Supplementary Material

Supplementary material for the paper: "Linear Convergence of Adaptive Stochastic Gradient Descent".

This appendix is organized as follows:

- Appendix A: Proof of Theorem 1 in the Stochastic Setting
- Appendix B: Proof of Theorem 2 and 3 in the batch Setting
- Appendix C: Proof of Lemmas in Stage I
- Appendix D: Proof of Lemmas in Stage II
- Appendix E: More Numerical Experiments

A Proof of Theorem 1 in the Stochastic Setting

From Lemma 1, let $C = \eta L$, after $N \geq \frac{\eta^2 L^2 - b_0^2}{\alpha\gamma\epsilon} + \frac{\delta}{\gamma}$ steps, if $\min_{0\leq i\leq N-1} \|\boldsymbol{x}_i - \boldsymbol{x}^*\|^2 > \epsilon$, then with high probability $1 - \exp(-\frac{\delta^2}{2(N\gamma(1-\gamma)+\delta)})$, $b_N > \eta L$. Then, there exists a first index $k_0 < N$, s.t. $b_{k_0} > \eta L$ but $b_{k_0-1} < \eta L$.
If $k_0 \geq 1$, then

$$\begin{aligned}
\|\boldsymbol{x}_{k_0+l} - \boldsymbol{x}^*\|^2 &= \|\boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*\|^2 + \frac{\eta^2}{b_{k_0+l}^2}\|G_{k_0-1+l}\|^2 - \frac{2\eta}{b_{k_0+l}}\langle \boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*, G_{k_0-1+l}\rangle \\
&\leq \|\boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*\|^2 + (\frac{\eta^2 L}{b_{k_0+l}^2} - \frac{2\eta}{b_{k_0+l}})\langle \boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*, G_{k_0-1+l}\rangle \\
&\leq \|\boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*\|^2 - \frac{\eta}{b_{k_0+l}}\langle \boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*, G_{k_0-1+l}\rangle \\
&\leq \|\boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*\|^2 - \frac{\eta}{b_{\max}}\langle \boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*, G_{k_0-1+l}\rangle
\end{aligned} \tag{1}$$

where the last second inequality is from the condition $b_{k_0} > \eta L$. The last inequality holds since $b_{k_0+l} \leq b_{\max}$ and $f_{\xi_{k_0-1+l}}(\boldsymbol{x})$ is convex, which implies $\langle \boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*, G_{k_0-1+l} - \nabla f_{\xi_{k_0-1+l}}(\boldsymbol{x}^*)\rangle \geq 0$, $\mathbb{P}(\nabla f_{\xi_{k_0-1+l}}(\boldsymbol{x}^*) = \mathbf{0}) = 1$ by Assumption (A4).

Take expectation regarding to ξ_{k_0-1+l}, and use the fact that when $j > k_0$, $b_j > L$, when $l \geq 1$ and $0 < \frac{\mu\eta}{b_{k_0-1+l}} < \frac{\mu}{L} < 1$, then we can get

$$\begin{aligned}
\mathbb{E}_{\xi_{k_0-1+l}}\|\boldsymbol{x}_{k_0+l} - \boldsymbol{x}^*\|^2 &\leq \|\boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*\|^2 - \frac{\eta}{b_{\max}}\langle \boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*, \nabla F(\boldsymbol{x}_{k_0-1+l})\rangle \\
&\leq (1 - \frac{\mu\eta}{b_{\max}})\|\boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*\|^2 \\
&\leq \prod_{j=0}^{l}(1 - \frac{\mu\eta}{b_{\max}})\|\boldsymbol{x}_{k_0-1} - \boldsymbol{x}^*\|^2 \\
&\leq \prod_{j=0}^{l}(1 - \frac{\mu\eta}{b_{\max}})(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 + \eta^2(\log(\frac{C^2}{b_0^2}) + 1) \\
&\leq (\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 + \eta^2(\log(\frac{\eta^2 L^2}{b_0^2}) + 1))\exp(-\frac{\mu l}{b_{\max}})
\end{aligned} \tag{2}$$

where the second inequality is from the strong convexity of $F(\boldsymbol{x})$, i.e. $\langle \boldsymbol{x} - \boldsymbol{y}, \nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\rangle \geq \mu\|\boldsymbol{x} - \boldsymbol{y}\|^2$ and $\nabla F(\boldsymbol{x}^*) = \mathbf{0}$. From Lemma 4, we can give an upper bound for $b_{\max} = \max_{l\geq 0} b_{k_0+l} = C + \frac{L}{\eta}(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 + \eta^2(\log\frac{C^2}{b_0^2} + 1)) = \eta L + \frac{L}{\eta}(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 + \eta^2(\log\frac{\eta^2 L^2}{b_0^2} + 1))$.

Then, take the iterated expectation and use Markov in inequality, with high probability $1-\delta_h$,

$$\|\boldsymbol{x}_{k_0+l}-\boldsymbol{x}^*\|^2 \leq \frac{1}{\delta_h}(\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2+\eta^2(\log\frac{\eta^2L^2}{b_0^2}+1))\exp(-\frac{\mu l}{b_{\max}})$$

Then, after $M \geq \frac{\eta L+\frac{L}{\eta}(\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2+\eta^2(\log\frac{\eta^2L^2}{b_0^2}+1))}{\mu}\log\frac{\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2+\eta^2(\log\frac{\eta^2L^2}{b_0^2}+1)}{\epsilon\delta_h}$ iterations, with high probability more than $1-\delta_h-\exp(-\frac{\delta^2}{2(N\gamma(1-\gamma)+\delta)})$

$$\|\boldsymbol{x}_{k_0+M}-\boldsymbol{x}^*\|^2 \leq \epsilon$$

Otherwise, if $k_0=0$, i.e. $b_0>\eta L$, then we use the same inequality as above,

$$\mathbb{E}_{\xi_{M-1}}\|\boldsymbol{x}_M-\boldsymbol{x}^*\|^2 \leq (1-\frac{\mu}{b'_{\max}})\|\boldsymbol{x}_{M-1}-\boldsymbol{x}^*\|^2 \leq \|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2\exp(-\frac{\mu M}{b'_{\max}})$$

