AutoDrop: Training Deep Learning Models with Automatic Learning Rate Drop (Supplementary Material)

A PROOFS

A.1 PROOF FOR THEOREM 1

Proof for Theorem 1. First note that if the learning rate is chosen as specified, then each of the trajectories is a contraction map. By Banach's fixed point theorem, they each have a unique fixed point. Clearly

$$\mathbb{E}^*_{SGD} = \lim_{t\to\infty} \mathbb{E}[x_t] = 0.$$

For the variance we can solve for the fixed points directly. Define $\mathbb{V}^*_{SGD} = \lim_{t\to\infty} \mathbb{V}[x_t]$,

$$\begin{aligned}
&\mathbb{V}^*_{SGD} = (I-\gamma A)^2 \mathbb{V}^*_{SGD} + \gamma A^2 \Sigma, \\
\Longrightarrow &\mathbb{V}^*_{SGD} = \frac{\gamma^2 A^2 \Sigma}{I-(I-\gamma A)^2} = diag(\frac{\alpha^2 a_1^2 \sigma_1^2}{1-(1-\alpha a_1)^2}, \cdots, \frac{\alpha^2 a_n^2 \sigma_n^2}{1-(1-\alpha a_n)^2}),
\end{aligned}$$

where σ_i^2 is the i-th diagonal element of the variance matrix Σ of a gaussian noise c_t. Because

$$\begin{aligned}
\mathbb{V}^*_{SGD} = \lim_{t\to\infty} \mathbb{V}[x_t] &= \lim_{t\to\infty} \mathbb{E}\left[(x_t - \mathbb{E}[x_t])(x_t-\mathbb{E}[x_t])^T\right] \\
&= \lim_{t\to\infty} \mathbb{E}[x_t x_t^T] \\
&= diag(\lim_{t\to\infty} \mathbb{E}[x_{t,1}^2], \lim_{t\to\infty} \mathbb{E}[x_{t,2}^2], \cdots, \lim_{t\to\infty} \mathbb{E}[x_{t,n}^2]),
\end{aligned}$$

we have

$$\lim_{t\to\infty} \mathbb{E}[x_{t,i}^2] = \frac{\alpha^2 a_i^2 \sigma_i^2}{1-(1-\alpha a_i)^2} \quad i = 1, \cdots, n. \tag{14}$$

Since $c_t \sim N(0,\Sigma)$,

$$\lim_{t\to\infty} \mathbb{E}[c_{t,i}^2] = \sigma_i^2 \quad i = 1, \cdots, n. \tag{15}$$

The update formula with learning rate α is

$$x_{t+1} = x_t - \alpha \nabla \hat{L}(x_t) = x_t - \alpha A(x_t - c_t), \quad c_t \sim N(0,\Sigma). \tag{16}$$

For the next iteration, the update formula can be written as

$$\begin{aligned}
x_{t+2} &= x_{t+1} - \alpha \nabla \hat{L}(x_{t+1}) \qquad (17)\\
&= x_{t+1} - \alpha A(x_{t+1} - c_{t+1}), \quad c_{t+1} \sim N(0,\Sigma) \\
&= x_{t+1} - \alpha A(x_t - \alpha A(x_t - c_t)), \quad c_t, c_{t+1} \sim N(0,\Sigma) \\
&= x_{t+1} - \alpha A(x_t - c_{t+1}) + \alpha^2 A^2 (x_t - c_t), \quad c_t, c_{t+1} \sim N(0,\Sigma).
\end{aligned}$$

Define the step at iteration t as $s_t = x_{t+1} - x_t$, then the inner product of two consecutive steps can be written as

$$\begin{aligned}
< s_t, s_{t+1} > &= < -\alpha A(x_t - c_t), -\alpha A(x_t - c_{t+1}) + \alpha^2 A^2 (x_t - c_t) > \qquad (18)\\
&= \alpha^2 (x_t - c_t)^T A^2 (x_t - c_{t+1}) - \alpha^3 (x_t - c_t)^T A^3 (x_t - c_t) \\
&= \alpha^2 \left[x_t^T A^2 x_t - x_t^T A^2 c_{t+1} - c_t^T A^2 x_t + c_t^T A^2 c_{t+1} - \alpha x_t^T A^3 x_t + 2\alpha x_t A^3 c_t - \alpha c_t^T A^3 c_t\right].
\end{aligned}$$

Therefore, the trajectory of the expectation of the inner product converges to

$$\begin{aligned} I^* = \lim_{t\to\infty} \mathbb{E}[< s_t, s_{t+1} >] &= \alpha^2 \left[\lim_{t\to\infty} \mathbb{E}[x_t^T A^2 (I-\alpha A)]x_t - \alpha \lim_{t\to\infty} \mathbb{E}[c_t^T A^3 c_t]\right] \\ &= \alpha^2 \left[\sum_{i=1}^{n} a_i^2 (1-\alpha a_i) \lim_{t\to\infty} \mathbb{E}[x_{t,i}^2] - \sum_{i=1}^{n} \alpha a_i^3 \lim_{t\to\infty} \mathbb{E}[c_{t,i}^2]\right] \\ &= \alpha^2 \sum_{i=1}^{n} \left[a_i^2 (1-\alpha a_i) \frac{\alpha a_i \sigma_i^2}{2-\alpha a_i} - \alpha a_i^3 \sigma_i^2\right] \\ &= \alpha^2 \sum_{i=1}^{n} \alpha a_i^3 \sigma_i^2 \left[\frac{1-\alpha a_i}{2-\alpha a_i} - 1\right] \\ &= -\alpha^3 \sum_{i=1}^{n} \frac{a_i^3 \sigma_i^2}{2-\alpha a_i}. \end{aligned} \tag{19}$$

