# ADAME: DYNAMIC EXPONENTIAL DECAY RATES FOR THE ADAM OPTIMIZER

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Adaptive optimizers such as Adam are widely used in deep learning. Their gradient statistics are governed by two fixed decay rates, most commonly 0.9 and 0.999. The best pair differs across tasks, and searching for it is costly. We show that under any fixed second-moment rate the effective per-step gradient scale can shrink over training, the failure mode behind known counterexamples where Adam fails to converge. We address this with AdamE, a drop-in variant of Adam whose decay rates follow dynamic, closed-form schedules: the first rate starts at one half and decays smoothly toward zero, while the second rises toward one, with both rates determined by the iteration index. The schedules introduce no new hyperparameters and provably restore the monotonicity that fixed rates lack, recovering convergence without altering the update rule as AMSGrad does. We prove that AdamE attains near-optimal regret in the online convex setting, converges to a stationary point on smooth non-convex objectives, and admits an unconditional error bound on least squares, where the corresponding Adam guarantee holds only in a restricted regime. Experiments on language modeling, node classification, graph clustering, image classification, and image recognition show that AdamE achieves competitive convergence and final performance relative to well-tuned Adam-type baselines, with no manual tuning of the decay rates.

## 1 INTRODUCTION

Adam (Kingma & Ba, 2014), building on RMSProp (Tieleman & Hinton, 2012) and AdaGrad (Duchi et al., 2011), is among the most widely used stochastic optimizers for training deep neural networks (DNNs), with applications spanning image recognition (He et al., 2016), sequence modeling (Hochreiter & Schmidhuber, 1997), and graph learning (Veličković et al., 2017). However, Reddi et al. (2019) constructed convex problems on which Adam *fails to converge*, and traced the failure to the fact that the key quantity $\sqrt{\hat{\mathbf{v}}_t/t}$ need not be monotone when the decay rates are fixed. Their remedy, AMSGrad, restores convergence by applying an elementwise maximum to the second moment estimate, which modifies the second-moment update. Subsequent variants, including AdaBound (Luo et al., 2019), RAdam (Liu et al., 2019), Padam (Chen et al., 2018), AdamW (Loshchilov & Hutter, 2017), AdaBelief (Zhuang et al., 2020), and AdaSwitch (Zou et al., 2026), improve other aspects of Adam, but they all retain *fixed* decay rates for the moment estimates; how the choice of these rates shapes convergence, generalization, and stability has received far less attention.

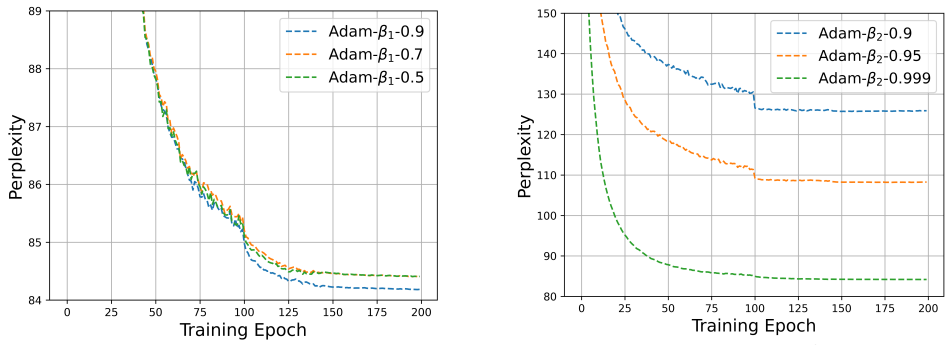
How much do these two fixed rates actually matter in practice? To find out, we ran Adam with $\beta_1 \in \{0.5, 0.7, 0.9\}$ and $\beta_2 \in \{0.9, 0.95, 0.999\}$ on two standard tasks drawn from different domains, holding all other hyperparameters fixed.

**Language modeling.** Training a one-layer LSTM on Penn TreeBank for 200 epochs (initial learning rate $0.001$), we find that perplexity is nearly invariant to $\beta_1$ when $\beta_2$ is fixed at $0.999$, whereas increasing $\beta_2$ with $\beta_1$ fixed at $0.9$ yields a clear improvement (Fig. 1); the best configuration on this task is the familiar $(\beta_1, \beta_2) = (0.9, 0.999)$.

**Node classification.** The picture reverses on graph data. Training Graph Attention Networks (GAT) (Veličković et al., 2017) on Cora with the protocol of Veličković et al. (2017) (initial learning rate $0.01$, decayed by a factor of $0.1$ after the 10th and 30th epochs, 50 hidden units), we observe



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




a. Adam with different $\beta_1$ b. Adam with different $\beta_2$

Figure 1: Adam with different $\beta_1$ and $\beta_2$ for language modeling.