Then, after $M \geq \frac{b'_{\max}}{\mu}\log\frac{\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2}{\epsilon\delta_h}$ iterations, by Markov's inequality,

$$\mathbb{P}(\|\boldsymbol{x}_M-\boldsymbol{x}^*\|^2 \geq \epsilon) \leq \frac{\mathbb{E}\|\boldsymbol{x}_M-\boldsymbol{x}^*\|^2}{\epsilon} \leq \delta_h$$

where $b'_{\max}=b_0+\frac{L}{\eta}\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2$ is derived as follows:

$$\begin{aligned}\|\boldsymbol{x}_{j+1}-\boldsymbol{x}^*\|^2 &\leq \|\boldsymbol{x}_j-\boldsymbol{x}^*\|^2-\frac{\eta}{L}\frac{\|G_j\|^2}{b_{j+1}}\\ &\leq \|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2-\frac{\eta}{L}\sum_{i=0}^{j}\frac{\|G_i\|^2}{b_{i+1}}\end{aligned}$$

Then, for any $j+1$:

$$b_{j+1}=b_0+\sum_{i=0}^{j}\frac{\|G_i\|^2}{b_i+b_{i+1}} \leq b_0+\frac{L}{\eta}(\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2-\|\boldsymbol{x}_{j+1}-\boldsymbol{x}^*\|^2) \tag{3}$$

Plugging in the value, we can get $M \geq \frac{b_0+\frac{L}{\eta}\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2}{\mu}\log\frac{\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2}{\epsilon\delta_h}$.

B Proof of Theorem 2 and 3 in the batch Setting

B.1 Proof of Theorem 2

Lemma B1. ***(Co-coercivity with Strong Convexity)*** *(Bubeck et al., 2015) If $F(\boldsymbol{x})$ is $\mu-$strongly convex and $L-$smooth, then*

$$\langle\nabla F(\boldsymbol{x})-\nabla F(\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{y}\rangle \geq \frac{\mu L}{\mu+L}\|\boldsymbol{x}-\boldsymbol{y}\|^2+\frac{1}{\mu+L}\|\nabla F(\boldsymbol{x})-\nabla F(\boldsymbol{y})\|^2$$

Proof. Let $\phi(\boldsymbol{x})=F(\boldsymbol{x})-\frac{\mu}{2}\|\boldsymbol{x}\|^2$, then $\phi(\boldsymbol{x})$ is convex and $(L-\mu)-$smooth. By Lemma C4,

$$\langle\nabla\phi(\boldsymbol{x})-\nabla\phi(\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{y}\rangle \geq \frac{1}{L-\mu}\|\nabla\phi(\boldsymbol{x})-\nabla\phi(\boldsymbol{y})\|^2$$

Plugging in $\nabla\phi(\boldsymbol{x})=\nabla F(\boldsymbol{x})-\mu\boldsymbol{x}$, we have

$$\begin{aligned}\langle\nabla F(\boldsymbol{x})-\nabla F(\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{y}\rangle-\mu\|\boldsymbol{x}-\boldsymbol{y}\|^2 \geq \frac{1}{L-\mu}&(\|\nabla F(\boldsymbol{x})-\nabla F(\boldsymbol{y})\|^2+\mu^2\|\boldsymbol{x}-\boldsymbol{y}\|^2\\ &-2\mu\langle\nabla F(\boldsymbol{x})-\nabla F(\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{y}\rangle)\end{aligned}$$

With simple algebra, we can get the result. □

By Lemma 2, after $N = \left\lceil \frac{\log(\eta^2(\mu+L)^2/4b_0^2)}{\log(1+4\mu^2\epsilon/(\mu+L)^2)} \right\rceil + 1$ iterations, if $\min_{0\le i\le N-1} \|\boldsymbol{x}_i - \boldsymbol{x}^*\|^2 > \epsilon$, then $\exists k_0 \le N$, such that k_0 is the first index s.t. $b_{k_0} > \eta\frac{\mu+L}{2}$.

If $k_0 > 1$, since $F(\boldsymbol{x})$ is μ-strongly convex and $L-$smooth, by Lemma B1,

$$\langle \nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \ge \frac{\mu L}{\mu+L}\|\boldsymbol{x}-\boldsymbol{y}\|^2 + \frac{1}{\mu+L}\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\|^2.$$

For $j \ge 0$, we have $0 < \eta\frac{2\mu L}{\mu+L}\frac{1}{b_{k_0+j}} < \frac{4\mu L}{(\mu+L)^2} < 1$, since $2\mu L < \mu^2 + L^2$.

We divide the analysis into two situations to get a better bound instead of using $b_{\max}$ for all the following steps, which is different from the proof of Theorem 1. First, assume that $\eta\frac{\mu+L}{2} < b_{k_0} < \eta L$ and after another l iterations, b_{k_0+l} is still less than ηL, then $\|\boldsymbol{x}_{k_0+l} - \boldsymbol{x}^*\|^2$ is bounded as follows:

$$\begin{aligned}
\|\boldsymbol{x}_{k_0+l} - \boldsymbol{x}^*\|^2 &= \|\boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*\|^2 + \frac{\eta^2}{b_{k_0+l}^2}\|\nabla F(\boldsymbol{x}_{k_0-1+l})\|^2 \\
&\quad - \frac{2\eta}{b_{k_0+l}}\langle \boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*, \nabla F(\boldsymbol{x}_{k_0-1+l})\rangle \\
&\le (1 - \frac{2\mu\eta L}{(\mu+L)b_{k_0+l}})\|\boldsymbol{x}_{k_0-1+l} - \boldsymbol{x}^*\|^2 \\
&\quad + \frac{\eta}{b_{k_0+l}}(\frac{\eta}{b_{k_0+l}} - \frac{2}{\mu+L})\|\nabla F(\boldsymbol{x}_{k_0-1+l})\|^2 \\
&\le \prod_{j=0}^{l}(1 - \frac{2\mu\eta L}{(\mu+L)b_{k_0+j}})\|\boldsymbol{x}_{k_0-1} - \boldsymbol{x}^*\|^2 \\
&\le \exp(-\sum_{j=0}^{l}\frac{2\mu\eta L}{(\mu+L)\eta L})\|\boldsymbol{x}_{k_0-1} - \boldsymbol{x}^*\|^2 \\
&\le \exp(-\frac{2\mu(l+1)}{\mu+L})\|\boldsymbol{x}_{k_0-1} - \boldsymbol{x}^*\|^2
\end{aligned}$$

where $\|\boldsymbol{x}_{k_0-1} - \boldsymbol{x}^*\|^2$ can be upper bounded according to Lemma 3 with $C = \frac{\mu+L}{2}$:

$$\|\boldsymbol{x}_{k_0-1} - \boldsymbol{x}^*\|^2 \le \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 + \eta^2(\log(\frac{(\mu+L)^2}{4b_0^2}) + 1)$$