The norm of step s_t at iteration t is written as

$$\begin{aligned} \|s_t\|^2 &= \|\alpha A (x_t - c_t)\|^2 \\ &= \alpha^2 (x_t - c_t)^T A^2 (x_t - c_t) \\ &= \alpha^2 (x_t^T A^2 x_t - 2 x_t^T A^2 c_t + c_t^T A^2 c_t). \end{aligned} \tag{20}$$

Therefore the trajectory of the expectation of the norm of s_t converges to

$$\begin{aligned} N^* = \lim_{t\to\infty} \mathbb{E}[\|s_t\|^2] &= \alpha^2 \lim_{t\to\infty} \mathbb{E}[x_t^T A^2 x_t] + \alpha^2 \lim_{t\to\infty} \mathbb{E}[c_t^T A^2 c_t] \\ &= \alpha^2 \sum_{i=1}^{n} a_i^2 \left(\mathbb{E}[x_{t,i}^2] + \mathbb{E}[c_{t,i}^2]\right) \\ &= \alpha^2 \sum_{i=1}^{n} a_i^2 \sigma^2 \left(\frac{\alpha a_i}{2-\alpha a_i} + 1\right) \\ &= 2\alpha^2 \sum_{i=1}^{n} \frac{a_i^2 \sigma^2}{2-\alpha a_i}. \end{aligned} \tag{21}$$

Here, in order to draw meaningful conclusions we make certain simplifications and proceed by approximating $\mathbb{E}[cos(\angle(s_t, s_{t+1}))] \approx \mathbb{E}[< s_t, s_{t+1} >]/\mathbb{E}[\|s_t\| \, \|s_{t+1}\|]$.

Because $cos(\angle(s_t, s_{t+1})) = \frac{<s_t, s_{t+1}>}{\|s_t\| \|s_{t+1}\|}$ and $\|s\|_t$ converges when t is large enough, then

$$\lim_{t\to\infty} \mathbb{E}[cos(\angle(s_t, s_{t+1}))] \approx \lim_{t\to\infty} \frac{\mathbb{E}[< s_t, s_{t+1} >]}{\mathbb{E}[\|s_t\|^2]}. \tag{22}$$

Since $I^* = \lim_{t\to\infty} \mathbb{E}[cos(\angle(s_t, s_{t+1}))]$ and $N^* = \lim_{t\to\infty} \mathbb{E}[\|s_t\|^2]$ are both bounded and not equal to 0,

$$\lim_{t\to\infty} \mathbb{E}[cos(\angle(s_t, s_{t+1}))] \approx \frac{\lim_{t\to\infty} \mathbb{E}[< s_t, s_{t+1} >]}{\lim_{t\to\infty} \mathbb{E}[\|s_t\|^2]}. \tag{23}$$

By combining formula (23), (19) and (21), we obtain that the expectation of cosine value converges to

$$C^* = \lim_{t\to\infty} \mathbb{E}[cos(\angle(s_t, s_{t+1}))] \approx \frac{I^*}{N^*} = -\frac{\alpha}{2} \frac{\sum_{i=1}^{n} \frac{a_i^3 \sigma_i^2}{2-\alpha a_i}}{\sum_{i=1}^{n} \frac{a_i^2 \sigma_i^2}{2-\alpha a_i}} \geq -\frac{\alpha}{2} \max_i a_i \frac{\sum_{i=1}^{n} \frac{a_i^2 \sigma_i^2}{2-\alpha a_i}}{\sum_{i=1}^{n} \frac{a_i^2 \sigma_i^2}{2-\alpha a_i}} = -\frac{\alpha \max_i a_i}{2} \tag{24}$$

Since $I - \alpha A \succ 0$ implies $\alpha a_i < 1$ for arbitrary i, then $C^* \in [-\frac{1}{2}, 0]$ and the angle is between 90 degree to 120 degrees. □

A.2 PROOF FOR THEOREM 2

Proof in this section in inspired by Yang et al. (2016).

Proof for Theorem 2. We denote $\mathcal{G}(x_t;\xi_t) = \mathcal{G}(x_t) = \mathcal{G}_t$. The update formula (7) implies the following recursions:

$$x_{t+1} + p_{t+1} = x_t + p_t - \frac{\alpha_t}{1-\beta}\mathcal{G}(x_t) \tag{25}$$

$$v_{t+1} = \beta v_t + ((1-\beta)s - 1)\alpha_t \mathcal{G}(x_t), \tag{26}$$

where $v_t = \frac{1-\beta}{\beta} p_t$ and p_t is given by

$$p_t = \begin{cases} \frac{\beta}{1-\beta}(x_t - x_{t-1} + s\alpha_{t-1}\mathcal{G}(x_{t-1})), & k \geq 1 \\ 0, & k = 0 \end{cases}. \tag{27}$$