**Algorithm 1** AdamE

**Input:** Initial parameter $\boldsymbol{\theta}_0$.
**Initialize:** $\lambda$ (base step size, tuned per task like Adam's learning rate; values used in our experiments are reported in Section 7), $\varepsilon = 10^{-8}$ (numerical-stability constant added to $\sqrt{s_t}$), $d_0 = 0$ (initial first moment estimate), and $s_0 = 0$ (initial second raw moment estimate).
**Output:** $\boldsymbol{\theta}_T$.
**for** $t = 1; t \leq T$ **do**
$g_t = \nabla_\theta f_t(\theta_{t-1})$; (Get gradients of stochastic objective function at the $t$th step)
$\alpha_t = \frac{t}{1+t^2}$; (Calculate the exponential decay rate of the first moment estimate)
$\beta_t = 1 - \alpha_t$; (Calculate the exponential decay rate of the second raw moment estimate)
$d_t = \alpha_t * d_{t-1} + (1 - \alpha_t) * g_t$; (Calculate the first moment estimate)
$s_t = \beta_t * s_{t-1} + (1 - \beta_t) * g_t^2$; (Calculate the second raw moment estimate)
$\theta_t = \theta_{t-1} - \lambda * \frac{d_t}{\sqrt{s_t}+\varepsilon}$ (Update parameters)
**end for**

in Figure F.1 of the supplementary material that $\beta_1 = 0.9$ converges more slowly and reaches lower test accuracy than smaller values, while test accuracy remains largely stable across $\beta_2$; the best final accuracy is attained with $(\beta_1, \beta_2) = (0.9, 0.9)$, even though smaller $\beta_1$ values converge noticeably faster.

Two findings emerge. First, the optimal $(\beta_1, \beta_2)$ pair is task-dependent: the configuration that is best for language modeling is mediocre for node classification, and vice versa, so no single fixed pair can serve as a safe default across applications. Second, tuning the pair requires additional runs: the search above already spans nine combinations on two tasks alone. These observations, together with our convergence analysis (Section 4), which shows that fixed $\beta_2$ leaves the key quantity $\sqrt{\hat{\mathbf{v}}_t/t}$ prone to decreasing, the failure mode identified by Reddi et al. (2019), motivate the question: *can the decay rates tune themselves?* We propose AdamE, a variant of Adam whose decay rates follow the closed-form schedules $\alpha_t = \frac{t}{1+t^2}$ and $\beta_t = 1 - \alpha_t$, requiring no additional hyperparameters (Algorithm 1). Throughout the paper, all quantities $\alpha_t, \beta_t, d_t, s_t$ refer to Algorithm 1 with this unified notation.

Our contributions are:

- **AdamE with dynamic exponential decay rates**. We propose AdamE (Algorithm 1), which replaces Adam's fixed decay rates with the closed-form schedules $\alpha_t = \frac{t}{1+t^2}$ and $\beta_t = 1 - \alpha_t$. The schedules introduce *no* new hyperparameters: both rates are computed exactly at every step, in contrast to Adam-type variants that add tuning knobs. Compared with Adam and other mainstream adaptive methods, AdamE adapts more closely to the current gradient values, yielding faster convergence and improved stability.
- **Convergence analysis in convex and nonconvex settings**. We identify structural monotonicity properties of AdamE's dynamic rates: the effective second-moment weight $t\,\mathbf{s}_{t,i}$ is provably nondecreasing (Lemma A.3 of the supplementary material) and the effective step

size $\lambda_t \hat{\mathbf{s}}_{t,i}^{-1/2}$ is provably nonincreasing (Lemma A.4 of the supplementary material). We show that their failure under fixed rates underlies the non-convergence of Adam. Building on these properties, we prove that (i) on a least-squares problem, AdamE admits an unconditional, parameter-free error bound, while Adam's bound requires $\beta_2 > \beta_1^2$, has constants that explode as $\beta_2 \downarrow \beta_1^2$, and admits no polynomial bound at all when $\beta_2 \leq \beta_1^2$, a regime that overlaps the classical AMSGrad parameter condition (Theorem 1); (ii) in the online convex setting, AdamE attains the optimal $O(\sqrt{T})$ regret up to a logarithmic factor, *without* the elementwise-maximum modification required by AMSGrad (Theorem 3); and (iii) in the nonconvex setting, AdamE converges to a stationary point at the rate $O(\log T/\sqrt{T})$ under a mild no-spike budget, with the step-size monotonicity that such analyses normally assume holding for AdamE identically (Theorem 4).

- **Experimental evaluation**. We evaluate AdamE on seven public benchmarks (WikiText-2, BBBP, Cora, Citeseer, DBLP, CIFAR-100, and ImageNet) spanning language modeling, node classification, graph clustering, image classification, and image recognition, against seven well-tuned Adam-type baselines. Across these settings, AdamE achieves faster convergence on several benchmarks and competitive or superior final performance.

