|\sum_{k=j}^{m}\frac{\beta_1^{k-j}}{1-\beta_1^k}\left(V_k^{(t)}\right)^{-1/2}\nabla F(\widetilde{w}_t)\right\|_2 &\leq \sum_{k=j}^{m}\frac{\beta_1^{k-j}}{1-\beta_1^k}\left\|\left(V_k^{(t)}\right)^{-1/2}\nabla F(\widetilde{w}_t)\right\|_2 \\
&= \sum_{k=j}^{m}\frac{\beta_1^{k-j}}{1-\beta_1^k}\sqrt{\sum_{i=1}^{d}\frac{\nabla_i F(\widetilde{w}_t)^2}{v_{k,i}^{(t)}+\epsilon}} \\
&\leq \sum_{k=j}^{m}\frac{\beta_1^{k-j}}{1-\beta_1^k}\frac{\|\nabla F(\widetilde{w}_t)\|_2}{\sqrt{\epsilon}} \\
&\leq \frac{G}{\sqrt{\epsilon}}\sum_{k=j}^{m}\frac{\beta_1^{k-j}}{1-\beta_1^k} \\
&\leq \frac{G}{\sqrt{\epsilon}}\frac{1}{1-\beta_1}\sum_{k=j}^{m}\beta_1^{k-j} \leq \frac{G}{(1-\beta_1)^2\sqrt{\epsilon}},
\end{aligned} \tag{12}$$

where the second inequality employs the assumption that the gradients of F are bounded. Secondly, according to L-smoothness of gradients of every loss function, we derive

$$\begin{aligned}
\left\|\nabla F^{\mathcal{B}_j^{(t)}}\left(w_j^{(t)}\right)-\nabla F^{\mathcal{B}_j^{(t)}}\left(\widetilde{w}_t\right)\right\|_2 &\leq L\left\|w_j^{(t)}-w_1^{(t)}\right\|_2 \\
&= L\left\|\sum_{l=1}^{j-1}\alpha_t\left(V_l^{(t)}\right)^{-1/2}\widetilde{m}_l^{(t)}\right\|_2 \\
&\leq \alpha_t L\sum_{l=1}^{j-1}\left\|\left(V_l^{(t)}\right)^{-1/2}\widetilde{m}_l^{(t)}\right\|_2 \\
&= \alpha_t L\sum_{l=1}^{j-1}\frac{1}{1-\beta_1^l}\sqrt{\sum_{i=1}^{d}\frac{\left(m_{l,i}^{(t)}\right)^2}{v_{l,i}^{(t)}+\epsilon}} \\
&\leq \frac{\alpha_t L}{1-\beta_1}\sum_{l=1}^{j-1}\frac{\left\|m_l^{(t)}\right\|_2}{\sqrt{\epsilon}} \leq \frac{3GL}{(1-\beta_1)\sqrt{\epsilon}}(j-1)\alpha_t, \qquad (13)
\end{aligned}$$

where the first inequality applies the Cauchy-Schwartz inequality and the last one applies Lemma 6. By plugging equations (12) and (13) into equation (11), we obtain

$$T_1 \geq -\sum_{j=1}^{m}\frac{3G^2L}{(1-\beta_1)^3\epsilon}(j-1)\alpha_t = -\frac{3G^2L}{2(1-\beta_1)^3\epsilon}m(m-1)\alpha_t. \qquad (14)$$