Define $\delta_t = \mathcal{G}_t - \partial f(x_t)$ and let x^* be the optimal point. From the above recursions we have

$$\begin{aligned}
&\|x_{t+1} + p_{t+1} - x^*\|^2 \\
=& \|x_t + p_t - x^*\|^2 - \frac{2\alpha_t}{1-\beta}(x_t + p_t - x^*)^T \mathcal{G}_t + \left(\frac{\alpha_t}{1-\beta}\right)^2 \|\mathcal{G}_t\|^2 \\
=& \|x_t + p_t - x^*\|^2 - \frac{2\alpha_t}{1-\beta}(x_t - x^*)^T \mathcal{G}_t - \frac{2\alpha_t\beta}{(1-\beta)^2}(x_t - x_{t-1})^T \mathcal{G}_t \\
&- \frac{2s\alpha_t\alpha_{t-1}\beta}{(1-\beta)^2}\mathcal{G}_{t-1}^T\mathcal{G}_t + \left(\frac{\alpha_t}{1-\beta}\right)^2 \|\mathcal{G}_t\|^2 \\
=& \|x_t + p_t - x^*\|^2 - \frac{2\alpha_t}{1-\beta}(x_t - x^*)^T(\delta_t + \partial f(x_t)) - \frac{2\alpha_t\beta}{(1-\beta)^2}(x_t - x_{t-1})^T(\delta_t + \partial f(x_t)) \\
&- \frac{2s\alpha_t\alpha_{t-1}\beta}{(1-\beta)^2}(\delta_{t-1} + \partial f(x_{t-1}))^T(\delta_t + \partial f(x_t)) + \left(\frac{\alpha_t}{1-\beta}\right)^2 \|\delta_t + \partial f(x_t)\|^2.
\end{aligned} \tag{28}$$

Note that

$$\begin{aligned}
&\mathbb{E}[(x_t - x^*)^T(\delta_t + \partial f(x_t))] = \mathbb{E}[(x_t - x^*)^T \partial f(x_t)] \\
&\mathbb{E}[(x_t - x_{t-1})^T(\delta_t + \partial f(x_t))] = \mathbb{E}[(x_t - x_{t-1})^T \partial f(x_t)] \\
&\mathbb{E}[(\delta_{t-1} + \partial f(x_{t-1}))^T(\delta_t + \partial f(x_t))] = \mathbb{E}[(\delta_{t-1} + \partial f(x_{t-1}))^T \partial f(x_t)] = \mathbb{E}[\mathcal{G}_{t-1}^T \partial f(x_t)] \\
&\mathbb{E}[\|\delta_t + \partial f(x_t)\|^2] = \mathbb{E}[\|\delta_t\|^2] + \mathbb{E}[\|\partial f(x_t)\|^2].
\end{aligned}$$

Taking the expectation on both sides gives the following

$$\begin{aligned}
&\mathbb{E}[\|x_{t+1} + p_{t+1} - x^*\|^2] \\
=& \mathbb{E}[\|x_t + p_t - x^*\|^2] - \frac{2\alpha_t}{1-\beta}\mathbb{E}[(x_t - x^*)^T \partial f(x_t)] - \frac{2\alpha_t\beta}{(1-\beta)^2}\mathbb{E}[(x_t - x_{t-1})^T \partial f(x_t)] \\
&- \frac{2s\alpha_t\alpha_{t-1}\beta}{(1-\beta)^2}\mathbb{E}[\mathcal{G}_{t-1}^T \partial f(x_t)] + \left(\frac{\alpha_t}{1-\beta}\right)^2 (\mathbb{E}[\|\delta_t\|^2] + \mathbb{E}[\|\partial f(x_t)\|^2]).
\end{aligned} \tag{29}$$

Moreover, since f is convex,$\mathbb{E}\left[\|\mathcal{G}(x;\xi) - \mathbb{E}[\mathcal{G}(x;\xi)]\|\right] \leq \delta^2$, and $\|\nabla f(x)\| \leq G$, then for any x

$$\begin{aligned}
&f(x_t) - f(x^*) \leq (x_t - x^*)^T \partial f(x_t) \\
&f(x_t) - f(x_{t-1}) \leq (x_t - x_{t-1})^T \partial f(x_t) \\
&- \mathbb{E}[\mathcal{G}_{t-1}^T \partial f(x_t)] \leq \frac{\mathbb{E}[\|\mathcal{G}_{t-1}\|^2 + \|\partial f(x_t)\|^2]}{2} \leq \delta^2/2 + G^2 \leq \delta^2 + G^2 \\
&\mathbb{E}[\|\delta_t\|^2] \leq \delta^2, \quad \mathbb{E}[\|\partial f(x_t)\|^2] \leq G^2.
\end{aligned}$$

Therefore, (29) can be rewritten as

$$\begin{aligned}\mathbb{E}[\|x_{t+1}+p_{t+1}-x^*\|^2] \le& \mathbb{E}[\|x_t+p_t-x^*\|^2] - \frac{2\alpha_t}{1-\beta}\mathbb{E}[f(x_t)-f(x^*)] \\ &- \frac{2\alpha_t\beta}{(1-\beta)^2}\mathbb{E}[f(x_t)-f(x_{t-1})] + \frac{2s\beta\alpha_t\alpha_{t-1}+\alpha_t^2}{(1-\beta)^2}(G^2+\delta^2).\end{aligned} \tag{30}$$

Since $\hat{\alpha}_i$ is decreasing, it implies that α_t is non-increasing. Thus, (30) could be upper-bounded as

$$\begin{aligned}\mathbb{E}[\|x_{t+1}+p_{t+1}-x^*\|^2] \le& \mathbb{E}[\|x_t+p_t-x^*\|^2] - \frac{2\alpha_t}{1-\beta}\mathbb{E}[f(x_t)-f(x^*)] \\ &- \frac{2\alpha_t\beta}{(1-\beta)^2}\mathbb{E}[f(x_t)-f(x_{t-1})] + \frac{(2s\beta+1)\alpha_t\alpha_{t-1}}{(1-\beta)^2}(G^2+\delta^2).\end{aligned} \tag{31}$$

