Supplementary Material for Momentum is All You Need for Data-Driven Adaptive Optimization

VII. Technical details of Subsection III-B

Here we provide more construction details and technical proofs for the Lévy-driven SDE in Adam-alike adaptive gradient algorithm (2). In the beginning, we introduce a detailed derivation of the process (4) as well as its corresponding escaping set Υ in definition 3. Then we give some auxiliary theorems and lemmas and summarize the proof of Theorem 1. Finally, we prove the Lemma 1 and give a more detailed analysis of the conclusion that the expected escaping time of AdaM^3 is longer than that of Adam in a comparatively flat basin.

A. Derivation of the Lévy-driven SDE (4)

To derive the SDE of Adam-alike algorithms (2), we firstly define $m_t' = \beta_1 m_{t-1}' + (1-\beta_1)\nabla f(\theta_t)$ with $m_0' = 0$. Then by the definition, it holds that

$$m_t' - m_t = (\beta_1 - 1)\sum_{i=0}^{t} \beta_1^{t-i}\zeta_t.$$

Following [29], the gradient noise ζ_t has heavy tails in reality and hence we assume that $\frac{1}{1-\beta_1^t}(m_t' - m_t)$ obeys $\mathcal{S}\widetilde{\alpha}\mathcal{S}$ distribution with time-dependent covariance matrix Σ_t. Since we can formulate (2) as

$$\begin{aligned} \theta_{t+1} &= \theta_t - \alpha\frac{m_t'}{z_t} + \alpha\frac{(m_t' - m_t)}{z_t} \\ \text{where } z_t &= (1-\beta_1^t)\sqrt{\frac{v_t}{(1-\beta_2^t)}}, \end{aligned} \tag{5}$$

and we can replace the term $(m_t' - m_t)$ by $\alpha^{-\frac{1}{\widetilde{\alpha}}}(1-\beta_1^t)\Sigma_t S$ where each coordinate of S is independent and identically distributed as $\mathcal{S}\widetilde{\alpha}\mathcal{S}(1)$ based on the property of centered symmetric $\widetilde{\alpha}$-stable distribution. Let $R_t = \text{diag}(\sqrt{\frac{v_t}{(1-\beta_2^t)}})$, and we further assume that the step size α is small, then the continuous-time version of the process (5) becomes the following SDE:

$$\begin{aligned} d\theta_t &= -R_t^{-1}\frac{m_t' dt}{(1-\beta_1^t)} + \alpha^{1-\frac{1}{\widetilde{\alpha}}}R_t^{-1}\Sigma_t dL_t, \\ dm_t &= (1-\beta_1)(\nabla f(\theta_t) - m_t),\ dv_t = (1-\beta_2)(k_t^2 - v_t). \end{aligned}$$

After replacing m_t' with m_t for brevity, we get the SDE (4) consequently.

B. Proof of Theorem 1

To prove Theorem 1, we first introduce Theorem 2.

Theorem 2. Suppose Assumptions 1-3 hold. We define $\kappa_1 = \frac{c_1 L}{v_-|\tau_m - 1|}$ and $\kappa_2 = \frac{2\mu\tau}{\beta_1 v_+ + \mu\tau}\left(\beta_1 - \frac{\beta_2}{4}\right)$ with a constant c_1. Let $\upsilon^{\widetilde{\alpha}+1} = \Theta(\widetilde{\alpha})$, $\rho_0 = \frac{1}{16(1+c_2)}$ and $\ln\left(\frac{2\Delta}{\mu\upsilon^{1/3}}\right) \le \kappa_2\upsilon^{-1/3}$ where $\Delta = f(\theta_0) - f(\theta^*)$ and a constant c_2. Then for any $\theta_0 \in \mathbf{\Omega}^{-2\upsilon^\gamma}$, $u > -1$, $\upsilon \in (0, \upsilon_0]$, $\gamma \in (0, \gamma_0]$ and $\rho \in (0, \rho_0]$ satisfying $\upsilon^\gamma \le \rho_0$ and $\lim_{\upsilon\to 0}\rho = 0$, the Adam-alike algorithm in (2) obey

$$\frac{1-\rho}{1+u+\rho} \le \mathbb{E}\left[\exp\left(-um(\Upsilon)\Theta(\upsilon^{-1})\Gamma\right)\right] \le \frac{1+\rho}{1+u-\rho}.$$

From Theorem 2, by setting υ small, it holds that for any adaptive gradient algorithm the upper and lower bounds of its expected escaping time Γ is at the order of $\left(\frac{\upsilon}{m(\Upsilon)}\right)$, which directly implies Theorem 1 conclusively. Therefore, it suffices to validate Theorem 2.

The proof of Theorem 2 is given in Section VII-B3. Before we proceeed, we first provide some prerequisite notations in Section VII-B1 and list some useful theorems and lemmas in Section VII-B2.

1) Preliminaries: For analyzing the uniform Lévy-driven SDEs in (4), we first introduce the Lévy process L_t into two components ξ_t and ε_i, namely

$$L_t = \xi_t + \varepsilon_t, \tag{6}$$

whose characteristic functions are respectively defined as

$$\begin{aligned}\mathbb{E}\left[e^{i\langle\lambda,\xi_t\rangle}\right] =& e^{t\int_{\mathbb{R}^d\setminus\{0\}}\varepsilon I\left\{\|y\|_2\le\frac{1}{\upsilon^\delta}\right\}\nu(dy)},\\ \mathbb{E}\left[e^{i\langle\lambda,\varepsilon_t\rangle}\right] =& e^{t\int_{\mathbb{R}^d\setminus\{0\}}\varepsilon I\left\{\|y\|_2\le\frac{1}{\upsilon^\delta}\right\}\nu(dy)},\end{aligned}$$

where $\varepsilon = e^{i\langle\lambda,y\rangle} - 1 - i\langle\lambda,y\rangle I\left\{\|y\|_2 \le 1\right\}$ with υ defined in (4) and a constant δ s.t. $\upsilon^{-\delta} < 1$. Accordingly, the Lévy measure ν of the stochastic processes ξ and ε are

$$\begin{aligned}\nu_\xi &= \nu\left(A\cap\left\{\|y\|_2 \le \frac{1}{\upsilon^\delta}\right\}\right),\\ \nu_\varepsilon &= \nu\left(A\cap\left\{\|y\|_2 \ge \frac{1}{\upsilon^\delta}\right\}\right),\end{aligned}$$

where $A\in\mathcal{B}(\mathbb{R}^d)$. Besides, for analysis, we should consider affects of the Lévy motion L_t to the Lévy-driven SDE of Adam variants. Here we define the Lévy-free SDE accordingly:

$$\begin{cases} d\widehat{\theta}_t = & -\mu_t\widehat{Q}_t^{-1}\widehat{m}_t,\\ d\widehat{m}_t = & (1-\beta_1)(\nabla f(\widehat{\theta}_t) - \widehat{m}_t),\\ d\widehat{v}_t = & (1-\beta_2)(\nabla(f\widehat{\theta}_t)^2 - \widehat{v}_t).\end{cases} \tag{7}$$

where $\widehat{Q}_t = \mathrm{diag}(\sqrt{\widehat{v}_t})$.

2) Auxiliary theorems and lemmas:

Theorem 3 (Adapted from [9])**.** Suppose Assumptions 1-3 hold. Assume the sequence $\{(\widehat{\theta}_t,\widehat{m}_t,\widehat{v}_t)\}$ are produced by (7). Let $\widehat{s}_t = \frac{h_t}{q_t}\left(\sqrt{\omega_t\widehat{v}_t}\right)$ with $h_t = 1-\beta_1$, $q_t = (1-(1-\beta_1)^t)^{-1}$ and $\omega_t = (1-(1-\beta_2)^t)^{-1}$. We define $\|x\|_y^2 = \sum_i y_i x_i^2$. Then for Lévy-driven Adam SDEs in (7), its Lyapunov function $\mathcal{L}(t) = f(\widehat{\theta}_t) - f(\widehat{\theta}^*) + \frac{1}{2}\|\widehat{m}_t\|_{\widehat{s}_t^{-1}}$ with the optimum solution θ^* in the current local basin $\mathbf{\Omega}$ obeys

$$\mathcal{L}(t) \le \Delta\exp\left(-\frac{2\mu\tau}{(1-\beta_1)v_+ + \mu\tau}\left(\frac{3}{4} - \beta_1 + \frac{\beta_2}{4}\right)t\right),$$

where $\Delta = f(\widehat{\theta}_0) - f(\widehat{\theta}^*)$ due to $\widehat{m}_0 = 0$. The sequence $\{\widehat{\theta}_t\}$ produced by (7) obeys

$$\left\|\widehat{\theta}_t - \theta^*\right\|_2^2 \le \frac{2\Delta}{\mu}\exp\left(-\frac{2\mu\tau}{(1-\beta_1)v_+ + \mu\tau}\left(\frac{3}{4} - \beta_1 + \frac{\beta_2}{4}\right)t\right).$$

Lemma 2 ([9])**.** (1) The process ξ in the Lévy process decomposition can be decomposed into two processes $\widehat{\xi}$ and linear drift, namely,

$$\xi_t = \widehat{\xi}_t + \mu_\upsilon t, \tag{8}$$

where $\widehat{\xi}$ is a zero mean Lévymartingale with bounded jumps.
(2) Let $\delta\in(0,1)$, $\mu_\upsilon = \mathbb{E}(\xi_1)$ and $T_\upsilon = \upsilon^{-\theta}$ for some $\theta > 0$, $\rho_0 = \rho_0(\delta) = \frac{1-\delta}{4} > 0$ and $\theta_0 = \theta_0(\delta) = \frac{1-\delta}{3} > 0$. Suppose υ is sufficiently small such that $\Theta(1) \le \upsilon^{-\frac{1-\delta}{6}}$ and $\upsilon^{-\rho} - 2(C+\Theta(1))\upsilon^{\frac{7}{6}(1-\delta)+\frac{\rho}{2}} \ge 1$ with a constant $C = |\int_{0<u\le1} u^2 d\Theta(u)| \in (0,+\infty)$. Then for all $\delta\in(0,\delta_0), \theta\in(0,\theta_0)$ there are $p_0 = p_0(\delta) = \frac{\delta}{2}$ and $\upsilon_0 = \upsilon_0(\delta,\rho)$ such that the estimates

$$\|\upsilon\xi_{T_\upsilon}\|_2 = \upsilon\|\mu_\upsilon\|_2 T_\upsilon < \upsilon^{2\rho} \text{ and } P([\upsilon\xi]_{T_\upsilon}^d \ge \upsilon^\rho) \le \exp(-\upsilon^{-p})$$

hold for all $p\in(0,p_0]$ and $\upsilon\in(0,\upsilon_0]$.