Second, if $b_{k_0+M_0} > \eta L$, M_0 can be $0, 1, 2, \ldots$, then for $l \ge 0$,

$$\begin{aligned}
&\|\boldsymbol{x}_{k_0+M_0+l} - \boldsymbol{x}^*\|^2 = \|\boldsymbol{x}_{k_0+M_0-1+l} - \boldsymbol{x}^*\|^2 + \frac{\eta^2}{b_{k_0+l}^2}\|\nabla F(\boldsymbol{x}_{k_0+M_0-1+l})\|^2 \\
&- \frac{2\eta}{b_{k_0+M_0+l}}\langle \boldsymbol{x}_{k_0+M_0-1+l} - \boldsymbol{x}^*, \nabla F(\boldsymbol{x}_{k_0+M_0-1+l})\rangle \\
&\le \|\boldsymbol{x}_{k_0+M_0-1+l} - \boldsymbol{x}^*\|^2 + (\frac{\eta^2 L}{b_{k_0+M_0+l}^2} - \frac{2\eta}{b_{k_0+M_0+l}})\langle \boldsymbol{x}_{k_0+M_0-1+l} - \boldsymbol{x}^*, \nabla F(\boldsymbol{x}_{k_0+M_0-1+l})\rangle \\
&\le (1 - \frac{\mu\eta}{b_{\max}})\|\boldsymbol{x}_{k_0+M_0+l} - \boldsymbol{x}^*\|^2 \\
&\le \exp(-\frac{\mu\eta l}{b_{\max}})\|\boldsymbol{x}_{k_0+M_0} - \boldsymbol{x}^*\|^2 \\
&\le \exp(-\frac{\mu\eta l}{b_{\max}})\exp(-\frac{2\mu(M_0+1)}{\mu+L})\|\boldsymbol{x}_{k_0-1} - \boldsymbol{x}^*\|^2
\end{aligned}$$

where $b_{\max}$ can be upper bounded according to Lemma 4, here $b_{\max} \le \eta L + \frac{L}{\eta}\|\boldsymbol{x}_{k_0+M_0} - \boldsymbol{x}^*\|^2$.

Once $b_t > \frac{\mu+L}{2} > \frac{L}{2}$, by Lemma 5, AdaGrad-Norm is indeed a decent algorithm for $\|\boldsymbol{x}_j - \boldsymbol{x}^*\|^2$, so $\|\boldsymbol{x}_{k_0+M_0-1} - \boldsymbol{x}^*\|^2 \le \|\boldsymbol{x}_{k_0-1} - \boldsymbol{x}^*\|^2$. Hence,

$$b_{\max} \le \eta L + \frac{L}{\eta}\|\boldsymbol{x}_{k_0-1} - \boldsymbol{x}^*\|^2$$

Combining the two situations above, we have

$$\begin{aligned}\|\boldsymbol{x}_{k_0+M}-\boldsymbol{x}^*\|^2 &\le \exp(-M\min\left\{\frac{\mu\eta}{b_{\max}},\frac{2\mu}{\mu+L}\right\})\|\boldsymbol{x}_{k_0-1}-\boldsymbol{x}^*\|^2\\ &\le \exp(-M\min\left\{\frac{\mu}{L(1+\Delta/\eta^2)},\frac{2\mu}{\mu+L}\right\})\|\boldsymbol{x}_{k_0-1}-\boldsymbol{x}^*\|^2\end{aligned}$$

where $\Delta=\|\boldsymbol{x}_{k_0-1}-\boldsymbol{x}^*\|^2$.

After $M\ge\max\left\{\frac{L(1+\Delta/\eta^2)}{\mu},\frac{\mu+L}{2\mu}\right\}\log\frac{\Delta}{\epsilon}-1$ iterations,

$$\|\boldsymbol{x}_{k_0+M}-\boldsymbol{x}^*\|^2\le\epsilon$$

Otherwise, if $k_0=1$, then

$$\|\boldsymbol{x}_M-\boldsymbol{x}^*\|^2\le\exp(-\sum_{j=1}^{M}\min\left\{\frac{\mu\eta}{b'_{\max}},\frac{2\mu}{\mu+L}\right\})\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2$$

where $b'_{\max}=\eta L+\frac{L}{\eta}\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2$.

Then, after $M\ge\max\left\{\frac{L(1+\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2/\eta^2)}{\mu},\frac{\mu+L}{2\mu}\right\}\log\frac{\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2}{\epsilon}$ iterations, we can assure that

$$\|\boldsymbol{x}_M-\boldsymbol{x}^*\|^2\le\epsilon$$

B.2 Proof of Theorem 3

By Lemma 2, after $N\ge\frac{\log(\eta^2L^2/b_0^2)}{\log(1+2\mu\epsilon/(\eta L)^2)}$ iterations, if $\min_{0\le i\le N-1}F(\boldsymbol{x}_i)-F^*>\epsilon$, then $\exists k_0\le N$, such that k_0 is the first index s.t. $b_{k_0}>\eta L$.

If $k_0>1$, then for $j\ge0$, from Assumption (A2), we have

$$\begin{aligned}F(\boldsymbol{x}_{k_0+j}) &\le F(\boldsymbol{x}_{k_0+j-1})-\frac{\eta}{b_{k_0+j}}(1-\frac{\eta L}{2b_{k_0+j}})\|\nabla F(\boldsymbol{x}_{k_0+j-1})\|^2\\ &\le F(\boldsymbol{x}_{k_0+j-1})-\frac{\eta}{2b_{k_0+j}}\|\nabla F(\boldsymbol{x}_{k_0+j-1})\|^2 \qquad (4)\\ &\le F(\boldsymbol{x}_{k_0+j-1})+\frac{\mu\eta}{b_{k_0+j}}(F^*-F(\boldsymbol{x}_{k_0+j-1}))\end{aligned}$$

The last inequality is from $\mu-$PL inequality (Assumption (A1b)): $-\|F(\boldsymbol{x})\|^2\le2\mu(F*-F(\boldsymbol{x}_j)),\forall x$. Then, add $-F^*$ on both sides, we can get

$$F(\boldsymbol{x}_{k_0+j})-F^*\le(1-\frac{\mu\eta}{b_{k_0+j}})(F(\boldsymbol{x}_{k_0+j-1})-F^*) \qquad (5)$$