## 2 Related Work

**Adaptive gradient methods.** AdaGrad (Duchi et al., 2011) pioneered per-coordinate scaling by accumulated squared gradients, with logarithmic regret guarantees in online convex optimization; RMSProp (Tieleman & Hinton, 2012) and Adam (Kingma & Ba, 2014) refined this idea with exponential moving averages governed by fixed decay rates. Reddi et al. (2019) proved that Adam can diverge on simple convex problems and proposed AMSGrad, which enforces monotonicity of the second moment estimate through an elementwise maximum; Mukkamala & Hein (2017) obtained logarithmic-regret variants of RMSProp and AdaGrad under analogous modifications. In AdamE, monotonicity is achieved *structurally* by the closed-form dynamic rates $\alpha_t, \beta_t$, rather than by modifying the update rule, so the analyzed algorithm is exactly the algorithm that is run.

**Variants with fixed decay rates.** Other methods keep the exponential moving averages but change other components: AdamW (Loshchilov & Hutter, 2017) decouples weight decay from the adaptive scaling; RAdam (Liu et al., 2019) introduces a variance-rectification term to warm up the step size; AdaBound (Luo et al., 2019) dynamically bounds the learning rate between SGD-like and Adam-like regimes; Padam (Chen et al., 2018) partially adapts the step size to close the generalization gap; AdaBelief (Zhuang et al., 2020) rescales steps by the variance of the gradient prediction; EAdam (Yuan & Gao, 2020) moves $\varepsilon$ inside the second-moment estimate; and AdaSwitch (Zou et al., 2026) switches from Adam to SGDM through an exponential function. All of these methods retain fixed $\beta_1$ and $\beta_2$. AdamE is orthogonal to and composable with several of these directions; our results show that dynamic decay rates suffice to restore the convergence guarantees that fixed rates cannot deliver.

**Convergence analyses of Adam-type methods.** Other analyses study *when* the original updates converge: via growing minibatches (Zaheer et al., 2018), via a sufficient condition that the second-moment weight be nondecreasing toward 1 (Zou et al., 2019), via a simple unified nonconvex proof (Défossez et al., 2022), or via the hyperparameter regime $\beta_1 < \sqrt{\beta_2}$ (Zhang et al., 2022a), the boundary our Theorem 1 identifies. AdamE follows the philosophy of Zou et al. (2019): the second-moment weight is nondecreasing toward 1 *by construction* ($\beta_t \uparrow 1$), so the sufficient condition holds at every step without tuning.

## 3 Numerical Verification

### 3.1 Results on Quadratic Function Optimization

We first evaluate AdamE on a quadratic fitting problem. The target function is $f(x) = (x-1)^2 + 2$ on the search domain $-5 \leq x \leq 7$, whose global minimum is at $\tilde{x} = 1$ with $f(\tilde{x}) = 2$. For Adam with different $(\beta_1, \beta_2)$ settings, the learning rate is 0.8 and the number of iterations is 200. As shown in Figure F.7 of the supplementary material, the overshooting common to optimization

methods is mitigated by dynamically adjusting the step size through the exponential decay rates of the two moment estimates, without bias correction and with the same initial learning rate.

Figure F.8 of the supplementary material further demonstrates that increasing $\beta_2$ in Adam gradually shrinks the effective per-parameter step size $\frac{\alpha}{\sqrt{v_t}}$, identifying $\beta_2$ as the key factor behind the extremes of Adam's learning rates; AdamE's dynamic rate removes this pathology by construction. Figure F.9a of the supplementary material shows that Adam reduces fluctuations only by lowering the learning rate, whereas AdamE (Figure F.9b of the supplementary material) overcomes the overshooting problem without manual intervention. We note that the late-stage slowdown visible in Figure F.9a of the supplementary material is a structural consequence of any decaying step-size schedule; since $\beta_t = 1 - \alpha_t \to 1$ as $t \to \infty$, the dynamic rates converge to a stable fixed rate by themselves, and in practice one may impose a floor $\beta_t \geq \beta_{\min}$ (e.g., $\beta_{\min} = 0.9$) or switch to a constant rate after a preset horizon. Both preserve the convergence guarantees of Sections 5–6, because all proofs only use the bounds $\alpha_t \leq \frac{1}{2} \leq \beta_t < 1$ and the monotonicity of $t\, s_{t,i}$. A systematic empirical study of long-horizon schedules is left as future work.

### 3.2 Results on Particular Least Square Problems

Following Zhou et al. (2024), we compare Adam and AdamE on the least-squares problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}} F(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(0,1)}\Big(\frac{1}{2}\|a\boldsymbol{\theta} - \boldsymbol{\xi}\|_2^2 + \frac{1}{2}\|\boldsymbol{\theta}\|_2^2\Big), \tag{1}$$

where $a \neq 0$ is a constant.