Taking $t = 0, ..., T-1$ and $x_{-1} = x_0$, and then summing all the inequalities gives

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}[\|x_{t+1}+p_{t+1}-x^*\|^2] \le& \sum_{t=0}^{T-1}\mathbb{E}[\|x_t+p_t-x^*\|^2] - \sum_{t=0}^{T-1}\frac{2\alpha_t}{1-\beta}\mathbb{E}[f(x_t)-f(x^*)] \\ &- \sum_{t=0}^{T-1}\frac{2\alpha_t\beta}{(1-\beta)^2}\mathbb{E}[f(x_t)-f(x_{t-1})] + \frac{(2s\beta+1)(G^2+\delta^2)}{(1-\beta)^2}\sum_{t=0}^{T-1}\alpha_t\alpha_{t-1}.\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{2}{1-\beta}\sum_{t=0}^{T-1}\alpha_t\mathbb{E}[f(x_t)-f(x^*)] \le& \|x_0-x^*\|^2 - \left\|x^T+p_T-x^*\right\| + \frac{2\beta}{(1-\beta)^2}\sum_{t=0}^{T-1}\alpha_t\mathbb{E}[f(x_{t-1})-f(x_t)] \\ &+ \frac{(2s\beta+1)(G^2+\delta^2)}{(1-\beta)^2}\sum_{t=0}^{T-1}\alpha_t\alpha_{t-1},\end{aligned}$$

since $\alpha_{T-1} \le ... \le \alpha_1 \le \alpha_0 < 1$, $\min_{t=0,...,T-1}\{\mathbb{E}[f(x_t)-f(x^*)]\} \le \mathbb{E}[f(x_t)-f(x^*)] (\forall t = 0, ..., T-1)$. Then

$$\begin{aligned}\frac{2}{1-\beta}\min_{t=0,...,T-1}\{\mathbb{E}[f(x_t)-f(x^*)]\}\sum_{t=0}^{T-1}\alpha_t \le& \|x_0-x^*\|^2 + \frac{2\beta}{(1-\beta)^2}\sum_{t=0}^{T-1}\alpha_t\mathbb{E}[f(x_{t-1})-f(x_t)] \\ &+ \frac{(2s\beta+1)(G^2+\delta^2)\sum_{t=0}^{T-1}\alpha_t\alpha_{t-1}}{(1-\beta)^2}.\end{aligned}$$

Moreover, $\alpha_t = \hat{\alpha}_i (t_i \le t < t_{i+1})$ implies that

$$\begin{aligned}\frac{2}{1-\beta}\min_{t=0,...,T-1}\{\mathbb{E}[f(x_t)-f(x^*)]\}\sum_{t=0}^{T-1}\alpha_t \le& \|x_0-x^*\|^2 + \frac{2\beta}{(1-\beta)^2}\sum_{i=0}^{n-1}\hat{\alpha}_i\mathbb{E}[f(x_{t_i})-f(x_{t_{i+1}})] \\ &+ \frac{(2s\beta+1)(G^2+\delta^2)\sum_{t=0}^{T-1}\alpha_t\alpha_{t-1}}{(1-\beta)^2}.\end{aligned}$$

Since $\mathbb{E}[f(x_{t_i})-f(x_{t_{i+1}})]$ is always upper-bounded by $f(x_0)-f(x^*)$, we have

$$\begin{aligned}\frac{2}{1-\beta}\min_{t=0,...,T-1}\{\mathbb{E}[f(x_t)-f(x^*)]\}\sum_{t=0}^{T-1}\alpha_t \le& \|x_0-x^*\|^2 + \frac{2\beta}{(1-\beta)^2}[f(x_0)-f(x^*)]\sum_{i=0}^{n-1}\hat{\alpha}_i \\ &+ \frac{(2s\beta+1)(G^2+\delta^2)\sum_{t=0}^{T-1}\alpha_t\alpha_{t-1}}{(1-\beta)^2}.\end{aligned}$$

After simplification, we have

$$\begin{aligned}\min_{t=0,...,T-1}\{\mathbb{E}[f(x_t)-f(x^*)]\} \le& \frac{(1-\beta)\|x_0-x^*\|^2}{2\sum_{t=0}^{T-1}\alpha_t} + \frac{\beta[f(x_0)-f(x^*)]\sum_{i=0}^{n-1}\hat{\alpha}_i}{(1-\beta)\sum_{t=0}^{T-1}\alpha_t} \\ &+ \frac{(2s\beta+1)(G^2+\delta^2)\sum_{t=0}^{T-1}\alpha_t\alpha_{t-1}}{2(1-\beta)\sum_{t=0}^{T-1}\alpha_t}.\end{aligned} \tag{32}$$

Because $\hat{\alpha}_i \leq (i+2)^{-1}, k_i\hat{\alpha}_i \geq \kappa_1(i+2)^{-\frac{1}{3}}, k_i\hat{\alpha}_i\hat{\alpha}_{i-1} \leq \kappa_2(i+1)^{-\frac{2}{3}}, \forall i = 0, 1, ..., n-1 (n \gg 1)$,

$$\sum_{i=0}^{n-1} \hat{\alpha}_i \leq \sum_{i=0}^{n-1} (i+2)^{-\frac{2}{3}} = \int_0^{n-1} (i+2)^{-\frac{2}{3}} = 3[(n+1)^{\frac{1}{3}} - 2^{\frac{1}{3}}] \tag{33}$$

$$\sum_{t=0}^{T-1} \alpha_t = \sum_{i=0}^{n-1} k_i\hat{\alpha}_i \geq \sum_{i=0}^{n-1} \kappa_1(i+2)^{-\frac{1}{3}} = \kappa_1 \int_0^{n-1} (i+2)^{-\frac{1}{2}} = \frac{3\kappa_1}{2}[(n+1)^{\frac{2}{3}} - 2^{\frac{2}{3}}] \tag{34}$$

$$\sum_{t=0}^{T-1} \alpha_t\alpha_{t-1} \leq \sum_{i=0}^{n-1} k_i\hat{\alpha}_i\hat{\alpha}_{i-1} \leq \kappa_2 \sum_{i=0}^{n-1} (i+1)^{-1} = \kappa_2 \int_0^{n-1} (i+1)^{-1} = \kappa_2 \log n. \tag{35}$$