Since $b_{k_0+j}>\eta L\ge\eta\mu$, $1-\frac{\mu\eta}{b_{k_0+j}}\in(0,1)$ holds for all $j\ge0$, it is a contraction at every step. Then,

$$\begin{aligned}F(\boldsymbol{x}_{k_0+j})-F^* &\le(\prod_{l=0}^{j}(1-\frac{\mu\eta}{b_{k_0+l}}))(F(\boldsymbol{x}_{k_0-1})-F^*)\\ &\le\exp(-\sum_{l=0}^{j}\frac{\mu\eta}{b_{k_0+l}})(F(\boldsymbol{x}_{k_0-1})-F^*) \qquad (6)\\ &\le\exp(-\sum_{l=0}^{j}\frac{\mu\eta}{b_{k_0+l}})(F(\boldsymbol{x}_0)-F^*+\frac{\eta^2L}{2}(1+\log(\frac{b_{k_0-1}^2}{b_0^2}))\end{aligned}$$

where we use the fact that $1-x\le e^{-x},\forall x\in(0,1)$ and the lemma in Ward et al. (2018): $F(\boldsymbol{x}_{k_0-1})\le F(\boldsymbol{x}_0)+\frac{\eta^2L}{2}(1+\log(\frac{b_{k_0-1}^2}{b_0^2}))$.

The upper bound of b_j is also from Ward et al. (2018):

$$b_{\max}=b_{k_0-1}+\frac{2}{\eta}(F_{k_0-1}-F^*)\le\eta L+\frac{2}{\eta}(F(\boldsymbol{x}_0)-F^*+\frac{\eta^2L}{2}(1+\log(\frac{\eta^2L^2}{b_0^2}))) \qquad (7)$$

Then,

$$F(\boldsymbol{x}_{k_0+M-1}) - F^* \leq \exp(-\frac{\mu\eta M}{b_{\max}})(F(\boldsymbol{x}_0) - F^* + \frac{\eta^2 L}{2}(1 + 2\log\frac{\eta L}{b_0}))$$

Hence, we need

$$M \geq \frac{b_{\max}}{\mu\eta}\log\frac{F(\boldsymbol{x}_0) - F^* + \frac{\eta^2 L}{2}(1 + 2\log\frac{\eta L}{b_0})}{\epsilon}$$

It is sufficient that

$$M \geq \frac{\eta L + \frac{2}{\eta}(F(\boldsymbol{x}_0) - F^* + \frac{\eta^2 L}{2}(1 + \log\frac{\eta^2 L^2}{b_0^2}))}{\mu\eta}\log\frac{F(\boldsymbol{x}_0) - F^* + \frac{\eta^2 L}{2}(1 + 2\log\frac{\eta L}{b_0})}{\epsilon}$$

Then,

$$\min_{0\leq i\leq N+M-1} F(\boldsymbol{x}_i) - F^* \leq \epsilon$$

where $N = \left\lceil \frac{\log(\eta^2 L^2/b_0^2)}{\log(1+2\mu\epsilon/(\eta L)^2)} \right\rceil + 1$.

Otherwise, if $k_0 = 1$, the upper bound of b_j degenerates to

$$b'_{\max} = b_0 + \frac{2}{\eta}(F(\boldsymbol{x}_0) - F^*)$$

Then, using the same procedure, we have

$$\begin{aligned} F(\boldsymbol{x}_M) - F^* &\leq \exp(-\sum_{k=0}^{M-1}\frac{\mu\eta}{b_{k+1}})(F(\boldsymbol{x}_0) - F^*) \\ &\leq \exp(\frac{-\mu\eta M}{b_{\max}})(F(\boldsymbol{x}_0) - F^*) \end{aligned} \tag{8}$$

Once the number of iterations satisfies

$$M \geq \frac{b'_{\max}}{\mu\eta}\log\frac{F(\boldsymbol{x}_0) - F^*}{\epsilon} = \frac{b_0 + \frac{2}{\eta}(F(\boldsymbol{x}_0) - F^*)}{\mu\eta}\log\frac{F(\boldsymbol{x}_0) - F^*}{\epsilon}$$

we can get the expected result: $F(\boldsymbol{x}_M) - F^* \leq \epsilon$.

C Proof of Lemmas in Stage I

C.1 Proof of Lemma 1

Lemma C2. ***(Bernstein's Inequality)*** *(Wainwright, 2019) Let X be a random variable, $\mathbb{E}[X] = \mu$, $Var(X) = \sigma^2$, if X satisfies Bernstein condition with parameter $b > 0$, i.e. if $|\mathbb{E}(X-\mu)^k| \leq \frac{1}{2}k!\sigma^2 b^{k-2}, \forall k \geq 2$, then*

$$\mathbb{P}(|X - \mu| \geq t) \leq 2\exp(-\frac{t^2}{2(\sigma^2 + bt)})$$

Lemma C3. *(Wainwright, 2019) Let $X_i \sim$ Bernoulli$(p), i.i.d. \forall i = 1, 2, \ldots n$, and $X = \sum_{i=1}^n X_i$. Since $X_i \in [0, 1]$, $\{X_i\}$ satisfy Bernstein condition, then*

$$P(|X - np| > t) \leq 2\exp(-\frac{t^2}{2(np(1-p) + t)})$$

Proof of Lemma 1 If $\min_j \|\boldsymbol{x}_j - \boldsymbol{x}^*\|^2 \leq \epsilon$, we are done.
Otherwise, we have $\|\boldsymbol{x}_j - \boldsymbol{x}^*\|^2 > \epsilon, \forall j = 0, 1, 2, \ldots, N$. Assume that $F(\boldsymbol{x})$ satisfies $(\epsilon, \alpha, \gamma)-$ RUIG (Assumption (A3)), we can use independent identical Bernoulli random variables $\{Z_j\}$ to represent them with the following distribution:

$$Z_j = \begin{cases} 1 & if\ \|\nabla f_{\xi_j}(\boldsymbol{x}_j)\|^2 \geq \alpha\|\boldsymbol{x}_j - \boldsymbol{x}^*\|^2 \\ 0 & else \end{cases} \tag{9}$$