Bounding T_2: We have

$$\begin{aligned}
T_2 &= \sum_{j=1}^{m}\sum_{k=j}^{m}\frac{\beta_1^{k-j}}{1-\beta_1^k}\nabla F(\widetilde{w}_t)^\top\left(V_k^{(t)}\right)^{-1/2}\nabla F(\widetilde{w}_t) \\
&= \sum_{j=1}^{m}\sum_{k=j}^{m}\frac{\beta_1^{k-j}}{1-\beta_1^k}\sum_{i=1}^{d}\frac{\nabla_i F(\widetilde{w}_t)^2}{\sqrt{v_{k,i}^{(t)}+\epsilon}} \\
&\geq \sum_{j=1}^{m}\sum_{k=j}^{m}\beta_1^{k-j}\sum_{i=1}^{d}\frac{\nabla_i F(\widetilde{w}_t)^2}{\sqrt{9G^2+\epsilon}} \\
&= \frac{\|\nabla F(\widetilde{w}_t)\|_2^2}{\sqrt{9G^2+\epsilon}}\sum_{j=1}^{m}\sum_{k=j}^{m}\beta_1^{k-j} \\
&\geq \frac{\|\nabla F(\widetilde{w}_t)\|_2^2}{\sqrt{9G^2+\epsilon}}\sum_{j=1}^{m}1 \\
&= \frac{1}{\sqrt{9G^2+\epsilon}}m\left\|\nabla F(\widetilde{w}_t)\right\|_2^2. \qquad (15)
\end{aligned}$$

Bounding T_3***:*** We obtain

$$\begin{aligned}
T_3 &\leq \left(\sum_{k=1}^{m}\left\|\left(V_k^{(t)}\right)^{-1/2}\widetilde{m}_k^{(t)}\right\|_2\right)^2 \\
&= \left(\sum_{k=1}^{m}\frac{1}{1-\beta_1^k}\sqrt{\sum_{i=1}^{d}\frac{\left(m_{k,i}^{(t)}\right)^2}{v_{k,i}^{(t)}+\epsilon}}\right)^2 \\
&\leq \left(\sum_{k=1}^{m}\frac{1}{\sqrt{\epsilon}}\left\|m_k^{(t)}\right\|_2\right)^2 \leq \left(\sum_{k=1}^{m}\frac{3G}{\sqrt{\epsilon}}\right)^2 \leq \frac{9G^2m^2}{\epsilon}. \qquad (16)
\end{aligned}$$

In summary, we get

$$\begin{aligned} F(\widetilde{w}_{t+1}) &\leq F(\widetilde{w}_t) - \alpha_t \frac{m(1-\beta_1)}{\sqrt{9G^2+\epsilon}} \left\|\nabla F(\widetilde{w}_t)\right\|_2^2 + \frac{3G^2Lm(m-1)/(1-\beta_1)^2 + 9G^2Lm^2}{2\epsilon}\alpha_t^2 \\ &= F(\widetilde{w}_t) - Q_8\alpha_t m \left\|\nabla F(\widetilde{w}_t)\right\|_2^2 + Q_9\alpha_t^2, \end{aligned}$$

where

$$\begin{aligned} Q_8 &= \frac{1-\beta_1}{\sqrt{9G^2+\epsilon}} \\ Q_9 &= \frac{3G^2Lm(m-1)/(1-\beta_1)^2 + 9G^2Lm^2}{2\epsilon}. \end{aligned}$$

Proof of Theorem 6: If $F(w)$ is c-strongly convex, we have

$$\left\|\nabla F(w)\right\|_2^2 \geq 2c\left(F(w) - F^*\right),$$

and thus according to Lemma 7, we have

$$F(\widetilde{w}_{t+1}) \leq F(\widetilde{w}_t) - 2cQ_8\alpha_t m\left(F(w_t) - F^*\right) + Q_9\alpha_t^2,$$

which is equivalent to

$$F(\widetilde{w}_{t+1}) - F^* \leq \left(1 - \frac{C_2 m\alpha}{t}\right)\left(F(\widetilde{w}_t) - F^*\right) + Q_9\alpha_t^2.$$

We obtain recursively

$$F(\widetilde{w}_T) - F^* \leq \prod_{t=1}^{T-1}\left(1 - \frac{C_2 m\alpha}{t}\right)\left(F(\widetilde{w}_1) - F^*\right) + \sum_{t=1}^{T-1}\alpha_t \prod_{j=t+1}^{T-1}\left(1 - \frac{C_2 m\alpha}{j}\right).$$