Substituting (33-35) into inequality (32) gives

$$\min_{t=0,...,T-1} \{\mathbb{E}[f(x_t) - f(x^*)]\} \leq \frac{2\beta(f(x_0) - f(x^*))[(n+1)^{\frac{1}{3}} - 2^{\frac{1}{3}}]}{2\kappa_1(1-\beta)[(n+1)^{\frac{2}{3}} - 2^{\frac{2}{3}}]} + \frac{(1-\beta)\left\|x_0 - x^*\right\|^2}{3\kappa_1[(n+1)^{\frac{2}{3}} - 2^{\frac{2}{3}}]}$$
$$+ \frac{(2s\beta+1)(G^2+\delta^2)\kappa_2 \log n}{3(1-\beta)\kappa_1[(n+1)^{\frac{2}{3}} - 2^{\frac{2}{3}}]}.$$

□

A.3 PROOF FOR THEOREM 3

First, we introduce Lemma 1 which will be used in the proof for Theorem 3. We prove this lemma later in this section.

Lemma 1. *If sequences* $\{\hat{\alpha}_i\}_{i=-1}^{n-1} \subset (0,1)$ *and* $\{k_i\}_{i=0}^{n} \subset \mathbb{N}$ *satisfy:*

$$\hat{\alpha}_i = (i+2)^{-\frac{2}{3}}, \quad \frac{\kappa_1}{\sqrt{\hat{\alpha}_i}} \leq k_i \leq \frac{\kappa_2}{\sqrt{\hat{\alpha}_i}},$$

where κ_1, κ_2 *are constants, then*

$$\hat{\alpha}_i \leq (i+2)^{-\frac{2}{3}}, \quad k_i\hat{\alpha}_i \geq \kappa_1(i+2)^{-\frac{1}{3}}, \quad k_i\hat{\alpha}_i\hat{\alpha}_{i-1} \leq \kappa_2(i+1)^{-1}, \quad \forall i = 0, 1, ..., n-1. \tag{36}$$

Moreover, suppose $T = \sum_{i=0}^{n-1} k_i$. *If* $n \gg 1$ *the following holds*

$$\frac{3\kappa_1}{5}[(n+1)^{\frac{5}{3}} - 2^{\frac{5}{3}}] \leq T \leq \frac{3\kappa_2}{5}[(n+1)^{\frac{5}{3}} - 2^{\frac{5}{3}}]. \tag{37}$$

Proof for Theorem 3. The derivative of the angular velocity model is:

$$v'_\alpha(t) = \frac{\pi(1+\epsilon\alpha)}{2\gamma t^2}.$$

Define the gaps of partition $\Pi : 0 = t_0 < t_1 < ... < t_n = T$ derived from the Algorithm 2 as

$$k_i = t_{i+1} - t_i, \quad \forall i = 0, ..., n-1.$$

Since we drop the learning rate every time the derivative of the angular velocity is smaller that δ, we have

$$v'_{\hat{\alpha}_i}(k_i) = \tau \Longrightarrow k_i = \sqrt{\frac{\pi(1+\epsilon\alpha)}{2\gamma\tau\hat{\alpha}_i}}.$$

Since $\epsilon \in (0, 1/3\hat{\alpha}_0)$, we have

$$\sqrt{\frac{\pi}{2\gamma\tau\hat{\alpha}_i}} \leq k_i \leq \sqrt{\frac{2\pi}{3\gamma\tau\hat{\alpha}_i}}. \tag{38}$$

Define $\kappa_1 = \sqrt{\frac{\pi}{2\gamma\tau}}$ and $\kappa_2 = \sqrt{\frac{2\pi}{3\gamma\tau}}$. By Lemma 1, we have

$$\hat{\alpha}_i \leq (i+2)^{-\frac{2}{3}}, \quad k_i\hat{\alpha}_i \geq \kappa_1(i+2)^{-\frac{1}{3}}, \quad k_i\hat{\alpha}_i\hat{\alpha}_{i-1} \leq \kappa_2(i+1)^{-1}, \quad \forall i = 0, 1, ..., n-1. \tag{39}$$

Then, by combining (39) with Theorem 2 we could conclude that the sequence $\{x_t\}_{t=0}^{T-1}$ generated by the Algorithm 2 satisfies

$$\begin{aligned}\min_{t=0,\ldots,T-1}\{\mathbb{E}[f(x_t)-f(x^*)]\} \leq & \frac{2\beta(f(x_0)-f(x^*))[(n+1)^{\frac{1}{3}}-2^{\frac{1}{3}}]}{2\kappa_1(1-\beta)[(n+1)^{\frac{2}{3}}-2^{\frac{2}{3}}]}+\frac{(1-\beta)\left\|x_0-x^*\right\|^2}{3\kappa_1[(n+1)^{\frac{2}{3}}-2^{\frac{2}{3}}]} \\ & +\frac{(2s\beta+1)(G^2+\delta^2)\kappa_2\log n}{3(1-\beta)\kappa_1[(n+1)^{\frac{2}{3}}-2^{\frac{2}{3}}]}. \end{aligned} \tag{40}$$