Lemma 3 ([9])**.** Let $\delta\in(0,1)$ and $g_{t\ge0}^t$ be a bounded adapted càdlàg stochastic process with values in $\mathbb{R}^d$, $T_\upsilon = \upsilon^{-\theta}$, $\theta > 0$. Suppose $\sup_{t\ge0}\|g^t\|$ is well bounded. Assume $\rho_0 = \rho_0(\delta) = \frac{1-\delta}{16} > 0$, $\theta_0 = \theta_0(\delta) = \frac{1-\delta}{3} > 0$, $p_0 = \frac{\rho}{2}$. For $\widehat{\xi}_t$ in 8, there is $\delta_0 = \delta_0(\delta) > 0$ such that for all $\rho\in(0,\rho_0)$ and $\theta\in(0,\theta_0)$, it holds

$$\mathbb{P}\left(\sup_{0\le t\le T_\upsilon}\upsilon\left|\sum_{i=1}^d\int_0^t g_{s-}^i d\widehat{\xi}_s^i\right| \ge \upsilon^\rho\right) \le 2\exp\left(-\upsilon^{-p}\right),$$

for all $p \in (0, p_0]$ and $0 < \upsilon \leq \upsilon_0$ with $\upsilon_0 = \upsilon(\rho)$, where $\widehat{\xi}_s^i$ represents the i-th entry in $\widehat{\xi}_s$.

Lemma 4 ([9]). Under Assumptions 1-3 hold, assume $\delta \in (0,1), \rho_0 = \rho_0(\delta) = \frac{1-\delta}{16(1+c_1\kappa_1)} > 0, \theta_0 = \theta_0(\delta) = \frac{1-\delta}{3} > 0, p_0 = \min(\frac{\widehat{\rho}(1+c_1\kappa_1)}{2}, p), \frac{1}{c_2}\ln\left(\frac{2\Delta}{\mu\upsilon^{\widehat{\rho}}}\right) \leq \upsilon^{-\theta_0}$ where $\kappa_1 = \frac{c_2 l}{\upsilon_- |\tau_m - 1|}$ and $c_2 = \frac{2\mu\tau}{(1-\beta_1)\upsilon_+ + \mu\tau}\left(\frac{3}{4} - \beta_1 + \frac{\beta_2}{4}\right)$ in Adam-alike adaptive gradient algorithms. For all $\widehat{\rho} \in (0, \rho_0)$, $p \in (0, p_0]$, $0 < \upsilon \leq \upsilon_0$ with $\upsilon_0 = \upsilon_0(\widehat{\rho})$, and $\theta_0 = \widehat{\theta}_0$, we have

$$\sup_{\theta_0 \in \mathbf{\Omega}} \mathbb{P}\left(\sup_{0 \leq t < \sigma_1} \left\|\theta_t - \widehat{\theta}_t\right\|_2 \geq 2\upsilon^{\widehat{\rho}}\right) \leq 2\exp(-\upsilon^{-\frac{p}{2}}), \tag{9}$$

where the sequences θ_t and $\widehat{\theta}_t$ are respectively produced by (4) and (7) in adaptive gradient method.

3) Proof of Theorem 2:

Proof. The idea of this proof comes from (9) we showed in Lemma 4 where the sequence θ_t and $\widehat{\theta}_t$ start from the same initialization. Based on Theorem 3, we know that the sequence $\{\widehat{\theta}_t\}$ from (7) exponentially converges to the minimum θ^* of the local basin $\mathbf{\Omega}$. To escape the local basin $\mathbf{\Omega}$, we can either take small steps in the process ζ or large jumps J_k in the process ε. However, (9) suggests that these small jumps might not be helpful for escaping the basin. And for big jumps, the escaping time Γ of the sequence $\{\theta_t\}$ most likely occurs at the time σ_1 if the big jump υJ_1 in the process ε is large.
The verification of our desired results can be divided into two separate parts, namely establishing upper bound and lower bound of $\mathbb{E}\left[\exp\left(-um(\Upsilon)\Theta(\upsilon^{-1})\Gamma\right)\right]$ for any $u > -1$. Both of them can be established based on the following facts:

$$\begin{aligned}
&\left|\mathbb{P}\left(R_\theta^{-1}\Sigma_\theta \upsilon J_k \notin \mathbf{\Omega}^{\pm\upsilon^\gamma}, \|\upsilon J_k\|_2 \leq R\right) - \mathbb{P}\left(R_{\theta^*}^{-1}\Sigma_{\theta^*}\upsilon J_k \notin \mathbf{\Omega}^{\pm\upsilon^\gamma}, \|\upsilon J_k\|_2 \leq R\right)\right| \\
&\leq \frac{\delta'}{4} \cdot \frac{\Theta(\upsilon^{-1})}{\Theta(\upsilon^{-\delta})}, \\
&\left|\mathbb{P}\left(R_\theta^{-1}\Sigma_\theta \upsilon J_k \notin \mathbf{\Omega}, \|\upsilon J_k\|_2 \leq R\right) - \mathbb{P}\left(R_{\theta^*}^{-1}\Sigma_{\theta^*}\upsilon J_k \notin \mathbf{\Omega}, \|\upsilon J_k\|_2 \leq R\right)\right| \leq \frac{\delta'}{4} \cdot \frac{\Theta(\upsilon^{-1})}{\Theta(\upsilon^{-\delta})}, \\
&\mathbb{P}\left(R_{\theta^*}^{-1}\Sigma_{\theta^*}\upsilon J_k \notin \mathbf{\Omega}\right) - \mathbb{P}\left(R_{\theta^*}^{-1}\Sigma_{\theta^*}\upsilon J_k \notin \mathbf{\Omega}, \|\upsilon J_k\|_2 \leq R\right) \leq \frac{\delta'}{4} \cdot \frac{\Theta(\upsilon^{-1})}{\Theta(\upsilon^{-\delta})}.
\end{aligned} \tag{10}$$

Specifically, for the upper bound of $\mathbb{E}\left[\exp\left(-um(\Upsilon)\Theta(\upsilon^{-1})\Gamma\right)\right]$, we consider both the big jumps in the process ε and small jumps in the process ζ which may escape the local minimum. Instead of estimating the escaping time Γ from $\mathbf{\Omega}$, we first estimate the escaping time $\widetilde{\Xi}$ from $\mathbf{\Omega}^{-\bar{\rho}}$. Here we define the inner part of $\mathbf{\Omega}$ as $\mathbf{\Omega}^{-\bar{\rho}} := \{y \in \mathbf{\Omega} : \mathrm{dis}(\partial\mathbf{\Omega}, y) \geq \bar{\rho}\}$. Then by setting $\bar{\rho} \to 0$, we can use $\widetilde{\Xi}$ for a decent estimation of Γ. We denote $\bar{\rho} = \upsilon^\gamma$ where γ is a constant such that the results of Lemma 2-4 hold. So for the upper bound, we mainly focus on $\widetilde{\Xi}$ in the beginning and then transfer the results to Γ. In the beginning, we can show that for any $u > -1$ it holds that,

$$\mathbb{E}\left[\exp\left(-um(\Upsilon)\Theta(\upsilon^{-1})\widetilde{\Xi}\right)\right] \leq \sum_{k=1}^{+\infty} \mathbb{E}\left[e^{-um(\Upsilon)\Theta(\upsilon^{-1})t_k} I\left\{\widetilde{\Xi} = t_k\right\} + Res_k\right],$$

where

$$Res_k \leq \begin{cases} \mathbb{E}\left[e^{-um(\Upsilon)\Theta(\upsilon^{-1})t_k} I\left\{\widetilde{\Xi} \in (t_{k-1}, t_k)\right\}\right], & \text{if } u \in (-1, 0] \\ \mathbb{E}\left[e^{-um(\Upsilon)\Theta(\upsilon^{-1})t_{k-1}} I\left\{\widetilde{\Xi} \in (t_{k-1}, t_k)\right\}\right], & \text{if } u \in (0, +\infty). \end{cases}$$

Then using the strong Markov property we can bound the first term $\mathbb{E}\left[e^{-um(\Upsilon)\Theta(\upsilon^{-1})t_k} I\left\{\widetilde{\Xi} = t_k\right\}\right]$ as

$$\begin{aligned}
R_1 = \sum_{k=1}^{+\infty} \mathbb{E}\left[e^{-um(\Upsilon)\Theta(\upsilon^{-1})t_k} I\left\{\Gamma = t_k\right\}\right] &\leq \frac{\alpha_\upsilon(1+\rho/3)}{1+u\alpha_\upsilon} \sum_{k=1}^{+\infty} \left(\frac{1-\alpha_\upsilon(1-\rho)}{1+u\alpha_\upsilon}\right)^{k-1} \\
&\leq \frac{\alpha_\upsilon(1+\rho/3)}{1+u\alpha_\upsilon} \sum_{k=0}^{+\infty} \left(\frac{1-\alpha_\upsilon(1-\rho)}{1+u\alpha_\upsilon}\right)^{k-1} \\
&= \frac{1+\rho/3}{1+u-\rho}.
\end{aligned}$$

On the other hand, for the lower bound of $\mathbb{E}\left[\exp\left(-um(\Upsilon)\Theta(\upsilon^{-1})\Gamma\right)\right]$, we only consider the big jumps in the process ε which could escape from the basin, and ignore the probability that the small jumps in the process ζ which may also lead to an escape from the local minimum θ^*. Specifically, we can find a lower bound by discretization:

$$\mathbb{E}\left[\exp\left(-um(\Upsilon)\Theta(\upsilon^{-1})\Gamma\right)\right] \geq \sum_{k=1}^{+\infty} \mathbb{E}\left[\exp\left(-um(\Upsilon)\Theta(\upsilon^{-1})t_k\right) I\{\Gamma = t_k\}\right].$$