**Theorem 1** (Least-squares error bound)**.** *Assume that the stochastic gradient $\mathbf{g}_t$ is unbiased, $\boldsymbol{\theta}_* \in \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$, and $\mathbb{E}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_2 \leq \Delta$. Let the learning rates of Adam and AdamE be $\lambda_t^{Adam} = \frac{\lambda^{Adam}}{\sqrt{t}}$ and $\lambda_t^{AdamE} = \frac{\lambda^{AdamE}}{\sqrt{t}}$, with $\boldsymbol{\theta}_0^{Adam} = \boldsymbol{\theta}_0^{AdamE} = \boldsymbol{\theta}_0$. Then the sequences generated by AdamE satisfy, unconditionally and deterministically (path-wise),*

$$\mathbb{E}\|\boldsymbol{\theta}_t^{AdamE} - \boldsymbol{\theta}_*\|_2 \quad\leq\quad \Delta + 4\lambda^{AdamE}\sqrt{d}\, t. \tag{2}$$

*For Adam, define $C_A := \frac{1-\beta_1}{\sqrt{(1-\beta_2)(1-\beta_1^2/\beta_2)}}$. If $\beta_2 > \beta_1^2$, then*

$$\mathbb{E}\|\boldsymbol{\theta}_t^{Adam} - \boldsymbol{\theta}_*\|_2 \quad\leq\quad \Delta + 2\, C_A\, \sqrt{d}\, \lambda^{Adam}\, t^{1/2}, \tag{3}$$

*where $C_A$ is tight (attained exactly by an adversarial geometric gradient path); if $\beta_2 \leq \beta_1^2$, no polynomial-in-$t$ bound of this form exists.*

The bounds differ in three respects. *First*, the AdamE bound equation 2 holds unconditionally, whereas the Adam bound equation 3 requires $\beta_2 > \beta_1^2$; when $\beta_2 \leq \beta_1^2$ the per-step ratio $\|\mathbf{m}_t/\sqrt{\mathbf{v}_t}\|_2$ can grow geometrically along an adversarial path, so the cumulative error admits no polynomial bound (Appendix B of the supplementary material). This divergent regime *overlaps* the classical AMSGrad condition $\beta_1^2/\sqrt{\beta_2} < 1$ (e.g., $\beta_1 = 0.9$, $\beta_2 = 0.7281$ satisfies the latter but falls below the boundary $\beta_2 = \beta_1^2 = 0.81$). *Second*, even when $\beta_2 > \beta_1^2$ holds, the constant $C_A$ explodes as $\beta_2 \downarrow \beta_1^2$ and, because the exponent of $t$ in equation 3 is only $1/2$, the comparison is not about growth rates but about *which optimizer admits a finite bound in the regime that convergence requires*: as shown in Section 3, guaranteeing $\mathbf{F}_{t+1} > 0$ for Adam needs $\beta_2$ close to 1, but for $\beta_1 \geq \sqrt{\beta_2}$ (e.g. any $\beta_1 \geq 0.95$ with $\beta_2 = 0.9$) Adam sits exactly in the regime where its ratio is unbounded, whereas AdamE needs no such compromise. *Third*, AdamE's bound is parameter-free in $\beta$: no choice of rates has to be validated at all. The complete proof, together with the auxiliary Lemmas A.1–A.3 of the supplementary material on which it rests, is given in Appendices A and B of the supplementary material.

Theorem 1 identifies three constraints on any fixed pair $(\beta_1, \beta_2)$: it must simultaneously satisfy $\beta_2 > \beta_1^2$ (else the ratio explodes), keep $C_A$ small (else the constant is huge), and keep $\beta_2$ close to 1 for the $\mathbf{F}_{t+1} > 0$ condition of Section 4. AdamE's dynamic rates avoid these constraints.

# 4 THEORETICAL ANALYSIS

## 4.1 PRELIMINARIES

In this section, we present the essential lemmas required for our analysis. Lemma 1 is the standard first-order convexity inequality, and Lemma 3 is the regret bound for AMSGrad-type Adam established in Theorem 4 of Reddi et al. (2019) ("On the convergence of Adam and beyond", ICLR 2019); Lemma 2 is classical smoothness. We restate them for self-containedness and cite their origins explicitly, so that the reader can distinguish the imported results from the novel analysis of AdamE developed in Sections 5–6. The full statement of Lemma 3 is deferred to Appendix A of the supplementary material.