where $\mathbb{P}(Z_j = 1) = \gamma, \forall j$. Note that the RUIG assumption is for any fixed x (conditional on $\boldsymbol{x}$), the probability distribution is over the random variable i (or ξ_i) (but not over $\boldsymbol{x}$). Every index ξ_i is sampled independently and uniformly at each iteration, so random variables $\{Z_i\}$ are independent. Then, from Lemma C3 and let $Z = \sum_j Z_j$, with high probability bigger than $1-\exp(-\frac{\delta^2}{2(N\gamma(1-\gamma)+\delta)})$, $Z \geq \gamma N - \delta, \forall N$. Thus, after $N \geq \frac{C^2-b_0^2}{\alpha\gamma\epsilon} + \frac{\delta}{\gamma}$ iterations, with $1 - \exp(-\frac{\delta^2}{2(N\gamma(1-\gamma)+\delta)})$, we have

$$b_N^2 = b_0^2 + \sum_{i=0}^{N-1} \|\nabla f_{\xi_i}(\boldsymbol{x}_i)\|^2 > b_0^2 + (\gamma N - \delta)\alpha\epsilon \geq C^2$$

Note that even if in the case that there is some correlation between Bernoulli random variables, since each of them is sub-Gaussian with $\sigma = 0.5$, then the upper bound of the sub-Gaussian parameter of the sum of them is $0.5N$, so the worst-case variance is $0.25N^2$. Hence, the result still holds under this setting.

C.2 Proof of Lemma 2

(a) If $b_0 > C$, we are done.
Otherwise if $b_0 < C$, and after $N \geq \frac{\log(C^2/b_0^2)}{\log(1+\mu^2\epsilon/C^2)}$ iterations, $b_N < C$ and $\min_{0\leq i\leq N-1} \|\boldsymbol{x}_i - \boldsymbol{x}^*\|^2 > \epsilon$. Since $F(\boldsymbol{x})$ is $\mu-$ strongly convex, $\epsilon < \|\boldsymbol{x}_i - \boldsymbol{x}^*\|^2 \leq \frac{1}{\mu^2}\|\nabla F(\boldsymbol{x}_i) - \nabla F(\boldsymbol{x}^*)\|^2, \forall \boldsymbol{x}_i$ and $\nabla F(\boldsymbol{x}^*) = \mathbf{0}$. Then,

$$\begin{aligned} b_N^2 &= b_{N-1}^2 + \|\nabla F(\boldsymbol{x}_{N-1})\|^2 \\ &= b_{N-1}^2(1 + \frac{\|\nabla F(\boldsymbol{x}_{N-1})\|^2}{b_{N-1}^2}) \\ &\geq b_0^2 \prod_{j=0}^{N-1}(1 + \frac{\|\nabla F(\boldsymbol{x}_j)\|^2}{b_j^2}) \\ &\geq b_0^2(1 + \frac{\mu^2\epsilon}{C^2})^N \geq C^2 \end{aligned} \tag{10}$$

Contradiction! Hence, at least one of $\min_{0\leq i\leq N-1} \|\boldsymbol{x}_i - \boldsymbol{x}^*\|^2 \leq \epsilon$ or $b_N > C$ holds. When μ is small and C is big, we have $\log(1 + \frac{\mu^2\epsilon}{C^2}) \approx \frac{\mu^2\epsilon}{C^2}$.

(b) With PL inequality $\frac{1}{2\mu}\|\nabla F(\boldsymbol{x})\|^2 \geq F(\boldsymbol{x}) - F(\boldsymbol{x}^*)$ instead of μ-strongly convex assumption, if $\min_{0\leq i\leq N-1} F(\boldsymbol{x}_i) - F^* > \epsilon$ and $b_N < C$, then after $N \geq \frac{\log(C^2/b_0^2)}{\log(1+2\mu\epsilon/C^2)}$ iterations, $b_N^2 \geq b_0^2(1 + \frac{2\mu\epsilon}{C^2})^N \geq C^2$, contradiction! Hence, either $\min_{0\leq i\leq N-1} F(\boldsymbol{x}_i) - F^* \leq \epsilon$ or $b_N > C$.

C.3 Proof of Lemma 3

Lemma C4. ***(Co-coercivity)*** *(Needell et al., 2016) For a $L-$smooth convex function $F(\boldsymbol{x})$, $\forall \boldsymbol{x}, \boldsymbol{y}$*

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\|^2 \leq L\langle \boldsymbol{x} - \boldsymbol{y}, \nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\rangle$$

Lemma C5. ***(Integral lemma)*** *(Ward et al., 2018) For any non-negative sequence $a_1, \ldots, a_T$, such that $a_1 \geq 1$,*

$$\sum_{l=1}^{T} \frac{a_l}{\sum_{i=1}^{l} a_i} \leq \log(\sum_{i=1}^{T} a_i) + 1 \tag{11}$$

$$\sum_{l=1}^{T} \frac{a_l}{\sqrt{\sum_{i=1}^{l} a_i}} \leq 2\sqrt{\sum_{i=1}^{T} a_i} \tag{12}$$

Proof. The lemma can be proved by induction. Besides, we can take above sums as Riemman sums, then the sums should be proportional to integrals, $\log(x)$ and $2\sqrt{x}$, respectively. □

Proof of Lemma 3 With above two lemmas, we can bound $\|\boldsymbol{x}_{J-1} - \boldsymbol{x}^*\|^2$ as follows:

$$\begin{aligned}
\|\boldsymbol{x}_{J-1} - \boldsymbol{x}^*\|^2 &= \|\boldsymbol{x}_{J-2} - \frac{\eta G_{J-2}}{b_{J-1}} - \boldsymbol{x}^*\|^2 \\
&= \|\boldsymbol{x}_{J-2} - \boldsymbol{x}^*\|^2 + \|\frac{\eta G_{J-2}}{b_{J-1}}\|^2 - 2\eta\langle\frac{G_{J-2}}{b_{J-1}}, \boldsymbol{x}_{J-2} - \boldsymbol{x}^*\rangle \\
&\leq \|\boldsymbol{x}_{J-2} - \boldsymbol{x}^*\|^2 + \|\frac{\eta G_{J-2}}{b_{J-1}}\|^2 - \frac{2\eta}{b_{J-1}L}\|G_{J-2} - \nabla f_{\xi_{J-2}}(\boldsymbol{x}^*)\|^2 \\
&\leq \|\boldsymbol{x}_{J-2} - \boldsymbol{x}^*\|^2 + \frac{\eta^2\|G_{J-2}\|^2}{b_{J-1}^2} \\
&\leq \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 + \eta^2\sum_{j=0}^{J-2}\frac{\|G_j\|^2}{b_{j+1}^2} \\
&\leq \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 + \eta^2\sum_{j=0}^{J-2}\frac{\|G_j\|^2/b_0^2}{\sum_{l=0}^{j}\|G_l\|^2/b_0^2} \\
&\leq \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 + \eta^2(\log(\sum_{j=0}^{J-2}\|G_j\|^2/b_0^2) + 1) \\
&\leq \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 + \eta^2(\log\frac{C^2}{b_0^2} + 1)
\end{aligned}$$