By Equation (37) in Lemma 1 we have that

$$\frac{3\kappa_1}{5}(n-1)^{\frac{5}{3}} \leq \frac{3\kappa_1}{5}[(n+1)^{\frac{5}{3}}-2^{\frac{5}{3}}] \leq T \leq \frac{3\kappa_2}{5}[(n+1)^{\frac{5}{3}}-2^{\frac{5}{3}}] \leq \frac{3\kappa_2}{5}(n+1)^{\frac{5}{3}}.$$

Therefore

$$(\frac{5T}{3\kappa_2})^{\frac{3}{5}}-1 \leq n \leq (\frac{5T}{3\kappa_1})^{\frac{3}{5}}+1. \tag{41}$$

Combining (41) with (40) gives

$$\begin{aligned}\min_{t=0,\ldots,T-1}\{\mathbb{E}[f(x_t)-f(x^*)]\} \leq & \frac{2\beta(f(x_0)-f(x^*))[((\frac{5T}{3\kappa_1})^{\frac{3}{5}}+1+1)^{\frac{1}{3}}-2^{\frac{1}{3}}]}{2\kappa_1(1-\beta)[((\frac{5T}{3\kappa_2})^{\frac{3}{5}}-1+1)^{\frac{2}{3}}-2^{\frac{2}{3}}]}+\frac{(1-\beta)\left\|x_0-x^*\right\|^2}{3\kappa_1[((\frac{5T}{3\kappa_2})^{\frac{3}{5}}-1+1)^{\frac{2}{3}}-2^{\frac{2}{3}}]} \\ & +\frac{(2s\beta+1)(G^2+\delta^2)\kappa_2\log((\frac{5T}{3\kappa_1})^{\frac{3}{5}}+1)}{3(1-\beta)\kappa_1[((\frac{5T}{3\kappa_2})^{\frac{3}{5}}-1+1)^{\frac{2}{3}}-2^{\frac{2}{3}}]} \\ = & \frac{2\beta(f(x_0)-f(x^*))[(\frac{5T}{3\kappa_1})^{\frac{3}{5}}+2)^{\frac{1}{3}}-2^{\frac{1}{3}}]}{2\kappa_1(1-\beta)[(\frac{5T}{3\kappa_2})^{\frac{2}{5}}-2^{\frac{2}{3}}]}+\frac{(1-\beta)\left\|x_0-x^*\right\|^2}{3\kappa_1[(\frac{5T}{3\kappa_2})^{\frac{2}{5}}-2^{\frac{2}{3}}]} \\ & +\frac{(2s\beta+1)(G^2+\delta^2)\kappa_2\log((\frac{5T}{3\kappa_1})^{\frac{3}{5}}+1)}{3(1-\beta)\kappa_1[(\frac{5T}{3\kappa_2})^{\frac{2}{5}}-2^{\frac{2}{3}}]} \\ = & O\left(T^{-\frac{1}{5}}\right). \end{aligned}$$

□

A.4 PROOF FOR LEMMA 1

Proof for Lemma 1. First, we show bounds from (36) one by one:

i) $\hat{\alpha}_i=(i+2)^{-\frac{2}{3}} \leq (i+2)^{-\frac{2}{3}}$.

ii) $k_i\hat{\alpha}_i=\kappa_1\sqrt{\hat{\alpha}_i}=\kappa_1(i+2)^{-\frac{1}{3}} \geq \kappa_1(i+2)^{-\frac{1}{3}}$.

iii) $k_i\hat{\alpha}_i\hat{\alpha}_{i=1} \leq \kappa_2\sqrt{\hat{\alpha}_i}\hat{\alpha}_{i-1}=\kappa_2(i+2)^{-\frac{1}{3}}(i+1)^{-\frac{2}{3}} \leq \kappa_2(i+1)^{-1}$.

Secondly, we compute $T=\sum_{i=0}^{n-1}k_i$ according to the definition of k_i. Because $n \gg 1$, the sum of the sequence could be treated as an integral:

$$T=\sum_{i=0}^{n-1}k_i \leq \kappa_2\sum_{i=0}^{n-1}\sqrt{\frac{1}{\hat{\alpha}_i}}=\kappa_2\sum_{i=0}^{n-1}(i+2)^{\frac{1}{3}}=\kappa_2\int_0^{n-1}(i+2)^{\frac{1}{3}}=\frac{3\kappa_2}{5}[(n+1)^{\frac{5}{3}}-2^{\frac{5}{3}}],$$

and

$$T=\sum_{i=0}^{n-1}k_i \geq \kappa_1\sum_{i=0}^{n-1}\sqrt{\frac{1}{\hat{\alpha}_i}}=\kappa_1\sum_{i=0}^{n-1}(i+2)^{\frac{1}{3}}=\kappa_1\int_0^{n-1}(i+2)^{\frac{1}{3}}=\frac{3\kappa_1}{5}[(n+1)^{\frac{5}{3}}-2^{\frac{5}{3}}].$$

□

B EXPERIMENTAL DETAILS

B.1 DATA SETS AND MODELS

The CIFAR-10 and CIFAR-100 data sets (Krizhevsky et al., 2009) consist of 50 K training images, with 10 and 100 different classes respectively. For CIFAR-10 experiments we used a ResNet-18 (He et al., 2016) and a WRN-28x10 (Zagoruyko & Komodakis, 2016) models. For CIFAR-100 experiments we used a ResNet-34 (He et al., 2016) and a WRN-40x10 (Zagoruyko & Komodakis, 2016) models. We do not use the dropout (Srivastava et al., 2014) layers for WRN models in our experiments. The implementation involving WRN architecture and CIFAR data set relies on publicly available codes[3].