Then we can lower bound each term by three equations (10) we just listed here, which implies that for any $\theta_0 \in \mathbf{\Omega}^{-\upsilon^\gamma}$,

$$\mathbb{E}\left[e^{-um(\Upsilon)\Theta\upsilon^{-1}\Gamma}\right] \geq \frac{\alpha_\upsilon(1-\rho)}{1+u\alpha_\upsilon} \sum_{k=1}^{+\infty} \left(\frac{1-\alpha_\upsilon(1+\rho)}{1+u\alpha_\upsilon}\right)^{k-1} = \frac{1-\rho}{1+u+\rho},$$

where $\rho \to 0$ as $\upsilon \to 0$. The proof is completed. □

C. Proof of Proposition 1

Proof. Since we assumed the minimizer $\theta^* = \mathbf{0}$ in the basin Ω which is usually small,we can employ second-order Taylor expansion to approximate Ω as a quadratic basin whose center is θ^*. In other words, we can write

$$\Omega = \left\{ y \in \mathbb{R}^d \;\middle|\; f(\theta^*) + \frac{1}{2} y^\top H(\theta^*) y \leq h(\theta^*) \right\},$$

where $H(\theta^*)$ is the Hessian matrix at θ^* of function f and $h(\theta^*)$ is the basin height. Then according to Definition 3, we have

$$\Upsilon = \left\{ y \in \mathbb{R}^d \;\middle|\; y^\top \Sigma_{\theta^*} R_{\theta^*}^{-1} H(\theta^*) R_{\theta^*}^{-1} \Sigma_{\theta^*} y \geq h_f^* \right\}.$$

Here $R_{\theta^*} = \lim_{\theta_t \to \theta^*} \operatorname{diag}(\sqrt{v_t/(1-\beta_2^t)})$ is a matrix depending on the algorithm, $h_f^* = 2(h(\theta^*) - f(\theta^*))$ and Σ_{θ^*} is independent of the algorithm, i.e. the same for Adam and AdaM^3. Firstly, we will prove that $v_t^{(\text{AdaM}^3)} \geq v_t^{(\text{Adam})}$ when $t \to \infty$. To clarify the notation, we use θ_t, m_t, v_t, g_t to denote the symbols for Adam and $\widetilde{\theta}_t, \widetilde{m}_t, \widetilde{v}_t, \widetilde{g}_t$ for AdaM^3, and ζ_t is the gradient noise. By using Lemma 1 and above results, we have $\theta_t \approx \widetilde{\theta}_t \approx \theta^*$ before escaping when t is large, and thus $v_t = \lim_{\theta_t \to \theta^*} [\nabla f(\theta_t) + \zeta_t]^2$ and $\widetilde{v}_t = \lim_{\theta_t \to \theta^*} [\beta_1 \widetilde{m}_{t-1} + (1-\beta_1)(\nabla f(\widetilde{\theta}_t) + \zeta_t)]^2$. We will firstly show that $\mathbb{E}(\widetilde{v}_t) \geq \mathbb{E}(v_t)$ when t is large.

$$\begin{aligned}
\mathbb{E}(v_t) &= \mathbb{E}(\lim_{\theta_t \to \theta^*} [\nabla f(\theta_t) + \zeta_t]^2) \overset{\text{(i)}}{=} \lim_{\theta_t \to \theta^*} \mathbb{E}([\nabla f(\theta_t) + \zeta_t]^2) \\
&= \lim_{\theta_t \to \theta^*} \left(\mathbb{E}(\nabla f(\theta_t)^2) + \mathbb{E}(2\nabla f(\theta_t)\zeta_t) + \mathbb{E}(\zeta_t^2)\right) \\
&\overset{\text{(ii)}}{=} \mathbb{E}(\lim_{\theta_t \to \theta^*} \nabla f(\theta_t)^2) + \lim_{\theta_t \to \theta^*} \mathbb{E}(2\nabla f(\theta_t)\zeta_t) + \lim_{\theta_t \to \theta^*} \mathbb{E}(\zeta_t^2) \\
&\overset{\text{(iii)}}{=} \lim_{\theta_t \to \theta^*} \mathbb{E}(\zeta_t^2),
\end{aligned}$$

where (i) and (ii) are due to the dominated convergence theorem (DCT) since we have that we know both $\|\nabla f(\theta_t)\|_2$ and $\|\nabla f(\theta_t) + \zeta_t\|_2$ could be bounded by H in Assumption 4. And (iii) is due to the fact that $\nabla f(\theta^*) = 0$ since function f attains its minimum point at θ^*, and ζ_t has zero mean, i.e.

$$\lim_{\theta_t \to \theta^*} \mathbb{E}(\nabla f(\theta_t)\zeta_t) = \lim_{\theta_t \to \theta^*} \mathbb{E}(\nabla f(\theta_t))\mathbb{E}(\zeta_t) = 0.$$

And similarly we can prove that,

$$\begin{aligned}
\mathbb{E}(\widetilde{v}_t) &= \mathbb{E}\left(\lim_{\theta_t \to \theta^*} [\beta_1 \widetilde{m}_{t-1} + (1-\beta_1)(\nabla f(\widetilde{\theta}_t) + \zeta_t)]^2\right) \\
&= \lim_{\theta_t \to \theta^*} \left(\mathbb{E}(\beta_1^2 \widetilde{m}_{t-1}^2) + \mathbb{E}((1-\beta_1)^2(\nabla f(\widetilde{\theta}_t) + \zeta_t)^2) + \mathbb{E}(2\beta_1(1-\beta_1)\widetilde{m}_{t-1}\nabla(f(\widetilde{\theta}_t) + \zeta_t))\right) \\
&\overset{\text{(i)}}{=} \beta_1^2 \lim_{\theta_t \to \theta^*} \mathbb{E}(\widetilde{m}_{t-1}^2) + (1-\beta_1)^2 \lim_{\theta_t \to \theta^*} \mathbb{E}(\zeta_t^2),
\end{aligned}$$

where we can get the equality (i) simply by the same argument with dominated convergence theorem we just used:

$$\begin{aligned}
\lim_{\widetilde{\theta}_t \to \theta^*} \mathbb{E}(\nabla(f(\widetilde{\theta}_t)^2) &= \mathbb{E}(\lim_{\widetilde{\theta}_t \to \theta^*} \nabla(f(\widetilde{\theta}_t)^2) \overset{\text{(i)}}{=} 0, \\
\lim_{\widetilde{\theta}_t \to \theta^*} \mathbb{E}(\nabla(f(\widetilde{\theta}_t)\zeta_t) &= \mathbb{E}(\lim_{\widetilde{\theta}_t \to \theta^*} \nabla(f(\widetilde{\theta}_t)\zeta_t) \overset{\text{(ii)}}{=} 0, \\
\lim_{\widetilde{\theta}_t \to \theta^*} \mathbb{E}(\widetilde{m}_{t-1}(\nabla f(\widetilde{\theta}_t) + \zeta_t)) &= \mathbb{E}(\lim_{\widetilde{\theta}_t \to \theta^*} \widetilde{m}_{t-1}\nabla f(\widetilde{\theta}_t)) + \lim_{\widetilde{\theta}_t \to \theta^*} \mathbb{E}(\widetilde{m}_{t-1})\mathbb{E}(\zeta_t) \overset{\text{(iii)}}{=} 0,
\end{aligned}$$

where we get the equality (i) and (ii) since the function $f(\widetilde{\theta}_t)^2$ and $f(\widetilde{\theta}_t)\zeta_t$ could be absolutely bounded by H^2. And the first term in equality (iii) is 0 since we have $\|\widetilde{m}_{t-1}\|_2 \leq H$ by its definition and $\nabla f(\theta^*) = 0$, and the second term vanishes since the noise ζ_t has zero mean. Based on the Assumption 5, we have

$$\mathbb{E}(\widetilde{m}_{t-1}^2) \geq \frac{2-\beta_1}{\beta_1}\mathbb{E}(\zeta_t^2),$$

which implies that $\mathbb{E}(\widetilde{v}_t) \geq \mathbb{E}(v_t)$ when t is large. It further indicates that $R_{\theta^*}^{(\text{AdaM}^3)} \geq R_{\theta^*}^{(\text{Adam})}$.
We consider the volume of the complementary set

$$\Upsilon^c = \left\{ y \in \mathbb{R}^d \;\middle|\; y^\top \Sigma_{\theta^*} R_{\theta^*}^{-1} H(\theta^*) R_{\theta^*}^{-1} \Sigma_{\theta^*} y < h_f^* \right\},$$

which can be viewed as a d-dimensional ellipsoid. We can further decompose the symmetric matrix $M := \Sigma_{\theta^*} R_{\theta^*}^{-1} H(\theta^*) R_{\theta^*}^{-1} \Sigma_{\theta^*}$ by SVD decomposition

$$M = U^\top A U,$$

where U is an orthogonal matrix and A is a diagonal matrix with nonnegative elements. Hence the transformation $y \to Uy$ is an orthogonal transformation which means the volume of Υ^c equals the volume of set

$$\left\{ y' \in \mathbb{R}^d \;\middle|\; y'^\top A y' < h_f^* \right\}.$$

Considering the fact that the volume of a d-dimensional ellipsoid centered at $\mathbf{0}$ $E_d(r) = \{(x_1, x_2, \cdots, x_n) : \sum_{i=1}^d \frac{x_i^2}{R_i^2} \leq 1\}$ is

$$V(E_d(r)) = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} \Pi_{i=1}^n R_i,$$

and the fact we just proved that $R_{\theta^*}^{(\text{AdaM}^3)} \geq R_{\theta^*}^{(\text{Adam})}$. Therefore we deduce the volume of $\Upsilon^{(\text{AdaM}^3)}$ is smaller than that of $\Upsilon^{(\text{Adam})}$, which indicates that for Radon measure $m(\cdot)$ we have $m(\Upsilon^{(\text{AdaM}^3)}) \geq m(\Upsilon^{(\text{Adam})})$. Based on Theorem 1, we consequently have $\mathbb{E}(\Gamma^{(\text{AdaM}^3)}) \geq \mathbb{E}(\Gamma^{(\text{Adam})})$. □