where the first inequality is from the co-coercivity (Lemma C4) and Assumption (A4) $\mathbb{P}(\nabla f_{\xi_{J-2}}(\boldsymbol{x}^*) = \mathbf{0}) = 1$; last second inequality is from lemma C5 and the last inequality is from the assumption that J is the first index s.t. $b_J > C$.

D Proof of Lemmas in Stage II

D.1 Proof of Lemma 4

Since $b_J > \eta L$, we have the following bound for $\|\boldsymbol{x}_{J+l} - \boldsymbol{x}^*\|^2$:

$$\begin{aligned}
\|\boldsymbol{x}_{J+l} - \boldsymbol{x}^*\|^2 &= \|\boldsymbol{x}_{J+l-1} - \boldsymbol{x}^*\|^2 + \frac{\eta^2\|G_{J+l-1}\|^2}{b_{J+l}^2} \\
&\quad - \frac{2\eta}{b_{J+l}}\langle G_{J+l-1} - \nabla f_{\xi_{J+l-1}}(\boldsymbol{x}^*), \boldsymbol{x}_{J+l-1} - \boldsymbol{x}^*\rangle \\
&\leq \|\boldsymbol{x}_{J+l-1} - \boldsymbol{x}^*\|^2 + \|\frac{\eta G_{J+l-1}}{b_{J+l}}\|^2 - \frac{2\eta}{b_{J+l}L}\|G_{J+l-1} - \nabla f_{J+l-1}(\boldsymbol{x}^*)\|^2 \\
&\leq \|\boldsymbol{x}_{J+l-1} - \boldsymbol{x}^*\|^2 + \frac{\eta}{b_{J+l}}(\frac{\eta}{b_{J+l}} - \frac{2}{L})\|G_{J+l-1}\|^2 \\
&\leq \|\boldsymbol{x}_{J+l-1} - \boldsymbol{x}^*\|^2 - \frac{\eta}{L}\frac{\|G_{J+l-1}\|^2}{b_{J+l}} \\
&\leq \|\boldsymbol{x}_{J-1} - \boldsymbol{x}^*\|^2 - \frac{\eta}{L}\sum_{j=0}^{l}\frac{\|G_{J+j-1}\|^2}{b_{J+j}}
\end{aligned} \tag{13}$$

inequalities are from $f_{\xi_{J+l-1}}(\boldsymbol{x})$ is L-smooth (Assumption (A2)) and co-coercivity (Lemma C4). Then, we have the bound of the sum:

$$\sum_{j=0}^{l}\frac{\|G_{J+j-1}\|^2}{b_{J+j}} \leq \frac{L}{\eta}(\|\boldsymbol{x}_{J-1} - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{J+l} - \boldsymbol{x}^*\|^2) \tag{14}$$

Therefore, $b_{\max}$ is bounded as follows:

$$\begin{aligned} b_{J+l} &= b_{J+l-1} + \frac{\|G_{J+l-1}\|^2}{b_{J+l}+b_{J+l-1}} \\ &\leq b_{J-1} + \sum_{j=1}^{l} \frac{\|G_{J+j-1}\|^2}{b_{J+j}} \\ &\leq C + \frac{L}{\eta}\|\boldsymbol{x}_{J-1} - \boldsymbol{x}^*\|^2 \\ &= C + \frac{L}{\eta}(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 + \eta^2(\log\frac{C^2}{b_0^2}+1)) \end{aligned} \tag{15}$$

where $\|\boldsymbol{x}_{J-1} - \boldsymbol{x}^*\|^2 \leq \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 + \eta^2(\log\frac{C^2}{b_0^2}+1)$ is from Lemma 3.

D.2 Proof of Lemma 5

Use similar technique as above,

$$\begin{aligned} \|\boldsymbol{x}_j - \boldsymbol{x}^*\|^2 &\leq \|\boldsymbol{x}_{j-1} - \boldsymbol{x}^*\|^2 + (\frac{\eta^2}{b_j^2} - \frac{2\eta}{b_j L})\|G_{j-1}\|^2 \\ &= \|\boldsymbol{x}_{j-1} - \boldsymbol{x}^*\|^2 - \frac{\eta}{b_j}(\frac{2}{L} - \frac{\eta}{b_j})\|G_{j-1}\|^2 \leq \|\boldsymbol{x}_{j-1} - \boldsymbol{x}^*\|^2 \end{aligned}$$

where the first inequality if from $f_j(\boldsymbol{x})$ is L-smooth (Assumption (A2)), $\nabla f_{j-1}(\boldsymbol{x}^*) = \mathbf{0}$ (Assumption (A4)) and Lemma C4. Therefore, once $b_j > \eta L/2$, AdaGrad-Norm is a descent algorithm.

E More Numerical Experiments

E.1 Numerical Experiments of AdaGrad-Norm with Extreme Initialization

In this section, we demonstrates the numerical experiments of AdaGrad-Norm with $\boldsymbol{x}_0 = \mathbf{0}$ (stochastic: Figure 6; batch: Figure 9) and the extreme case (stochastic: Figure 7; batch: Figure 10): $\boldsymbol{x}_0$ is far away from $\boldsymbol{x}^*$ and $\|\boldsymbol{x}_0\|$ is large. Then, we tune the hyperparameter η in the extreme case with $\eta = \Theta(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2)$ (stochastic: Figure 8) and $\eta = \Theta(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|)$ (batch: Figure 11). In these figures, the x-axis represents iteration t while y-axis is the approximation error $\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2$ in log scale for the first and third columns and it is b_t for the second and fourth columns.