By definition, we have $C_2 m\alpha < 1$, and thus we can use Lemma 4 to obtain

$$F(\widetilde{w}_T) - F^* \leq \mathcal{O}\left(T^{-C_2 m\alpha}\right).$$

Proof of Theorem 7: Let us consider the set of indices $A = \{t \in \mathbb{N} : \|\nabla F(\widetilde{w}_t)\| = 0\}$. If the set is infinite, there exists a sequence $\{t_k\}_{k=1}^{+\infty}$ such that $\|\nabla F(\widetilde{w}_{t_k})\| = 0$ for all k. Then we have

$$\liminf_{t\to\infty} \left\|\nabla F(\widetilde{w}_t)\right\|_2 = 0.$$

Otherwise, A is finite, and thus its maximum exists. For all $t > \tau := \max A$, we have $\left\|\nabla F(\tilde{w}_t)\right\|_2 > 0$. Applying Lemma 7, we have

$$F(\widetilde{w}_{t+1}) - F(\widetilde{w}_t) \leq -\alpha_t Q_8 m \left\|\nabla F(\tilde{w}_t)\right\|_2^2 + Q_9\alpha_t^2.$$

Then it follows

$$\begin{aligned} F_{\inf} - F(\widetilde{w}_{\tau+1}) &\leq F(\widetilde{w}_{T+1}) - F(\widetilde{w}_{\tau+1}) \\ &\leq \sum_{t=\tau+1}^{T} -Q_8 m\alpha_t \left\|\nabla F(\widetilde{w}_t)\right\|_2^2 + Q_9 \sum_{t=\tau+1}^{T}\alpha_t^2, \end{aligned}$$

and thus

$$\min_{\tau+1\leq t\leq T} \left\|\nabla F(\widetilde{w}_t)\right\|_2^2 \sum_{t=\tau+1}^{T}\alpha_t \leq \sum_{t=\tau+1}^{T}\alpha_t \left\|\nabla F(\widetilde{w}_t)\right\|_2^2 \leq \frac{F(\widetilde{w}_{\tau+1}) - F_{\inf}}{Q_8 m} + \frac{Q_9}{Q_8 m}\sum_{t=\tau+1}^{T}\alpha_t^2.$$

Then, we have

$$\min_{\tau+1\leq t\leq T}\|\nabla F(\widetilde{w}_t)\|_2^2 \leq \frac{1}{\sum_{t=\tau+1}^{T}\alpha_t}\left\{\frac{F(\widetilde{w}_{\tau+1})-F_{\inf}}{Q_8 m}+\frac{Q_9}{Q_8 m}\sum_{t=\tau+1}^{T}\alpha_t^2\right\},$$

which yields

$$\lim_{T\to+\infty}\min_{\tau+1\leq t\leq T}\|\nabla F(\widetilde{w}_t)\|_2^2 = 0. \tag{17}$$

For all $r > \tau$, there must exists an $s > r$ such that

$$\|\nabla F(\widetilde{w}_s)\|_2 < \|\nabla F(\widetilde{w}_r)\|_2 .$$

Otherwise, if there exists an $r_0 > \tau$, such that for all $s > r_0$ we have

$$\|\nabla F(\widetilde{w}_s)\|_2 \geq \|\nabla F(\widetilde{w}_{r_0})\|_2 ,$$

then for all $T \geq r_0$, we have

$$\min_{\tau+1\leq t\leq T}\|\nabla F(\widetilde{w}_t)\|_2^2 = \min_{\tau+1\leq t\leq r_0}\|\nabla F(\widetilde{w}_t)\|_2^2 = A > 0,$$

which contradicts (17), since A is a positive constant.