VIII. Proofs in Section IV

A. Convergence Analysis in Convex Optimization

We analyze the convergence of AdaM^3 in convex setting utilizing the online learning framework [60]. Given a sequence of convex cost functions $f_1(\theta), \cdots, f_T(\theta)$, the regret is defined as $R(T) = \sum_{t=1}^T [f_t(\theta_t) - f_t(\theta^*)]$, where $\theta^* = \text{argmin}_\theta \sum_{t=1}^T f_t(\theta)$ is the optimal parameter and f_t can be interpreted as the loss function at the t-th step. Then we have:

Theorem 4. Let $\{\theta_t\}$ and $\{v_t\}$ be the sequences yielded by AdaM^3. Let $\alpha_t = \alpha/\sqrt{t}, \beta_{1,1} = \beta_1, 0 < \beta_{1,t} \leq \beta_1 < 1, v_t \leq v_{t+1}$ for all $t \in [T]$ and $\gamma = \beta_1/\sqrt{\beta_2} < 1$. Assume that the distance between any θ_t generated by AdaM^3 is bounded, $\|\theta_m - \theta_n\|_\infty \leq D_\infty$ for any $m, n \in \{1, \cdots, T\}$. Then we have the following bound:

$$\begin{aligned} R(T) \leq & \frac{D_\infty^2 \sqrt{T}}{2\alpha(1-\beta_1)} \sum_{i=1}^d \sqrt{v_{T,i}} + \frac{D_\infty^2}{2(1-\beta_1)} \sum_{t=1}^T \sum_{i=1}^d \frac{\beta_{1,t}\sqrt{v_{t,i}}}{\alpha_t} \\ & + \frac{\alpha\sqrt{1+\log T}}{(1-\beta_1)^3(1-\gamma)\sqrt{1-\beta_2}} \sum_{i=1}^d \|g_{1:T,i}\|_2 . \end{aligned}$$

Theorem 4 implies that the regret of AdaM^3 can be bounded by $\widetilde{O}$[3]$(\sqrt{T})$, especially when the data features are sparse as Section 1.3 in [3] and then we have $\sum_{i=1}^d \sqrt{v_{T,i}} \ll \sqrt{d}$ and $\sum_{i=1}^d \|g_{1:T,i}\|_2 \ll \sqrt{dT}$. Imposing additional assumptions that $\beta_{1,t}$ decays exponentially and that the gradients of f_t are bounded [5], [16], we can obtain:

Corollary 2. Further Suppose $\beta_{1,t} = \beta_1 \lambda^t$ and the function f_t has bounded gradients, $\|\nabla f_t(\theta)\|_\infty \leq G_\infty$ for all $\theta \in \mathbb{R}^d$, AdaM^3 achieves the guarantee $R(T)/T = \widetilde{O}(1/\sqrt{T})$ for all $T \geq 1$:

$$\begin{aligned} \frac{R(T)}{T} \leq & \left[\frac{d\alpha\sqrt{1+\log T}}{(1-\beta_1)^3(1-\gamma)\sqrt{(1-\beta_2)T}} + \frac{dD_\infty^2}{2\alpha(1-\beta_1)\sqrt{T}} \right] \\ & \cdot (G_\infty + \sqrt{\epsilon/1-\beta_2}) + \frac{dD_\infty^2 G_\infty \beta_1}{2\alpha(1-\beta_1)(1-\lambda)^2 T}. \end{aligned}$$

From Corollary 2, the average regret of AdaM^3 converges to zero as T goes to infinity. The proofs of Theorem 4 and Corollary 2 are provided in Appendix VIII-B.

[3] $\widetilde{O}(\cdot)$ denotes $O(\cdot)$ with hidden logarithmic factors.

B. Proof of the convergence results for the convex case

1) Proof of Theorem 4:

Proof. Firstly, according to the definition of AdaM^3 in Algorithm 1, by algebraic shrinking we have

$$
\begin{aligned}
\sum_{t=1}^{T}\frac{m_{t,i}^2}{\sqrt{tv_{t,i}}} &= \sum_{t=1}^{T-1}\frac{m_{t,i}^2}{\sqrt{tv_{t,i}}} + \frac{\left(\sum_{j=1}^{T}(1-\beta_{1,j})\Pi_{k=1}^{T-j}\beta_{1,T-k+1}g_{j,i}\right)^2}{\sqrt{T\left[\sum_{j=1}^{T}(1-\beta_2)\beta_2^{T-j}m_{j,i}^2+\epsilon+\sum_{j=1}^{T-1}\prod_{i=1}^{j}\beta_2^i\epsilon\right]}} \\
&\leq \sum_{t=1}^{T-1}\frac{m_{t,i}^2}{\sqrt{tv_{t,i}}} + \frac{\left(\sum_{j=1}^{T}(1-\beta_{1,j})\Pi_{k=1}^{T-j}\beta_{1,T-k+1}g_{j,i}\right)^2}{\sqrt{T\sum_{j=1}^{T}(1-\beta_2)\beta_2^{T-j}m_{j,i}^2}} \\
&\leq \sum_{t=1}^{T-1}\frac{m_{t,i}^2}{\sqrt{tv_{t,i}}} + \frac{(\sum_{j=1}^{T}\Pi_{k=1}^{T-j}\beta_{1,T-k+1})(\sum_{j=1}^{T}\Pi_{k=1}^{T-j}\beta_{1,T-k+1}g_{j,i}^2)}{\sqrt{T\sum_{j=1}^{T}(1-\beta_2)\beta_2^{T-j}m_{j,i}^2}} \\
&\overset{\text{(i)}}{\leq} \sum_{t=1}^{T-1}\frac{m_{t,i}^2}{\sqrt{tv_{t,i}}} + \frac{(\sum_{j=1}^{T}\beta_1^{T-j})(\sum_{j=1}^{T}\beta_1^{T-j}g_{j,i}^2)}{\sqrt{T(1-\beta_2)\sum_{j=1}^{T}\beta_2^{T-j}m_{j,i}^2}} \\
&\leq \sum_{t=1}^{T-1}\frac{m_{t,i}^2}{\sqrt{tv_{t,i}}} + \frac{1}{1-\beta_1}\frac{\sum_{j=1}^{T}\beta_1^{T-j}g_{j,i}^2}{\sqrt{T(1-\beta_2)\sum_{j=1}^{T}\beta_2^{T-j}m_{j,i}^2}} \\
&= \sum_{t=1}^{T-1}\frac{m_{t,i}^2}{\sqrt{tv_{t,i}}} \\
&\quad + \frac{1}{(1-\beta_1)\sqrt{T(1-\beta_2)}}\sum_{j=1}^{T}\frac{\beta_1^{T-j}g_{j,i}^2}{\sqrt{\sum_{j=1}^{T}\beta_2^{T-j}\left(\sum_{l=1}^{j}(1-\beta_{1,l})\Pi_{k=1}^{j-l}\beta_{1,j-k+1}g_{l,i}\right)^2}} \\
&\leq \sum_{t=1}^{T-1}\frac{m_{t,i}^2}{\sqrt{tv_{t,i}}} + \frac{1}{(1-\beta_1)\sqrt{T(1-\beta_2)}}\sum_{j=1}^{T}\frac{\beta_1^{T-j}g_{j,i}^2}{\sqrt{\sum_{j=1}^{T}\beta_2^{T-j}\left((1-\beta_{1,j})g_{j,l}\right)^2}} \\
&\leq \sum_{t=1}^{T-1}\frac{m_{t,i}^2}{\sqrt{tv_{t,i}}} + \frac{1}{(1-\beta_1)\sqrt{T(1-\beta_2)}}\sum_{j=1}^{T}\frac{\beta_1^{T-j}g_{j,i}^2}{\sqrt{\beta_2^{T-j}(1-\beta_{1,j})^2g_{j,i}^2}} \\
&\overset{\text{(ii)}}{\leq} \sum_{t=1}^{T-1}\frac{m_{t,i}^2}{\sqrt{tv_{t,i}}} + \frac{1}{(1-\beta_1)^2\sqrt{T(1-\beta_2)}}\sum_{j=1}^{T}\gamma^{T-j}g_{j,i},
\end{aligned}
$$

where (i) arises from $\beta_{1,t} \leq \beta_1$, and (ii) comes from the definition that $\gamma = \frac{\beta_1}{\sqrt{\beta_2}}$. Then by induction, we have

$$\begin{aligned}
\sum_{t=1}^{T} \frac{m_{t,i}^2}{\sqrt{t v_{t,i}}} &\leq \sum_{t=1}^{T} \frac{1}{(1-\beta_1)^2\sqrt{t(1-\beta_2)}} \sum_{j=1}^{t} \gamma^{t-j} g_{j,i} \\
&\leq \frac{1}{(1-\beta_1)^2\sqrt{1-\beta_2}} \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \sum_{j=1}^{t} \gamma^{t-j} g_{j,i} \\
&\overset{(i)}{\leq} \frac{1}{(1-\beta_1)^2\sqrt{1-\beta_2}} \sum_{t=1}^{T} g_{t,i} \sum_{j=t}^{T} \frac{\gamma^{j-t}}{\sqrt{j}} \\
&\leq \frac{1}{(1-\beta_1)^2\sqrt{1-\beta_2}} \sum_{t=1}^{T} g_{t,i} \cdot \frac{1}{(1-\gamma)\sqrt{t}} \\
&\leq \frac{1}{(1-\beta_1)^2(1-\gamma)\sqrt{1-\beta_2}} \sum_{t=1}^{T} \frac{g_{t,i}}{\sqrt{t}} \\
&\overset{(ii)}{\leq} \frac{1}{(1-\beta_1)^2(1-\gamma)\sqrt{1-\beta_2}} \left\| g_{1:T,i} \right\|_2 \sqrt{\sum_{t=1}^{T} \frac{1}{t}} \\
&\overset{(iii)}{\leq} \frac{\sqrt{1+\log T}}{(1-\beta_1)^2(1-\gamma)\sqrt{1-\beta_2}} \left\| g_{1:T,i} \right\|_2,
\end{aligned}$$

where (i) exchanges the indices of summing, (ii) employs Cauchy-Schwarz Inequality and (iii) comes from the following bound on harmonic sum:

$$\sum_{t=1}^{T} \frac{1}{t} \leq 1 + \log T.$$

Due to convexity of f_t, we get

$$\begin{aligned}
f_t(\theta_t) - f_t(\theta^*) &\leq g_t^\top (\theta_t - \theta^*) \\
&= \sum_{i=1}^{d} g_{t,i}(\theta_{t,i} - \theta_{,i}^*).
\end{aligned} \tag{11}$$