We show that when starting from $\boldsymbol{x}_0 = \mathbf{0}$, the result is close to the experiment we show in Figure 2. When initialize $\boldsymbol{x}_0$ with extremely bad one, $\boldsymbol{x}_0 = 100 * \boldsymbol{w}_0$, where $\boldsymbol{w}_0$ is a randomly generated vector $\boldsymbol{w}_0$ and $\boldsymbol{w}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, AdaGrad-Norm takes much more iterations than before. However, after tuning $\eta = 10000$ in stochastic setting and $\eta = 100$ in batch setting, the convergence rate of AdaGrad-Norm is better again. In this case, b_0 plays a small role.

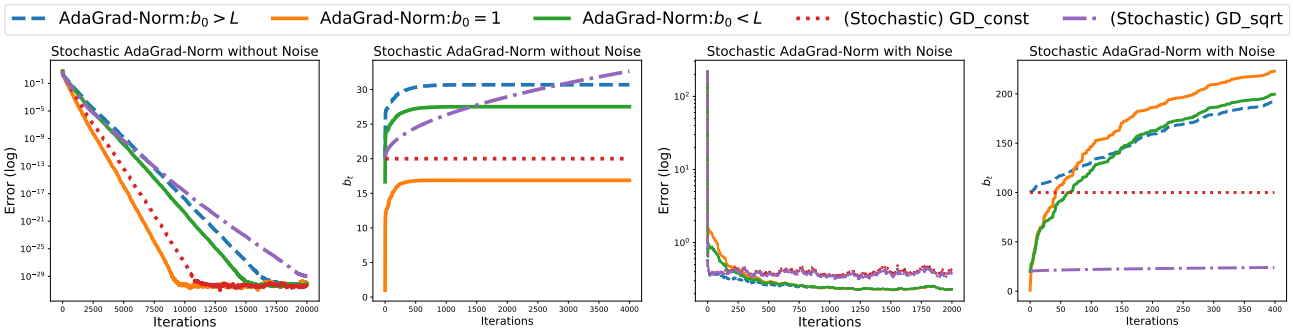

Figure 6: Error and growth of b_t with $\boldsymbol{x}_0 = \mathbf{0}$ and $\eta = 1$ in stochastic setting.

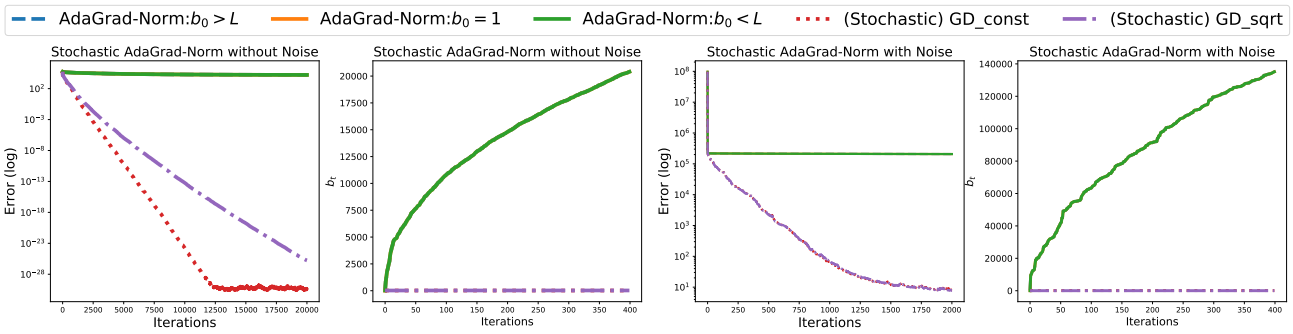

Figure 7: Error and growth of b_t with extremely bad initialization and $\eta = 1$ in stochastic setting.

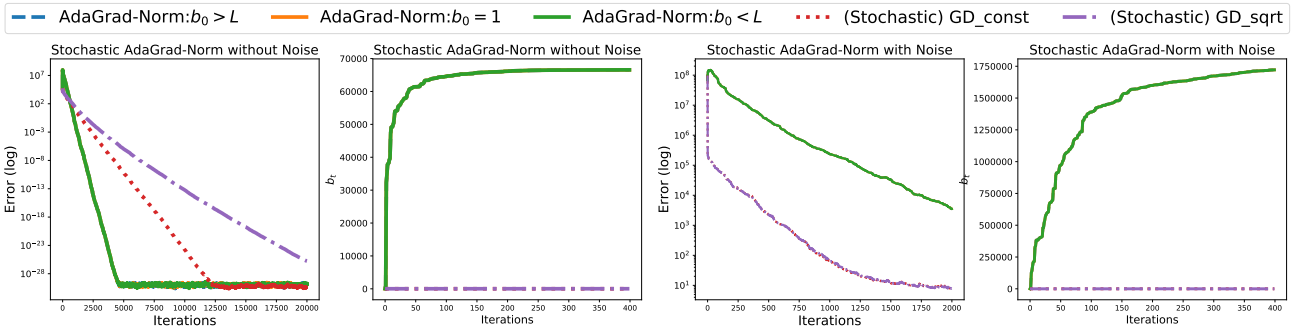

Figure 8: Error and growth of b_t with extremely bad initialization and tuning $\eta = \Theta(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2)$ in stochastic setting.

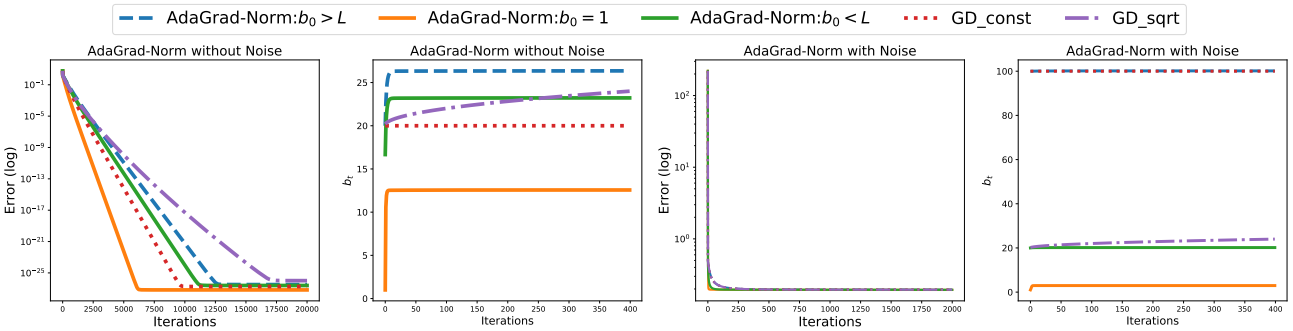

Figure 9: Error and growth of b_t with $\boldsymbol{x}_0 = \mathbf{0}$ and $\eta = 1$ in batch setting.