Let $t_1 = \tau + 1$ and for all $k \in \mathbb{N}$, let $t_{k+1} = \inf\left\{s > t_k : \|\nabla F(\widetilde{w}_s)\|_2 < \|\nabla F(\widetilde{w}_{t_k})\|_2\right\}$. This implies a sub-sequence $\left\{\|\nabla F(\widetilde{w}_{t_k})\|_2\right\}_k$ of sequence $\{\|\nabla F(\widetilde{w}_t)\|_2\}_t$. Since

$$\|\nabla F(\widetilde{w}_{t_k})\|_2 = \min_{\tau+1\leq t\leq t_k}\|\nabla F(\widetilde{w}_t)\|_2 ,$$

by employing (17), we have

$$\lim_{k\to\infty}\|\nabla F(\widetilde{w}_{t_k})\|_2 = 0,$$

which implies that

$$\liminf_{t\to\infty}\|\nabla F(\widetilde{w}_t)\| = 0.$$

This completes the proof.

2 Experiments

2.1 Network Structure

Dataset	Input dimension	Hidden dimension	Output Dimension
CovType	98	100	7
MNIST	784	100	10

Table 1: Feedforward network structure

The feedforward networks used in the experiments have two fully connected layers with the dimensions described in Table 1.

The structure of the CNN used in the experiments is described as follows. The CNN is mainly composed of two convolution layers, two max pooling layers and one fully connected layer. The kernel size of the convolution layers is 4 and the kernel size of the pooling layers is 2. The numbers of channels of the two convolution layers are 16 and 32, respectively, and the dimensions of the fully connected layer are 32 for input and 10 for output.

2.2 Additional Results

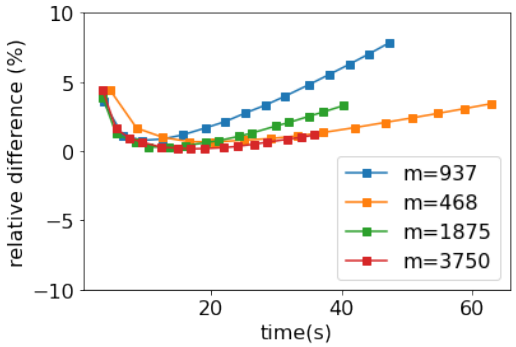

Figure 1: Relative difference of VRADAM in classifying Embedded CIFAR10 with Logistic regression