According to the updating rule, we have

$$\begin{aligned}
\theta_{t+1} &= \theta_t - \alpha_t \frac{m_t}{\sqrt{v_t}} \\
&= \theta_t - \alpha_t \left(\frac{\beta_{1,t}}{\sqrt{v_t}} m_{t-1} + \frac{1-\beta_{1,t}}{\sqrt{v_t}} g_t \right).
\end{aligned} \tag{12}$$

Substracting θ^*, squaring both sides and considering only the i-th element in vectors, we obtain

$$(\theta_{t+1,i} - \theta_{,i}^*)^2 = (\theta_{t,i} - \theta_{,i}^*)^2 - 2\alpha_t \left(\frac{\beta_{1,t}}{\sqrt{v_{t,i}}} m_{t-1,i} + \frac{1-\beta_{1,t}}{\sqrt{v_{t,i}}} g_{t,i} \right) (\theta_{t,i} - \theta_{,i}^*) + \alpha_t^2 \left(\frac{m_{t,i}}{\sqrt{v_{t,i}}} \right)^2.$$

By rearranging the terms, we have

$$\begin{aligned}
&2\alpha_t \frac{1-\beta_{1,t}}{\sqrt{v_{t,i}}} g_{t,i}(\theta_{t,i} - \theta_{,i}^*) \\
&= (\theta_{t,i} - \theta_{,i}^*)^2 - (\theta_{t+1,i} - \theta_{,i}^*)^2 - 2\alpha_t \cdot \frac{\beta_{1,t}}{\sqrt{v_{t,i}}} \cdot m_{t-1,i}(\theta_{t,i} - \theta_{,i}^*) + \alpha_t^2 \left(\frac{m_{t,i}}{\sqrt{v_{t,i}}} \right)^2.
\end{aligned}$$

Further we have

$$
\begin{aligned}
g_{t,i}(\theta_{t,i}-\theta^*_{,i}) =& \frac{\sqrt{v_{t,i}}}{2\alpha_t(1-\beta_{1,t})}[(\theta_{t,i}-\theta^*_{,i})^2-(\theta_{t+1,i}-\theta^*_{,i})^2]+\frac{\alpha_t\sqrt{v_{t,i}}}{2(1-\beta_{1,t})}\left(\frac{m_{t,i}}{\sqrt{v_{t,i}}}\right)^2 \\
&+\frac{\beta_{1,t}}{1-\beta_{1,t}}(\theta^*_{,i}-\theta_{t,i})m_{t-1,i} \\
=& \frac{\sqrt{v_{t,i}}}{2\alpha_t(1-\beta_{1,t})}[(\theta_{t,i}-\theta^*_{,i})^2-(\theta_{t+1,i}-\theta^*_{,i})^2]+\frac{\alpha_t\sqrt{v_{t,i}}}{2(1-\beta_{1,t})}\left(\frac{m_{t,i}}{\sqrt{v_{t,i}}}\right)^2 \\
&+\frac{\beta_{1,t}}{1-\beta_{1,t}}\cdot\frac{v_{t,i}^{\frac{1}{4}}}{\sqrt{\alpha_t}}\cdot(\theta^*_{,i}-\theta_{t,i})\cdot\sqrt{\alpha_t}\cdot\frac{m_{t-1,i}}{v_{t,i}^{\frac{1}{4}}} \\
\le& \frac{\sqrt{v_{t,i}}}{2\alpha_t(1-\beta_1)}[(\theta_{t,i}-\theta^*_{,i})^2-(\theta_{t+1,i}-\theta^*_{,i})^2]+\frac{\alpha}{2(1-\beta_1)}\cdot\frac{m_{t,i}^2}{\sqrt{tv_{t,i}}} \qquad (13)\\
&+\frac{\beta_{1,t}}{2\alpha_t(1-\beta_{1,t})}(\theta^*_{,i}-\theta_{t,i})^2\sqrt{v_{t,i}}+\frac{\beta_1\alpha}{2(1-\beta_1)}\cdot\frac{m_{t-1,i}^2}{\sqrt{tv_{t,i}}}, \qquad (14)
\end{aligned}
$$

where (14) bounds the last term of (13) by Cauchy-Schwarz Inequality and plugs in the value of α_t. Plugging (14) into (12) and summing from $t=1$ to T, we obtain

$$
\begin{aligned}
R(T) =& \sum_{t=1}^{T}\sum_{i=1}^{d} g_{t,i}(\theta_{t,i}-\theta^*_{,i}) \\
\le& \sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\sqrt{v_{t,i}}}{2\alpha_t(1-\beta_1)}[(\theta_{t,i}-\theta^*_{,i})^2-(\theta_{t+1,i}-\theta^*_{,i})^2]+\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\alpha}{2(1-\beta_1)}\cdot\frac{m_{t,i}^2}{\sqrt{tv_{t,i}}} \\
&+\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1,t}}{2\alpha_t(1-\beta_{1,t})}(\theta^*_{,i}-\theta_{t,i})^2\sqrt{v_{t,i}}+\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_1\alpha}{2(1-\beta_1)}\cdot\frac{m_{t-1,i}^2}{\sqrt{tv_{t,i}}} \\
\le& \sum_{i=1}^{d}\frac{\sqrt{v_{1,i}}}{2\alpha_1(1-\beta_1)}(\theta_{1,i}-\theta^*_{,i})^2+\frac{1}{2(1-\beta_1)}\sum_{t=2}^{T}\sum_{i=1}^{d}(\theta_{t,i}-\theta^*_{,i})^2\left(\frac{\sqrt{v_{t,i}}}{\alpha_t}-\frac{\sqrt{v_{t-1,i}}}{\alpha_{t-1}}\right) \qquad (15)\\
&+\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1,t}}{2\alpha_t(1-\beta_1)}(\theta^*_{,i}-\theta_{t,i})^2\sqrt{v_{t,i}}+\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\alpha}{1-\beta_1}\cdot\frac{m_{t,i}^2}{\sqrt{tv_{t,i}}}, \qquad (16)
\end{aligned}
$$

where (16) rearranges the first term of (15). Finally utilizing the assumptions in Theorem 4, we get

$$
\begin{aligned}
R(T) \le& \sum_{i=1}^{d}\frac{\sqrt{v_{1,i}}}{2\alpha_1(1-\beta_1)}D_\infty^2+\frac{1}{2(1-\beta_1)}\sum_{t=2}^{T}\sum_{i=1}^{d}D_\infty^2\left(\frac{\sqrt{v_{t,i}}}{\alpha_t}-\frac{\sqrt{v_{t-1,i}}}{\alpha_{t-1}}\right) \\
&+\frac{D_\infty^2}{2(1-\beta_1)}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1,t}v_{t,i}^{\frac{1}{2}}}{\alpha_t}+\sum_{i=1}^{d}\frac{\alpha\sqrt{1+\log T}}{(1-\beta_1)^3(1-\gamma)\sqrt{1-\beta_2}}\left\|g_{1:T,i}\right\|_2 \\
=& \sum_{i=1}^{d}\frac{\sqrt{v_{T,i}}}{2\alpha_T(1-\beta_1)}D_\infty^2+\frac{D_\infty^2}{2(1-\beta_1)}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1,t}v_{t,i}^{\frac{1}{2}}}{\alpha_t} \\
&+\sum_{i=1}^{d}\frac{\alpha\sqrt{1+\log T}}{(1-\beta_1)^3(1-\gamma)\sqrt{1-\beta_2}}\left\|g_{1:T,i}\right\|_2, \qquad (17)
\end{aligned}
$$

which is our desired result. □

2) Proof of Corollary 2:

Proof. Plugging $\alpha_t=\frac{\alpha}{\sqrt{t}}$ and $\beta_{1,t}=\beta_1\lambda^t$ into Theorem 17, we get

$$
\begin{aligned}
R(T) \le& \frac{D_\infty^2\sqrt{T}}{2\alpha(1-\beta_1)}\sum_{i=1}^{d}\sqrt{v_{T,i}}+\frac{D_\infty^2}{2\alpha(1-\beta_1)}\sum_{t=1}^{T}\sum_{i=1}^{d}\beta_1\lambda^t\sqrt{tv_{t,i}} \\
&+\sum_{i=1}^{d}\frac{\alpha\sqrt{1+\log T}}{(1-\beta_1)^3(1-\gamma)\sqrt{1-\beta_2}}\left\|g_{1:T,i}\right\|_2. \qquad (18)
\end{aligned}
$$