The ImageNet (ILSVRC-2012) data set (Deng et al., 2009) consists of 1.2 M images divided into 1 K categories. We train a ResNet-18 (He et al., 2016) model. We use model implementation from PyTorch official model zoo[4].

B.2 TRAINING SETUP

For CIFAR-10 and CIFAR-100 experiments we refer to (Zhang et al., 2019b) and (Zagoruyko & Komodakis, 2016) for ResNet and WRN models respectively. For ImageNet experiments we follow the training procedure proposed by (He et al., 2016).

In all our experiments, for the baseline we use the same setting of hyperparameters (including the learning rate schedule) as recommended in the referenced literature.

---

[3] https://github.com/meliketoy/wide-resnet.pytorch

[4] https://pytorch.org/vision/stable/models.html

B.3 ADDITIONAL RESULTS

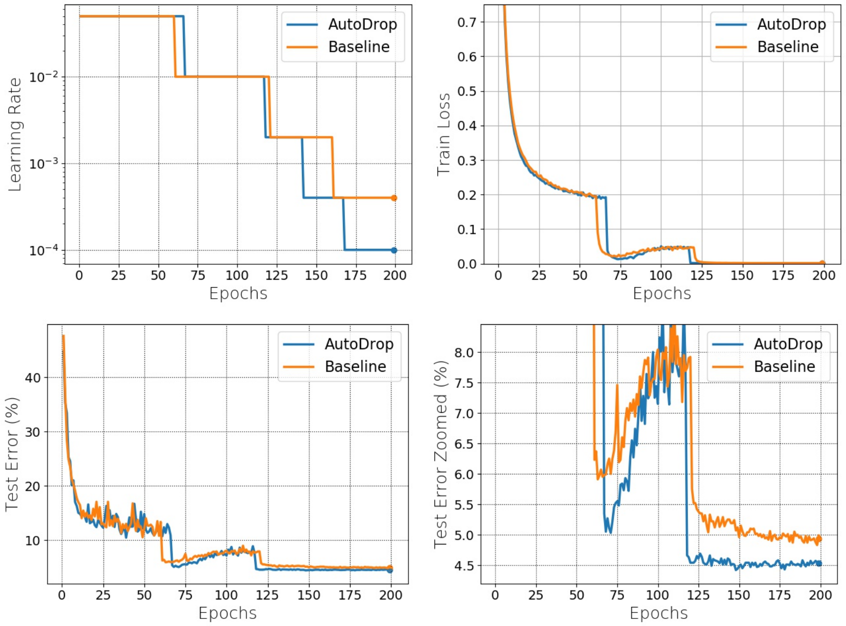

Figure 8: Experimental curves for ResNet-18 model and CIFAR-10 data set. Top (from left to right): learning rate and train loss. Bottom (from left to right): test error and zoomed test error.

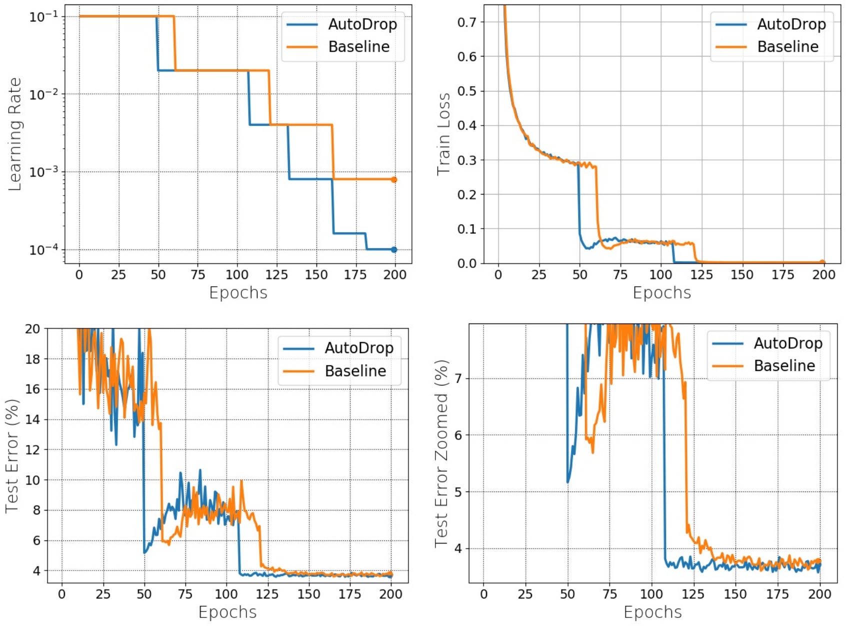

Figure 9: Experimental curves for WRN-28x10 model and CIFAR-10 data set. Top (from left to right): learning rate and train loss. Bottom (from left to right): test error and zoomed test error.