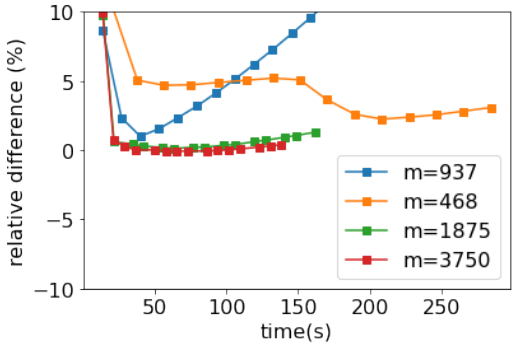

Figure 2: Relative difference of VRADAM in classifying MNIST with CNN

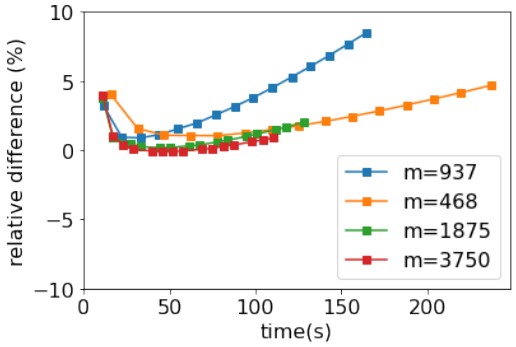

Figure 3: Relative difference of VRADAM in classifying MNIST with FFN

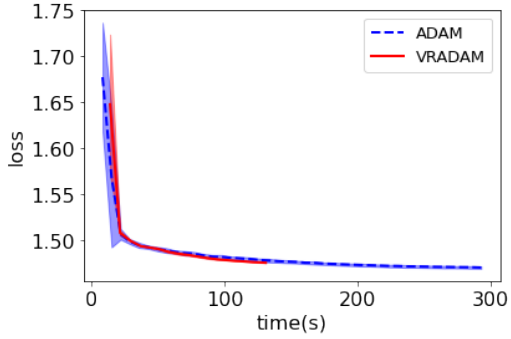

Figure 4: Deviation of VRADAM and ADAM for MNIST with CNN

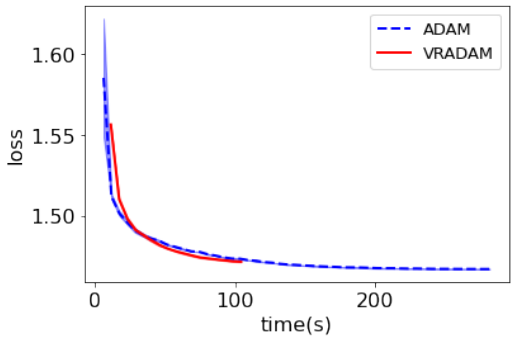

Figure 5: Deviation of VRADAM and ADAM for MNIST with FFN

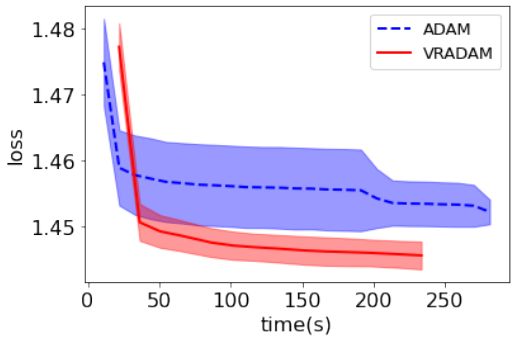

Figure 6: Deviation of VRADAM and ADAM for CovType with logistic regression

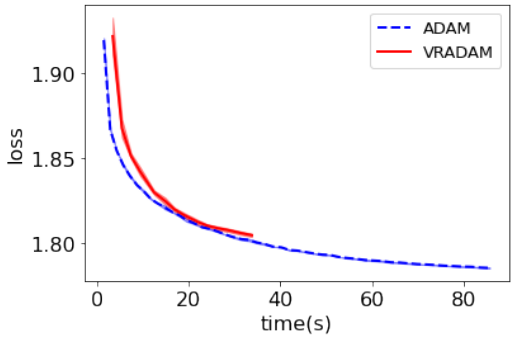

Figure 7: Deviation of VRADAM and ADAM for Embedded CIFAR-10 with logistic regression