Next, we employ Mathematical Induction to prove that $v_t, i \leq G_\infty$ for any $0 \leq t \leq T, 1 \leq i \leq d$. $\forall i$, we have $m_{0,i}^2 = 0 \leq G_\infty^2$. Suppose $m_{t-1,i} \leq G_\infty$, we have

$$\begin{aligned} m_{t,i}^2 &= (\beta_{1,t} m_{t-1,i} + (1-\beta_{1,t}) g_{t,i})^2 \\ &\overset{\text{(i)}}{\leq} \beta_{1,t} m_{t-1,i}^2 + (1-\beta_{1,t}) g_{t,i}^2 \\ &\leq \beta_{1,t} G_\infty^2 + (1-\beta_{1,t}) G_\infty^2 = G_\infty^2, \end{aligned}$$

where (i) comes from the convexity of function $f = x^2$. Hence by induction, we have $m_{t,i}^2 \leq G_\infty^2$ for all $0 \leq t \leq T$. Furthermore, $\forall i$, we have $v_{0,i} = 0 \leq G_\infty^2$. Suppose $v_{t-1,i} \leq G_\infty^2 + (1-\beta_2^{t-1})\epsilon/(1-\beta_2)$, we have

$$\begin{aligned} v_{t,i} &= \beta_2 v_{t-1,i} + (1-\beta_2) m_{t,i}^2 + \epsilon \\ &\leq \beta_2 G_\infty^2 + (1-\beta_2) G_\infty^2 + \left(\frac{\beta_2 - \beta_2^t}{1-\beta_2} + 1 \right) \epsilon = G_\infty^2 + \frac{1-\beta_2^t}{1-\beta_2} \epsilon. \end{aligned}$$

Therefore, by induction, we have $v_{t,i} \leq G_\infty^2 + (1-\beta_2^t)\epsilon/(1-\beta_2) \leq G_\infty^2 + \epsilon/(1-\beta_2), \forall i, t$. Combining this with the fact that $\sum_{i=1}^d \|g_{1:T,i}\|_2 \leq d G_\infty \sqrt{T}$ and (18), we obtain

$$\begin{aligned} R(T) \leq & \frac{d(G_\infty + \sqrt{\frac{\epsilon}{1-\beta_2}}) D_\infty^2 \sqrt{T}}{2\alpha(1-\beta_1)} \\ & + \frac{d(G_\infty + \sqrt{\frac{\epsilon}{1-\beta_2}}) D_\infty^2 \beta_1}{2\alpha(1-\beta_1)} \sum_{t=1}^T \lambda^t \sqrt{t} + \frac{d G_\infty \alpha \sqrt{1+\log T}}{(1-\beta_1)^3 (1-\gamma) \sqrt{(1-\beta_2)T}}. \end{aligned} \tag{19}$$

For $\sum_{t=1}^T \lambda^t \sqrt{t}$, we apply arithmetic geometric series upper bound:

$$\sum_{t=1}^T \lambda^t \sqrt{t} \leq \sum_{t=1}^T t \lambda^t \leq \frac{1}{(1-\lambda)^2}. \tag{20}$$

Plugging (20) into (19) and dividing both sides by T, we obtain

$$\frac{R(T)}{T} \leq \frac{d(G_\infty + \sqrt{\frac{\epsilon}{1-\beta_2}}) \alpha \sqrt{1+\log T}}{(1-\beta_1)^3 (1-\gamma) \sqrt{(1-\beta_2)T}} + \frac{d D_\infty^2 (G_\infty + \sqrt{\frac{\epsilon}{1-\beta_2}})}{2\alpha(1-\beta_1)\sqrt{T}} + \frac{d D_\infty^2 G_\infty \beta_1}{2\alpha(1-\beta_1)(1-\lambda)^2 T},$$

which concludes the proof. □

C. Proof of the convergence results for the non-convex case

1) Useful Lemma:

Lemma 5. ([34], [61]) Consider a moving average sequence $m_{t+1} = \beta_{1,t} m_t + (1-\beta_t) g_{t+1}$ for tracking $\nabla f(\theta_t)$, where $\mathbb{E}(g_{t+1}) = \nabla f(\theta_t)$ and f is an L-Lipschits continuous mapping. Then we have

$$\begin{aligned} \mathbb{E}_t(\|m_{t+1} - \nabla f(\theta_t)\|_2^2) \leq & \beta_{1,t} \|m_t - \nabla f(\theta_{t-1})\|_2^2 + 2(1-\beta_{1,t})^2 \mathbb{E}_t(\|g_{t+1} - \nabla f(\theta_t)\|_2^2) \\ & + \frac{L^2}{1-\beta_{1,t}} \|\theta_t - \theta_{t-1}\|_2^2. \end{aligned}$$

Based on the above Lemma 5, we could derive the following convergence result in Theorem 1.

2) Proof of Theorem 1:

We denote $\Delta_t = \|m_{t+1} - \nabla f(\theta_2)\|_2^2$, and by applying Lemma 5 we can get:

$$\mathbb{E}_t(\Delta_{t+1}) \leq \beta_{1,t+1} \Delta_t + 2(1-\beta_{1,t+1})^2 \mathbb{E}_t(\|g_{t+2} - \nabla f(\theta_{t+1})\|_2^2) + \frac{L^2}{1-\beta_{1,t+1}} \|\theta_{t+1} - \theta_t\|_2^2. \tag{21}$$

Based on some simple calculation, we can verify that $\sum_{i=0}^{t-1} \beta_2^i \epsilon = (1-\beta_2^t)\epsilon/(1-\beta_2)$, which implies that $1/\sqrt{v_t} \leq b_{u,t}$ holds for all $t \in [T]$ elementwisely. On the other hand, since we have $m_{t+1} = \beta_{1,t} m_{t-1} + (1-\beta_{1,t}) g_t$ with the condition $\|g_t\|_\infty \leq G$ for all $t \in [T]$. Therefore, we can deduce that

$$\|m_t\|_\infty \leq \beta_{1,t} \|m_{t-1}\|_\infty + (1-\beta_{1,t}) G \leq \beta \|m_{t-1}\|_\infty + (1-\beta) G, \quad m_0 = 0,$$

which implies that $\|m_t\|_\infty \le G(1-\beta^t)$ after some simple calculation, and hence we have $m_t^2 \le G^2(1-\beta^T)^2$ elementwise. Next, since we have $v_{t+1} = \beta_2 v_{t-1} + (1-\beta_2)m_t^2 + \epsilon$, we can similarly get

$$\|v_t\|_\infty \le \beta_2 \|v_{t-1}\|_\infty + (1-\beta_2)\left(G^2(1-\beta_{1,1}^T) + \frac{\epsilon}{1-\beta_2}\right), \quad v_0 = 0,$$

which implies that $\|v_t\|_\infty \le \left(G^2(1-\beta^T) + \frac{\epsilon}{1-\beta_2}\right)(1-\beta_2^t)$ and hence $1/\sqrt{v_t} \ge b_{l,t}$. After some simplification of (21), we have

$$\begin{aligned}
&\mathbb{E}_t(\sum_{t=0}^{T}(1-\beta_{1,t+1})\Delta_t) \\
&\le \mathbb{E}\left[\sum_{t=0}^{T}(\Delta_t - \Delta_{t-1}) + \sum_{t=0}^{T} 2\sigma^2(1-\beta_{1,t+1})^2 + \sum_{t=0}^{T}\frac{L^2}{1-\beta_{1,t+1}}\|\theta_{t+1}-\theta_t\|_2^2\right] \\
&\overset{\text{(i)}}{=} \mathbb{E}\left[\sum_{t=0}^{T}(\Delta_t - \Delta_{t-1}) + \sum_{t=0}^{T} 2\sigma^2(1-\beta_{1,t+1})^2 + \sum_{t=0}^{T}\frac{L^2\alpha_t^2 b_{u,t+1}^2}{1-\beta_{1,t+1}}\|m_{t+1}\|_2^2\right],
\end{aligned} \tag{22}$$

where (i) comes from the Lipschitz property of ∇f. On the other hand, since f has Lipschitz gradient, we have:

$$\begin{aligned}
f(\theta_{t+1}) &\le f(\theta_t) + \nabla f(\theta_t)^\top(\theta_{t+1}-\theta_t) + \frac{L}{2}\|\theta_{t+1}-\theta_t\|_2^2 \\
&= f(\theta_t) - \nabla f(\theta_t)^\top(\frac{\alpha_t}{\sqrt{v_t}}m_{t+1}) + \frac{L}{2}\left\|\frac{\alpha_t}{\sqrt{v_t}}m_{t+1}\right\|_2^2 \\
&= f(\theta_t) + \frac{\alpha_t}{2\sqrt{v_t}}\|\nabla f(\theta_t) - m_{t+1}\|_2^2 + \frac{L}{2}\left\|\frac{\alpha_t}{\sqrt{v_t}}m_{t+1}\right\|_2^2 - \frac{\alpha_t}{2\sqrt{v_t}}\|\nabla f(\theta_t)\|_2^2 \\
&\quad - \frac{\alpha_t}{2\sqrt{v_t}}\|m_{t+1}\|_2^2 \\
&\le f(\theta_t) + \frac{\alpha_t b_{u,t}}{2}\Delta_t + \frac{L\alpha_t^2 b_{u,t}^2 - \alpha_t b_{l,t}}{2}\|m_{t+1}\|_2^2 - \frac{\alpha_t b_{l,t}}{2}\|\nabla f(\theta_t)\|_2^2.
\end{aligned} \tag{23}$$

Since we know that $T_0 \lesssim \frac{1}{\alpha_T}$, then we know the overall loss of the first T_0 terms would be $\mathbb{E}(\sum_{t=1}^{T_0}\|\nabla f(\theta)\|_2^2) \lesssim 1/\alpha_T$, and hence

$$\mathbb{E}\left(\frac{1}{T+1}\sum_{t=1}^{T_0}\|\nabla f(\theta)\|_2^2\right) \lesssim \frac{1}{\alpha_T(T+1)}. \tag{24}$$