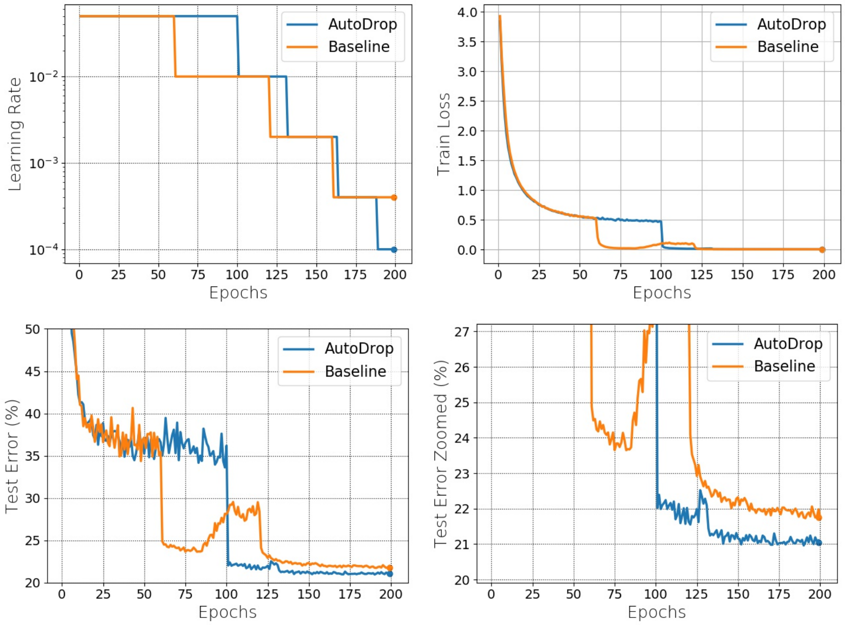

Figure 10: Experimental curves for ResNet-34 model and CIFAR-100 data set. Top (from left to right): learning rate and train loss. Bottom (from left to right): test error and zoomed test error.

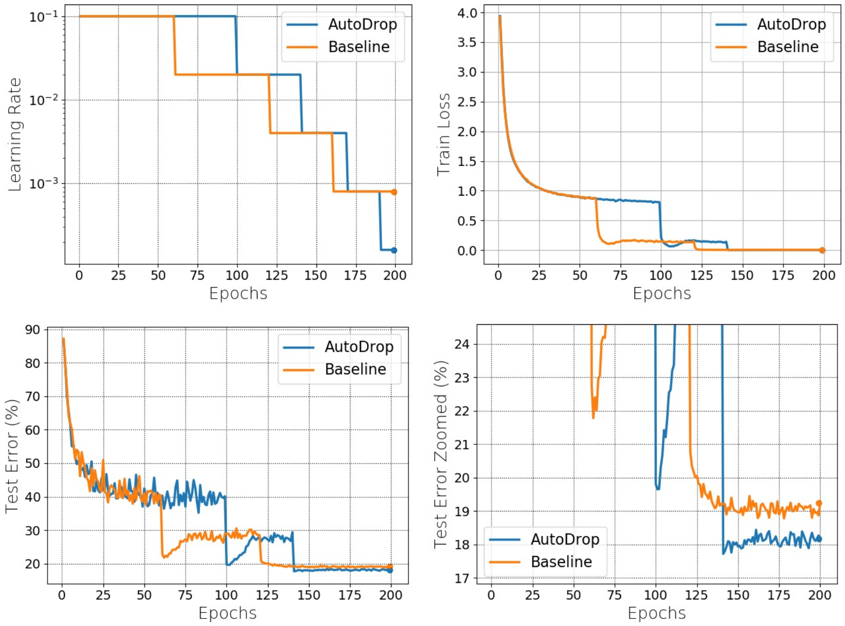

Figure 11: Experimental curves for WRN-40x10 model and CIFAR-100 data set. Top (from left to right): learning rate and train loss. Bottom (from left to right): test error and zoomed test error.

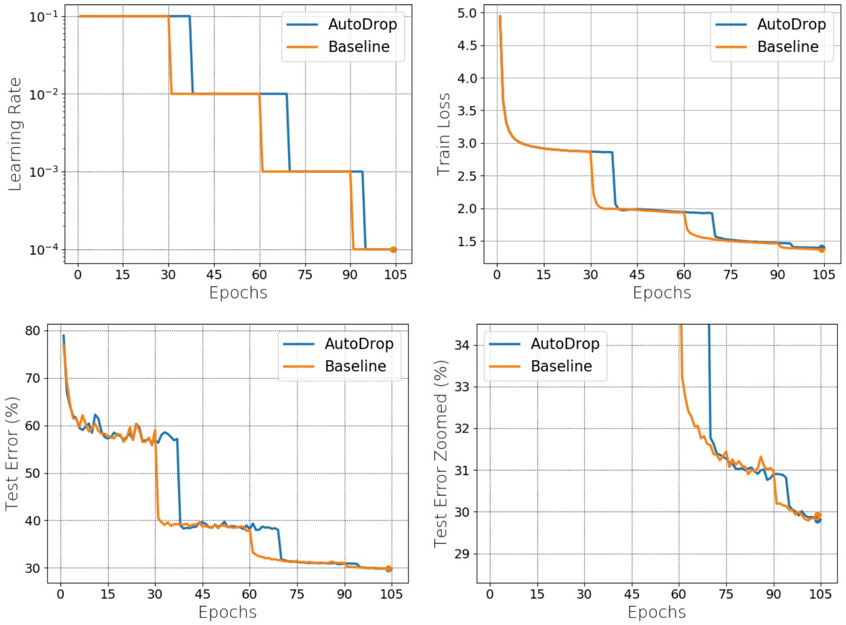

Figure 12: Experimental curves for ResNet-18 model and ImageNet data set. Top (from left to right): learning rate and train loss. Bottom (from left to right): test error and zoomed test error.