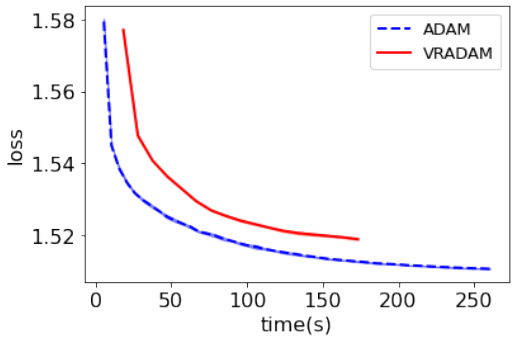

Figure 8: Deviation of VRADAM and ADAM for MNIST with logistic regression

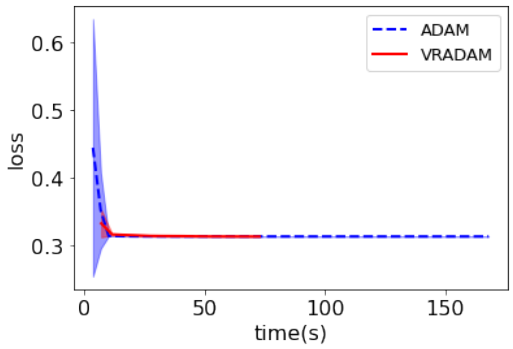

Figure 9: Deviation of VRADAM and ADAM for NSL-KDD with logistic regression

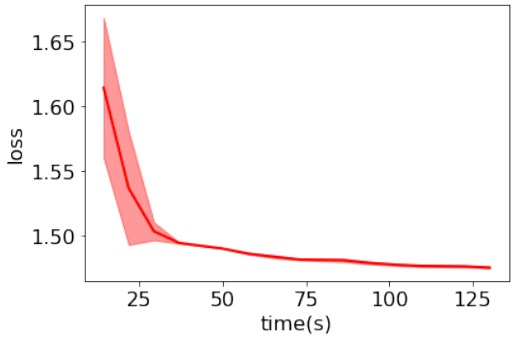

Figure 10: Sensitivity on initial point for MNIST with CNN

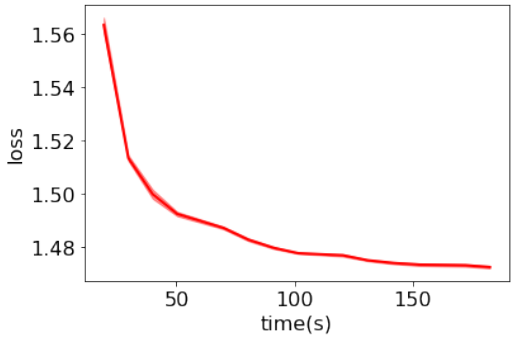

Figure 11: Sensitivity on initial point for MNIST with FFN

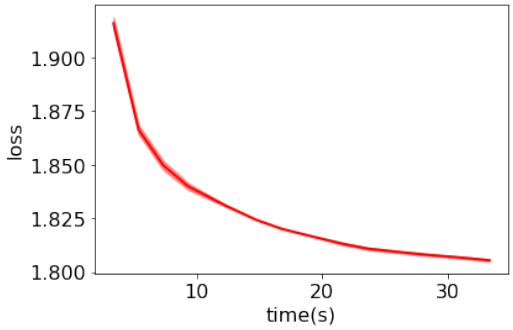

Figure 12: Sensitivity on initial point for Embedded CIFAR-10 with logistic regression

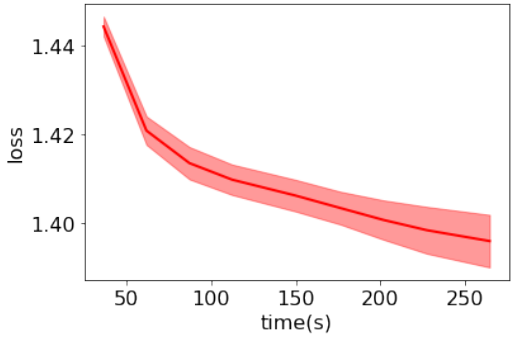

Figure 13: Sensitivity on initial point for CovType with logistic regression

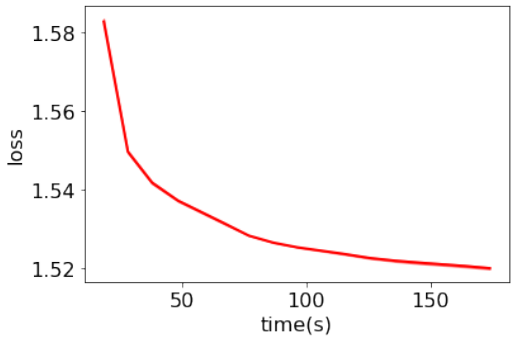

Figure 14: Sensitivity on initial point for MNIST with logistic regression

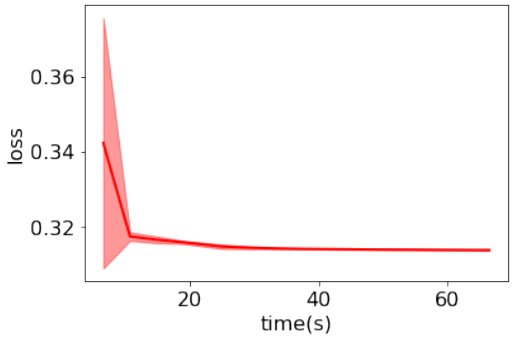

Figure 15: Sensitivity on initial point for NSL-KDD with logistic regression