For the other case when $t > T_0$, without loss of generality we can assume that $T_0 = 0$ for the above argument. We denote $A = \sqrt{\frac{b_{l,T}}{2L^2 b_{u,1}^3}}$ and $\theta^* = \arg\min_\theta f(\theta)$. From (23), we have

$$
\begin{aligned}
&\mathbb{E}\left(\sum_{t=0}^{T} \frac{\alpha_t b_{l,t}}{2} \|\nabla f(\theta_t)\|_2^2\right) \\
&\leq \mathbb{E}\left[\sum_{t=0}^{T}(f(\theta_t) - f(\theta_{t+1})) + \sum_{t=0}^{T} \frac{\alpha_t b_{u,t}}{2}\Delta_t + \sum_{t=0}^{T} \frac{(L\alpha_t^2 b_{u,t}^2 - \alpha b_{l,t})}{2} \|m_{t+1}\|_2^2\right] \\
&\leq f(\theta_0) - f(\theta^*) + \mathbb{E}\left(\sum_{t=0}^{T} \frac{(L\alpha_t^2 b_{u,t}^2 - \alpha b_{l,t})}{2} \|m_{t+1}\|_2^2\right) + \mathbb{E}\left(\sum_{t=0}^{T} \frac{\alpha_t b_{u,t}}{2(1-\beta_{1,t+1})}(1-\beta_{1,t+1})\Delta_t\right) \\
&\overset{\text{(i)}}{\leq} f(\theta_0) - f(\theta^*) + \mathbb{E}\left(\sum_{t=0}^{T} \frac{(L\alpha_t^2 b_{u,t}^2 - \alpha b_{l,t})}{2} \|m_{t+1}\|_2^2\right) + \frac{A b_{u,1}}{2}\mathbb{E}\left(\sum_{t=0}^{T}(1-\beta_{1,t+1})\Delta_t\right) \\
&\overset{\text{(ii)}}{\leq} f(\theta_0) - f(\theta^*) + \mathbb{E}\left(\sum_{t=0}^{T} \frac{(L\alpha_t^2 b_{u,t}^2 - \alpha b_{l,t})}{2} \|m_{t+1}\|_2^2\right) \\
&\quad + \frac{A b_{u,1}}{2}\mathbb{E}\left[\Delta_0 + \sum_{t=0}^{T} 2(1-\beta_{1,t+1})^2\sigma^2 + \sum_{t=0}^{T} \frac{L^2\alpha_t^2 b_{u,t+1}^2}{1-\beta_{1,t+1}} \|m_{t+1}\|_2^2\right] \\
&\overset{\text{(iii)}}{\leq} f(\theta_0) - f(\theta^*) + \frac{A b_{u,1}}{2}(\sigma^2 + \|\nabla f(\theta_0)\|_2^2) + A b_{u,1}\sigma^2 \sum_{t=0}^{T}(1-\beta_{1,t+1})^2 \\
&\quad + \mathbb{E}\left[\sum_{t=0}^{T}\left(\frac{AL^2 b_{u,1}\alpha_t^2 b_{u,t+1}^2}{2(1-\beta_{1,t+1})} + \frac{L\alpha_t^2 b_{u,t}^2 - \alpha_t b_{l,t}}{2}\right)\|m_{t+1}\|_2^2\right] \\
&\overset{\text{(iv)}}{\leq} f(\theta_0) - f(\theta^*) + \frac{A b_{u,1}}{2}(\sigma^2 + \|\nabla f(\theta_0)\|_2^2) + A b_{u,1}\sigma^2 \sum_{t=0}^{T}(1-\beta_{1,t+1})^2,
\end{aligned}
$$

where (i) comes from the fact that $\alpha_t \leq (1-\beta_{1,t+1})A$ based on the conditions in Theorem 1; (ii) could be obtained after we apply (22) to the summation; (iii) is due to the fact that

$$
\begin{aligned}
\mathbb{E}(\Delta_0) &= \mathbb{E}(\|(1-\beta_{1,1})(g_1 - \nabla f(\theta_0)) - \beta_{1,1}\nabla f(\theta_0))\|_2^2) \\
&= (1-\beta_{1,1})^2\mathbb{E}(\|g_1 - \nabla f(\theta_0)\|_2^2) + \beta_{1,1}^2\mathbb{E}\|\nabla f(\theta_0)\|_2^2 \leq \sigma^2 + \|\nabla f(\theta_0)\|_2^2.
\end{aligned}
$$

And we can deduce (iv) by using the assumptions in Theorem 1

$$
\frac{AL^2 b_{u,1}\alpha_t^2 b_{u,t+1}^2}{2(1-\beta_{1,t+1})} \leq \frac{A^2 L^2 b_{u,1} b_{u,t+1}^2}{2} \leq \frac{b_{l,t}}{4}, \quad \frac{L\alpha_t b_{u,t}^2}{2} \leq \frac{b_{l,t}}{4}.
$$

Therefore, we have

$$
\begin{aligned}
&\mathbb{E}\left(\sum_{t=0}^{T} \frac{\alpha_T b_{l,T}}{2} \|\nabla f(\theta_t)\|_2^2\right) \\
&\leq \mathbb{E}\left(\sum_{t=0}^{T} \frac{\alpha_t b_{l,t}}{2} \|\nabla f(\theta_t)\|_2^2\right) \\
&\leq f(\theta_0) - f(\theta^*) + \frac{A b_{u,1}}{2}(\sigma^2 + \|\nabla f(\theta_0)\|_2^2) + A b_{u,1}\sigma^2 \sum_{t=0}^{T}(1-\beta_{1,t+1})^2.
\end{aligned}
$$

Algorithm	Adam	AdamW	Yogi	AdaBound	RAdam	AdaBelief	AdaM^3
Stepsize α	0.001	0.001	0.001	0.001	0.001	0.001	0.001
β_1	0.9	0.9	0.9	0.9	0.9	0.9	0.9
β_2	0.999	0.999	0.999	0.999	0.999	0.999	0.999
Weight decay	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}
ϵ	10^{-8}	10^{-8}	10^{-8}	10^{-8}	10^{-8}	10^{-8}	10^{-8}

TABLE VII: Well tuned hyperparameter configuration of the adaptive gradient methods for CNNs on CIFAR-10.

As a consequence, we can deduce that

$$\begin{aligned}
&\mathbb{E}\left(\sum_{t=0}^{T}\frac{1}{T+1}\left\|\nabla f(\theta_t)\right\|_2^2\right)\\
&\leq \frac{\frac{2(f(\theta_0)-f(\theta^*))}{b_{l,T}}+\sqrt{\frac{1}{2L^2b_{u,1}b_{l,T}}}(\sigma^2+\|\nabla f(\theta_0)\|_2^2)}{\alpha_T(T+1)}+\sqrt{\frac{2}{Lb_{u,1}b_{l,T}}}\frac{\sigma^2\sum_{t=0}^{T}(1-\beta_{1,t+1})^2}{\alpha_T(T+1)}\\
&:= \frac{1}{\alpha_T(T+1)}(Q_1+Q_2\eta(T)),
\end{aligned} \tag{25}$$

where

$$Q_1=\frac{2(f(\theta_0)-f(\theta^*))}{b_{l,T}}+\sqrt{\frac{1}{2L^2b_{u,1}b_{l,T}}}(\sigma^2+\|\nabla f(\theta_0)\|_2^2),\quad Q_2=\sqrt{\frac{2}{Lb_{u,1}b_{l,T}}}\sigma^2.$$

□

3) Proof of Corollary 1: Without loss of generality we choose $1-\beta_{1,t}=\beta/\sqrt{t}$ and $\alpha_t=\alpha/\sqrt{t},\forall t\in[T]$ for some constants α,β with all conditions in Theorem 1 hold, we have

$$T\alpha_T=\alpha\sqrt{T},\quad \eta(T)=\sum_{t=1}^{T}(1-\beta_{1,t})^2=\beta^2\sum_{t=1}^{T}\frac{1}{t}\leq\beta^2(1+\log(T)).$$

After combining this with (25) and making some rearrangement, we have:

$$\begin{aligned}
&\mathbb{E}\left(\sum_{t=0}^{T}\frac{1}{T+1}\left\|\nabla f(\theta_t)\right\|_2^2\right)\\
&\leq \frac{\frac{2(f(\theta_0)-f(\theta^*))}{b_{l,T}}+\sqrt{\frac{1}{2L^2b_{u,1}b_{l,T}}}(\sigma^2+\|\nabla f(\theta_0)\|_2^2)}{\alpha\sqrt{T}}+\sqrt{\frac{2}{Lb_{u,1}b_{l,T}}}\frac{\sigma^2\beta^2(1+\log(T))}{\alpha\sqrt{T}}\\
&:= \frac{1}{\alpha_T(T+1)}(Q_1^*+Q_2^*\eta(T)),
\end{aligned}$$

where

$$\begin{aligned}
Q_1^*&=\frac{2(f(\theta_0)-f(\theta^*))}{b_{l,T}\alpha}+\sqrt{\frac{1}{2L^2b_{u,1}b_{l,T}}}\frac{(\sigma^2+\|\nabla f(\theta_0)\|_2^2)}{\alpha}+\sqrt{\frac{2}{Lb_{u,1}b_{l,T}}}\frac{\sigma^2\beta^2}{\alpha},\\
Q_2^*&=\sqrt{\frac{2}{Lb_{u,1}b_{l,T}}}\frac{\sigma^2\beta^2}{\alpha}.
\end{aligned}$$

□

IX. Additional Experimental Details

A. Image classification

a) CIFAR datasets: The values of the hyperparameters after careful tuning of the reported results of the adaptive gradient methods on CIFAR-10 in the main paper is summarized in Table VII. For SGDM, the optimal hyperparameter setting is: the learning rate is 0.1, the momentum parameter is 0.9, the weight decay parameter is 5×10^{-4}. For Adabound, the final learning rate is set as 0.1 (matching SGDM) and the value of the hyperparameter gamma is 10^{-3}.

b) ImageNet: For SGDM, the tuned stepsize is 0.1, the tuned momentum parameter is 0.9 and the tuned weight decay is 1×10^{-4}. For Adam, the learning rate is 0.001, $\epsilon=1e^{-8}$, and the weight decay parameter is $1e^{-4}$. For AdaM^3, the learning rate is 0.001, $\epsilon=1e^{-16}$ and the weight decay parameter is $5e^{-2}$.

Algorithm	Adam	AdamW	Yogi	AdaBound	RAdam	AdaBelief	AdaM^3
Stepsize α	0.001	0.001	0.01	0.01	0.001	0.001	0.001
β_1	0.9	0.9	0.9	0.9	0.9	0.9	0.9
β_2	0.999	0.999	0.999	0.999	0.999	0.999	0.999
Weight decay	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}
ϵ	10^{-12}	10^{-12}	10^{-8}	10^{-8}	10^{-12}	10^{-16}	10^{-16}

TABLE VIII: Well tuned hyperparameter configuration of adaptive gradient methods for 1-layer-LSTM on Penn Treebank dataset.