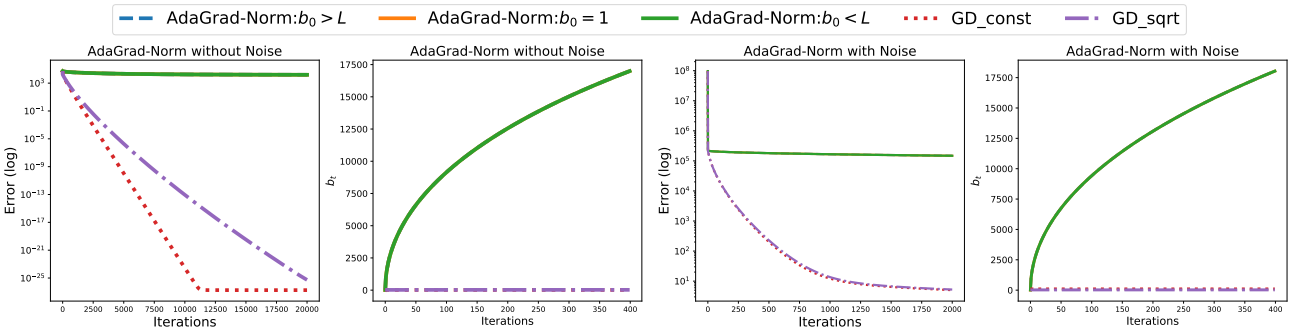

Figure 10: Error and growth of b_t with extremely bad initialization and $\eta = 1$ in batch setting.

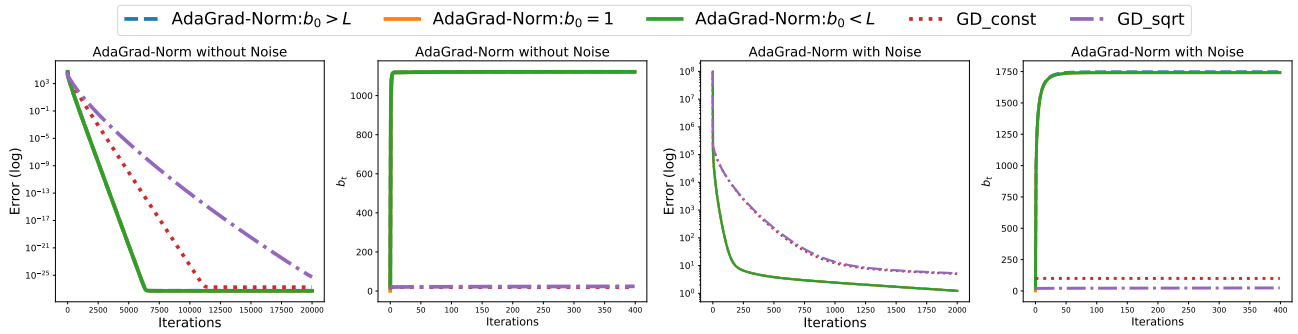

Figure 11: Error and growth of b_t with extremely bad initialization and tuning $\eta = \Theta(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|)$ in batch setting.

E.2 Numerical Experiment of Two Layer Neural Networks

We implement AdaGrad-Norm in a two-layer network. The experiment is mainly to show the stochastic AdaGrad-Norm (black curve) converges with a linear rate. We first define loss function as in Du et al. (2019):

$$L(\boldsymbol{W}) = \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{1}{\sqrt{m}}\sum_{r=1}^{m} a_r\sigma(\langle \boldsymbol{w}_r(k), \boldsymbol{x}_i\rangle) - \boldsymbol{y}_i\right)^2$$

where σ is a ReLU activation function; n is size of data; m is the width for the one-hidden layer. For our implementation, we set $n = 100$, $m = 200$ and $d = 10$. Set mini-batch size 20 for each iteration and the effective stepsize of AdaGrad-Norm with $100/b_t$ and $b_0 = 0.1$. We also run vanilla SGD (blue curve) with $\eta_t = 100$. For details, see the code here [4]. Figure 12 (left) clearly illustrates that AdaGrad-Norm (black curve) converges linearly. Figure 12 (right) shows the norm of the gradients at the first few iterations by AdaGrad-Norm are often big enough to accumulate to exceed ηL, which empirically verifies Assumption (A3).

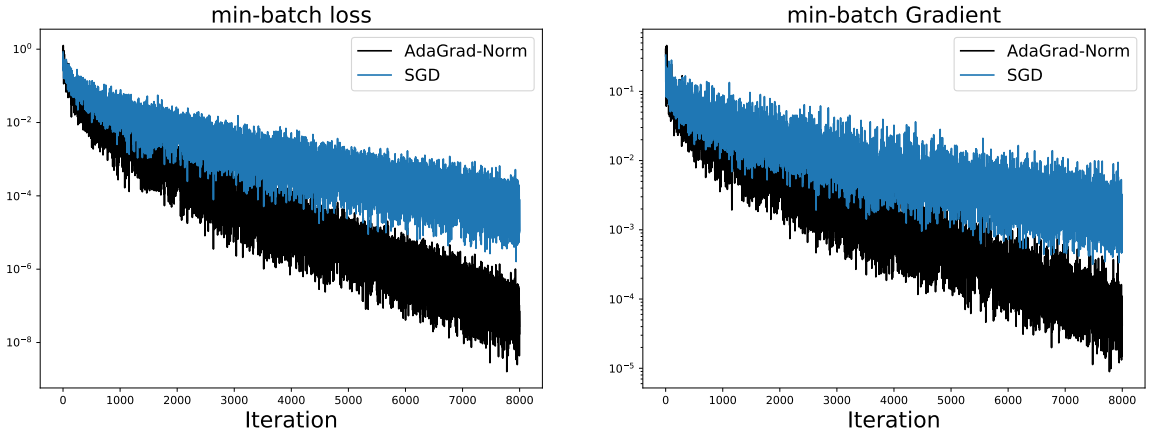

Figure 12: Error and the norm of gradient in a two-layer neural network

[4] https://colab.research.google.com/drive/1kv-XwUxvSogVfNyTO2w1aAoqlS2chlYH