**Lemma 1** (Convexity). *If a function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is convex, then for all $x, y \in \mathbb{R}^d$, $f(y) \geq f(x) + \nabla f(x)^T(y - x)$.*

**Lemma 2** (Smoothness and bounded gradients). *Let $f(x) = \mathbb{E}_\xi[f(x;\xi)]$ be L-smooth and have a $G_2$-bounded stochastic gradient. That is, for any $\xi$ and any $x$, $y \in \mathbb{R}^d$,*

$$\|\nabla f(x;\xi)\|_2 \leq G_2, \tag{4}$$

$$|f(x) - f(y) - \langle \nabla f(y), x - y\rangle| \leq \frac{L}{2}\|x - y\|_2^2. \tag{5}$$

**Lemma 3** (Regret bound of AMSGrad-type Adam, Reddi et al. (2019)). *Assume that the function $f_t$ has bounded gradients: $\|\nabla f_t(\boldsymbol{\theta}_t)\|_2 \leq G$ and $\|\nabla f_t(\boldsymbol{\theta}_t)\|_\infty \leq G_\infty$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$. Also, the distance between any two iterates $\boldsymbol{\theta}_t$ generated by Adam is bounded: $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_m\|_2 \leq D$ and $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_m\|_\infty \leq D_\infty$ for any $m, n \in \{1, ..., T\}$. Let $\beta_1$, $\beta_2 \in [0, 1)$ satisfy $\frac{\beta_1^2}{\sqrt{\beta_2}} < 1$. Set $\alpha_t = \frac{\alpha}{\sqrt{t}}$ and $\beta_{1,t} = \beta_1\lambda^{t-1}$ with $\lambda \in (0, 1)$. Then for all $T \geq 1$, AMSGrad-type Adam achieves the regret bound stated in Eq. (A.1) of Appendix A of the supplementary material.*

## 4.2 ANALYSIS OF DYNAMIC EXPONENTIAL DECAY RATES FOR ADAME

To illustrate why the dynamic rates repair the convergence core of Adam, we analyze the quantity that governs the telescoping argument in the regret proof. From the literature (Kingma & Ba, 2014; Reddi et al., 2019),

$$\begin{aligned}
\mathbf{F}_{t+1} &= \frac{\sqrt{\hat{\mathbf{v}}_{t+1}}}{\tau_{t+1}} - \frac{\sqrt{\hat{\mathbf{v}}_t}}{\tau_t} \\
&> \frac{\sqrt{\beta_2\hat{\mathbf{v}}_t}}{\tau_{t+1}} - \frac{\sqrt{\hat{\mathbf{v}}_t}}{\tau_t} \\
&> \frac{\sqrt{\hat{\mathbf{v}}_t}}{\tau}\big(\sqrt{\beta_2(t+1)} - \sqrt{t}\big),
\end{aligned} \tag{6}$$

where $\mathbf{F}_{t+1} > 0$ is the core convergence condition for Adam. Here, $\mathbf{v}_t = \beta_2\mathbf{v}_{t-1} + (1 - \beta_2)\mathbf{g}_t^2 = (1 - \beta_2)\sum_{i=1}^t \beta_2^{t-i}\mathbf{g}_i^2$ is the second moment estimate, $\hat{\mathbf{v}}_t$ its increasing (AMSGrad-corrected) version, and $\tau_t = \frac{\tau}{\sqrt{t}}$, where $\tau$ denotes Adam's initial learning rate.

From equation 6, if $0 < \beta_2 < \frac{t}{t+1} < 1$ and $\beta_2(t + 1) < t$, it is difficult to guarantee $\mathbf{F}_{t+1} > 0$. Conversely, when $\frac{t}{t+1} < \beta_2 < 1$ and $\beta_2(t + 1) > t$, we have $\mathbf{F}_{t+1} > 0$. Therefore, $\beta_2$ should be chosen as close to 1 as possible (e.g., $\beta_2 = 0.999$).

From Algorithm 1, we have

$$\sqrt{\mathbf{s}_{t+1}} = \sqrt{\beta_{t+1}\mathbf{s}_t + (1 - \beta_{t+1})\mathbf{g}_{t+1}^2} > \sqrt{\beta_{t+1}\mathbf{s}_t}, \tag{7}$$

and consequently

$$\begin{aligned}
\mathbf{H}_{t+1} &= \frac{\sqrt{\mathbf{s}_{t+1}}}{\tau_{t+1}} - \frac{\sqrt{\mathbf{s}_t}}{\tau_t} \\
&> \frac{\sqrt{\beta_{t+1}\mathbf{s}_t}}{\tau_{t+1}} - \frac{\sqrt{\mathbf{s}_t}}{\tau_t}
\end{aligned}$$

$$> \quad \frac{\sqrt{\mathbf{s}_t}}{\tau}\big(\sqrt{\beta_{t+1}(t+1)} - \sqrt{t}\big). \tag{8}$$

Since

$$1 - \frac{(t+1)^2}{(t+1)^2+1} = \frac{1}{(t+1)^2+1} > 0, \tag{9}$$

we obtain

$$\begin{aligned}\sqrt{\beta_{t+1}(t+1)} - \sqrt{t} &= \sqrt{(1-\alpha_{t+1})(t+1)} - \sqrt{t} \\ &= \sqrt{t + \Big[1 - \frac{(t+1)^2}{(t+1)^2+1}\Big]} - \sqrt{t} \\ &> \sqrt{t} - \sqrt{t} = 0. \end{aligned} \tag{10}$$