Algorithm	Adam	AdamW	Yogi	AdaBound	RAdam	AdaBelief	AdaM^3
Stepsize α	0.01	0.001	0.01	0.01	0.001	0.01	0.001
β_1	0.9	0.9	0.9	0.9	0.9	0.9	0.9
β_2	0.999	0.999	0.999	0.999	0.999	0.999	0.999
Weight decay	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}
ϵ	10^{-12}	10^{-12}	10^{-8}	10^{-8}	10^{-12}	10^{-12}	10^{-16}

TABLE IX: Well tuned hyperparameter configuration of adaptive gradient methods for 2-layer-LSTM on Penn Treebank dataset.

B. LSTM on language modeling

The training and testing perplexity curves are illustrated in Figure 6 and 7. We can clearly see that AdaM^3 is able to make the perplexity descent faster than SGDM and most other adaptive gradient methods during training and mean while generalize much better in testing phase. In experimental settings, the size of the word embeddings is 400 and the number of hidden units per layer is 1150. We employ dropout in training and the dropout rate for RNN layers is 0.25 and the dropout rate for input embedding layers is 0.4.

The optimal hyperparameters of adaptive gradient methods for 1-layer, 2-layer and 3-layer LSTM are listed in Tables VIII, IX and X respectively. For SGDM, the Well tuned stepsize is 30.0 and the momentum parameter is 0.9. For Adabound, the final learning rate is set as 30.0 (matching SGDM) and the value of the hyperparameter gamma is 10^{-3}.

C. Transformer on neural machine translation

For transformer on NMT task, the well tuned hyperparameter values are summarized in Table XI. The stepsize of SGDM is 0.1 and the momentum parameter of SGDM is 0.9. Initial learning rate is 10^{-7} and the minimum learning rate threshold is set as 10^{-9} in the warm-up process for all the optimizers.

D. Generative Adversarial Network

The optimal momentum parameters of SGD for all GANs are 0.9. For adaptive gradient methods, the well tuned hyperparameter values for BigGAN with consistency regularization are summarized in Table XII. We implement the GAN experiments adapting the code from public repository [4]. We sample two visualization results of generated samples of GAN training with AdaM^3 in Figure 8.

X. Hypotheis Test Results

Doing pair-wise hypothesis tests between our optimizer and each other optimizer helps performance comparison. Note that all the experiments reported in our main paper are run 5 times independently using random seeds. Since the sample size is comparatively small, we adopt non-parametric method Exact paired Wilcoxon signed rank test together with Paired t-test to conduct hypothesis tests on CIFAR image classification experiments (using best epoch test accuracy), LSTM language modeling experiments and BigGAN image generation experiment. The null hypothesis is that the compared baseline method is no worse than our AdaM^3 in performance, while the alternative hypothesis is that our proposed AdaM^3 is better than the compared baseline gradient method. The results (p-values) are summarized in Tab. XIII-XIX. We can conclude from the tables that most of the P-values are quite small, which demonstrates that our shown superiority of AdaM^3 is reliable and universal.

[4] https://github.com/POSTECH-CVLab/PyTorch-StudioGAN

Algorithm	Adam	AdamW	Yogi	AdaBound	RAdam	AdaBelief	AdaM^3
Stepsize α	0.01	0.001	0.01	0.01	0.001	0.01	0.001
β_1	0.9	0.9	0.9	0.9	0.9	0.9	0.9
β_2	0.999	0.999	0.999	0.999	0.999	0.999	0.999
Weight decay	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}	1.2×10^{-4}
ϵ	10^{-12}	10^{-12}	10^{-8}	10^{-8}	10^{-12}	10^{-12}	10^{-16}

TABLE X: Well tuned hyperparameter configuration of adaptive gradient methods for 3-layer-LSTM on Penn Treebank dataset.

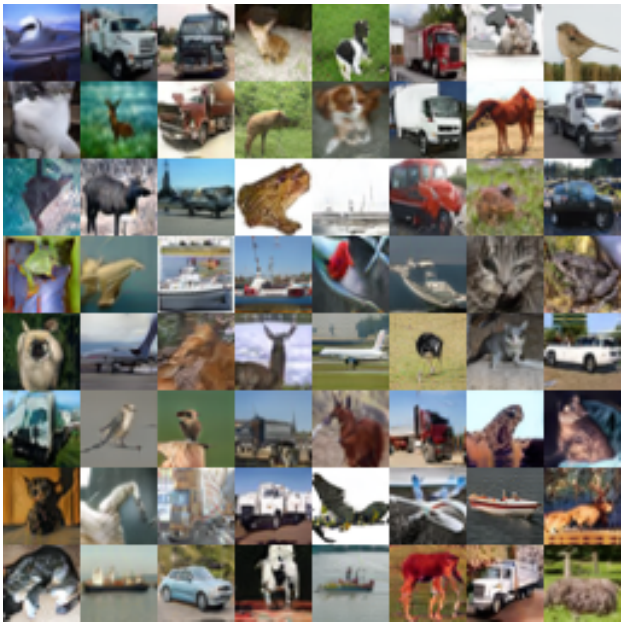
Algorithm	Adam	AdamW	AdaBelief	AdaM^3
Stepsize α	0.0015	0.0015	0.0015	0.0005
β_1	0.9	0.9	0.9	0.9
β_2	0.98	0.98	0.999	0.999
Weight decay	10^{-4}	10^{-4}	10^{-4}	10^{-4}
ϵ	10^{-8}	10^{-8}	10^{-16}	10^{-16}

TABLE XI: Well tuned hyperparameter configuration of adaptive gradient methods for transformer on IWSTL'14 DE-EN dataset.

Algorithm	Adam	Yogi	AdaBound	RAdam	AdaBelief	AdaM^3
β_1	0.5	0.5	0.5	0.5	0.5	0.5
β_2	0.999	0.999	0.999	0.999	0.999	0.999
Weight decay	0	0	0	0	0	0
ϵ	10^{-8}	10^{-8}	10^{-8}	10^{-8}	10^{-16}	10^{-16}

TABLE XII: Well tuned hyperparameter configuration of adaptive gradient methods for BigGAN with consistency regularization.

(a) DCGAN trained with random seed 0 (best FID score 43.52, iteration 26000).



(b) BigGAN trained with random seed 2 (best FID score 7.07, iteration 92000).

Fig. 8: Generated figures trained on CIFAR-10 optimizing with AdaM^3.

Test type	SGDM	Adam	AdamW	Yogi	AdaBound	RAdam	AdaBelief
Exact paired Wilcoxon signed rank test	0.313	0.031	0.031	0.031	0.031	0.031	0.031
Paried t-test	0.246	3.96e-6	8.34e-5	1.28e-4	1.66e-5	1.73e-5	0.012

TABLE XIII: P-values (↓) calculated using pair-wise hypothesis tests between AdaM and the baseline optimizers on VGGNet-16 for CIFAR-10 classification.

Test type	SGDM	Adam	AdamW	Yogi	AdaBound	RAdam	AdaBelief
Exact paired Wilcoxon signed rank test	0.901	0.031	0.031	0.031	0.031	0.031	0.031
Paried t-test	0.902	4.35e-4	3.77e-4	1.17e-4	3.44e-4	2.59e-4	0.006

TABLE XIV: P-values (↓) calculated using pair-wise hypothesis tests between AdaM and the baseline optimizers on ResNet-34 for CIFAR-10 classification.

Test type	SGDM	Adam	AdamW	Yogi	AdaBound	RAdam	AdaBelief
Exact paired Wilcoxon signed rank test	0.094	0.031	0.031	0.031	0.031	0.031	0.094
Paried t-test	0.112	2.17e-4	1.52e-3	4.99e-5	6e-4	0.003	0.04

TABLE XV: P-values (↓) calculated using pair-wise hypothesis tests between AdaM and the baseline optimizers on DenseNet-121 for CIFAR-10 classification.

Test type	SGDM	Adam	AdamW	Yogi	AdaBound	RAdam	AdaBelief
Exact paired Wilcoxon signed rank test	0.031	0.031	0.031	0.031	0.031	0.031	0.031
Paried t-test	4.54e-7	2.46e-6	2.21e-7	5.58e-7	6.07e-6	1.36e-5	1.36e-5

TABLE XVI: P-values (↓) calculated using pair-wise hypothesis tests between AdaM and the baseline optimizers on 1-layer LSTM for Penn Treebank dataset.

Test type	SGDM	Adam	AdamW	Yogi	AdaBound	RAdam	AdaBelief
Exact paired Wilcoxon signed rank test	0.031	0.031	0.031	0.031	0.031	0.031	0.031
Paried t-test	1.96e-5	2.50e-5	3.75e-9	1.88e-7	1.96e-5	2.16e-7	7.45e-6

TABLE XVII: P-values (↓) calculated using pair-wise hypothesis tests between AdaM and the baseline optimizers on 2-layer LSTM for Penn Treebank dataset.

Test type	SGDM	Adam	AdamW	Yogi	AdaBound	RAdam	AdaBelief
Exact paired Wilcoxon signed rank test	0.031	0.031	0.031	0.031	0.031	0.031	0.031
Paried t-test	5.76e-6	3.75e-6	6.56e-8	2.26e-8	1.76e-6	1.16e-8	2.16e-4

TABLE XVIII: P-values (↓) calculated using pair-wise hypothesis tests between AdaM and the baseline optimizers on 3-layer LSTM for Penn Treebank dataset.

Test type	SGDM	Adam(W)	Yogi	AdaBound	RAdam	AdaBelief
Exact paired Wilcoxon signed rank test	0.031	0.031	0.031	0.031	0.031	0.094
Paried t-test	3.38e-7	1.01e-3	1.67e-4	3.68e-7	0.215	0.076

TABLE XIX: P-values (↓) calculated using pair-wise hypothesis tests between AdaM and the baseline optimizers on BigGAN experiments on CIFAR-10.