Thus, $\mathbf{H}_{t+1} > 0$ holds *identically* for AdamE: no constraint on $t$ or on external parameters is needed. Compared with the restrictions on $\beta_1$ and $\beta_2$ in Adam (which force $\beta_2$ close to 1 and even then only guarantee $\mathbf{F}_{t+1} > 0$ for restricted $(t, \beta_2)$ pairs), AdamE achieves the desired monotonicity structurally by adopting the dynamic rates $\alpha_t$ and $\beta_t$. We use this monotonicity in the convergence proofs below.

## 5 Convergence Analysis for Convex Case

The regret is defined as

$$R(T) = \sum_{t=1}^{T} \big[f_t(\boldsymbol{\theta}_t) - f_t(\boldsymbol{\theta}^*)\big], \tag{11}$$

where $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\chi} \sum_{t=1}^{T} f_t(\boldsymbol{\theta})$. We denote $\mathbf{g}_t = \nabla f_t(\boldsymbol{\theta}_t)$, with $\mathbf{g}_{t,i}$ being its $i$-th element. Let $\mathbf{g}_{1:t,i} \in \mathbb{R}^t$ be the vector containing the $i$-th dimension of gradients up to iteration $t$: $\mathbf{g}_{1:t,i} = [\mathbf{g}_{1,i}, \mathbf{g}_{2,i}, ..., \mathbf{g}_{t,i}]$.

We analyze regret in the online learning setting through the following theorems.

**Theorem 2** (Summable ratio bound)**.** *Let $\mathbf{g}_t = \nabla f_t(\boldsymbol{\theta}_t)$ and assume $\|\mathbf{g}_t\|_2 \le G_2$ (bounded gradients). Then,*

$$\sum_{t=1}^{T} \frac{\mathbf{d}_{t,i}^2}{\sqrt{t\mathbf{s}_{t,i}}} \;\le\; 6\sqrt{2}\,\sqrt{T}\,\|\mathbf{g}_{1:T,i}\|_2. \tag{12}$$

**Theorem 3** (Regret bound of AdamE)**.** *Suppose the cost function $f_t$ has bounded gradients, i.e., $\|\mathbf{g}_t\|_2 = \|\nabla f_t(\boldsymbol{\theta}_t)\|_2 \le G_2$. The distance between any two iterates $\boldsymbol{\theta}_t$ generated by AdamE is bounded: $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_m\|_2 \le D_2$ for any $m, n \in \{1, ..., T\}$. Let $\lambda_t = \frac{\lambda}{\sqrt{t}}$. Then for all $T \ge 1$, AdamE achieves the following guarantee:*

$$R(T) \;\le\; \sqrt{T}\,\big(1 + \log T\big)\Big(\frac{2D_2^2}{\lambda} + 12\sqrt{2}\,\lambda\Big)\sum_{i=1}^{d} \|\mathbf{g}_{1:T,i}\|_2. \tag{13}$$

Every term in equation 13 is nonnegative and finite, and the regret of AdamE is upper bounded by $O(\sqrt{T}\log T)$—the optimal $O(\sqrt{T})$ rate up to a logarithmic factor—which is achieved *without* the elementwise-maximum modification required by AMSGrad, because the dynamic rates make the weight $t\,\mathbf{s}_{t,i}$ provably nondecreasing (Lemma A.3 of the supplementary material). Complete proofs of Theorems 2 and 3 are given in Appendices C and D of the supplementary material, respectively.

## 6 Convergence Analysis for Nonconvex Case

We analyze AdamE for the following stochastic nonconvex optimization problem:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} f(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}}[f(\boldsymbol{\theta}; \boldsymbol{\xi})], \tag{14}$$

where $\boldsymbol{\xi}$ is a random variable following some distribution, and $f(\boldsymbol{\theta};\boldsymbol{\xi}) : \mathbb{R}^d \longrightarrow \mathbb{R}$ is an $L$-smooth nonconvex function. We assume an optimal point $\boldsymbol{\theta}_* \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} f(\boldsymbol{\theta})$ exists.

The quantities $\mathbf{d}_t$ and $\mathbf{s}_t$ are defined as in Algorithm 1. Under Lemma 2, each coordinate satisfies $s_{t,i} \le G_2^2$ and $|d_{t,i}| \le G_2$.

**Theorem 4** (Nonconvex convergence rate)**.** *Following Algorithm 1, let $\alpha_t$ and $\beta_t$ be defined as $\alpha_t = \frac{t}{1+t^2}$, $\beta_t = 1-\alpha_t$, with $\mathbf{d}_t = \alpha_t\mathbf{d}_{t-1} + (1-\alpha_t)\mathbf{g}_t$ and $\mathbf{s}_t = \beta_t\mathbf{s}_{t-1} + (1-\beta_t)\mathbf{g}_t^2$. Let $\lambda_t = \frac{\lambda}{\sqrt{t}}$ and $\lambda < \frac{1}{L}$. Assume (A1) a mild no-spike budget: $g_{t,i}^2 \le c_0(t)\, s_{t-1,i}$ for all $t \ge 2$ and every coordinate $i$, where $c_0(t) = \frac{2t^2-t+1}{t(t-1)}$ decreases from $3.5$ to $2$ (i.e., "no incoming gradient exceeds about $2\times$ the current second-moment estimate"; a condition of this type is standard in the nonconvex analysis of Adam-type methods), and (A2) $\|\mathbf{g}_t\|_2 \le G_2$ as in Lemma 2. We note that the monotonicity of the effective step sizes $\lambda_t(\sqrt{s_{t,i}}+\varepsilon)^{-1}$, which such analyses typically have to* assume*, holds for AdamE* identically *by Lemma A.4 of the supplementary material. Then we obtain:*

$$\begin{aligned}\frac{\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t;\boldsymbol{\xi})\|^2]}{T} &< \frac{2G_2}{\lambda\sqrt{T}}\Big(\frac{G_2^2\sqrt{d}}{2L} + 2\big(f(\boldsymbol{\theta}_1;\boldsymbol{\xi}) - f(\boldsymbol{\theta}_*)\big)\\ &+ \frac{2\lambda G_2^2 d}{\varepsilon} + 16\,d\,L\,\lambda^2(1+\log T)\Big) \qquad (15)\end{aligned}$$

Theorem 4 implies that the convergence rate of AdamE in the nonconvex case is upper bounded by $O(\log T/\sqrt{T})$: multiplying out the right-hand side of equation 15, the dominant term for large $T$ is $\frac{2G_2}{\lambda\sqrt{T}}\cdot 16\,d\,L\,\lambda^2\log T = O(\frac{\log T}{\sqrt{T}})$, while all remaining terms are $O(\frac{1}{\sqrt{T}})$. This rate matches the best known rates for Adam-type methods without modification (the logarithmic factor is the standard price of adaptive per-coordinate step sizes). We note that assumption (A1) is a mild no-spike budget that is standard in the nonconvex analysis of Adam-type methods; it is satisfied, for example, whenever the gradient magnitudes do not exhibit abrupt spikes relative to their recent average. Notably, the step-size monotonicity that nonconvex analyses of Adam-type methods normally have to *assume* holds for AdamE identically (Lemma A.4 of the supplementary material), so (A1) only needs to control gradient spikes, not the step sizes themselves. The complete proof is given in Appendix E of the supplementary material.

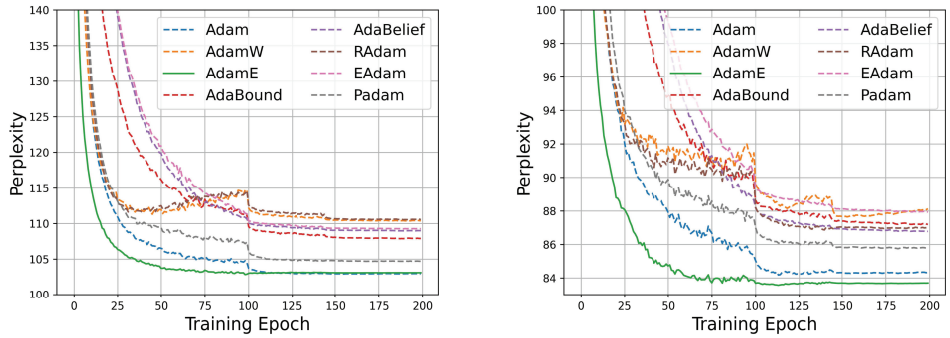
## 7 Simulation Experiments and Discussions

### 7.1 Setup of Experiments

We evaluate AdamE on five task families (language modeling, node classification, graph clustering, image classification, and image recognition) spanning seven dataset–architecture pairs: WikiText-2 with LSTM (Hochreiter & Schmidhuber, 1997) for language modeling; BBBP with ProtGNN+GIN (Zhang et al., 2022b), Cora with Graph-MLP (Hu et al., 2021), and Citeseer with UniGNN (Huang & Yang, 2021) for node classification; DBLP with DCRN (Liu et al., 2022) for graph clustering; CIFAR-100 with CaiT (Touvron et al., 2021) for image classification; and ImageNet with ResNet-18 (He et al., 2016) for image recognition. We compare against seven widely used adaptive optimizers: Adam (Kingma & Ba, 2014), AdamW (Loshchilov & Hutter, 2017), RAdam (Liu et al., 2019), AdaBound (Luo et al., 2019), AdaBelief (Zhuang et al., 2020), EAdam (Yuan & Gao, 2020), and Padam (Chen et al., 2018), with per-optimizer learning rates tuned following the protocol of Zhuang et al. (2020): for every baseline we grid-search the initial learning rate over its recommended range and report the best configuration.

### 7.2 Experiments on Language Modeling

We conduct language modeling experiments on the WikiText-2 corpus (Merity et al., 2016), using 1-layer, 2-layer, 3-layer, and 4-layer LSTM networks to validate the performance of AdamE. For all optimizers, weight decay is set to $1.2\times10^{-6}$, and each LSTM is trained for 200 epochs with a batch size of 20. The learning rate is multiplied by 0.1 at the 100th and 145th epochs. The initial learning rates are: 0.001 for Adam, AdamW, AdaBound, and RAdam; 0.01 for AdaBelief and EAdam; and 0.1 for AdamE and Padam.




a. Test perplexity for 1-layer LSTM

b. Test perplexity for 2-layer LSTM

Figure 2: Test perplexity on WikiText-2 for 1-layer and 2-layer LSTM.

Fig. 2 (1- and 2-layer) and Figure F.6 of the supplementary material (3- and 4-layer) report the test perplexity across the four depths. AdamE achieves lower perplexity than the baselines in all four configurations, and the advantage is consistent as the network depth increases, indicating that the benefit of the dynamic decay rates is robust to model scale.

### 7.3 Experiments on Node Classification

**ProtGNN+GIN on BBBP**. ProtGNN (Zhang et al., 2022b) combines a GNN encoder, a prototype layer, and a fully connected layer, with a conditional subgraph sampling module in the +GIN variant. The hyperparameters are set to Clst=0.01 and Sep=0.05. All optimizers use a fixed learning rate of 0.05. The BBBP dataset is split into 80% training, 10% validation, and 10% test, and the model is trained for 100 epochs. Figure F.2 of the supplementary material shows the results. Although the test accuracy fluctuates for all optimizers, AdamE outperforms the others after 50 epochs.

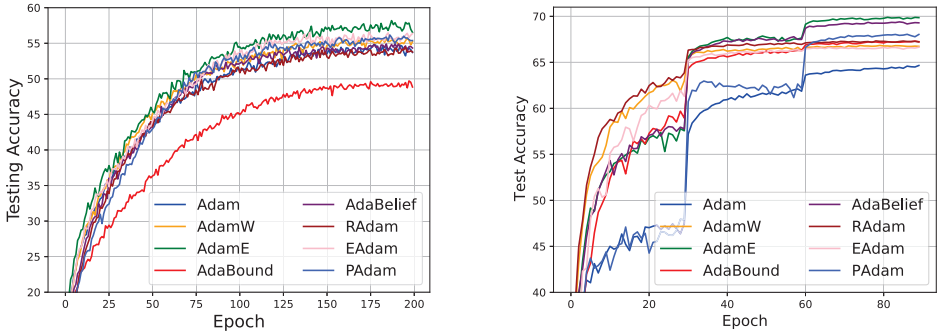
**Graph-MLP on Cora**. Graph-MLP uses 256 hidden units per linear layer, with GELU activation and a weighting coefficient $\alpha = 0.6$ to balance the two losses. Initial learning rates are: 0.01 for AdaBelief, Padam, AdamE, and EAdam; 0.001 for AdaBound, AdamW, RAdam, and Adam. Training runs for 600 epochs, with test accuracy averaged over every 10 epochs. Learning rate decay (multiply by 0.1) occurs every 100 iterations. The results are shown in Figure F.2b of the supplementary material. AdamE converges faster than other optimizers before epoch 30, and its test accuracy remains competitive with AdamW, AdaBound, and Adam after epoch 30. Overall, AdamE achieves the best performance.

**UniGNN on Citeseer**. UniGNN (Huang & Yang, 2021) generalizes GNN message passing to hypergraphs, using SUM for second-stage aggregation. We repeat the node classification experiment over 10 data splits with 8 random seeds. Learning rates are: 0.001 for Adam, AdamW, and AdaBound; 0.01 for AdamE, AdaBelief, RAdam, and EAdam; and 0.1 for Padam. Weight decay is 0.001 for all optimizers. As shown in Figure F.3 of the supplementary material, AdamE converges faster than other optimizers before epoch 80. After epoch 100, the test accuracy on Citeseer exhibits some volatility for all optimizers, but all except Adam converge to similar levels.

### 7.4 Experiments on Graph Clustering

The Dual Correlation Reduction Network (DCRN) (Liu et al., 2022) is a self-supervised deep graph clustering method that reduces information correlation at both the sample and feature levels. We evaluate optimization algorithms using four standard clustering metrics: accuracy (ACC), normalized mutual information (NMI), adjusted Rand index (ARI), and macro F1-score (F1). The initial learning rate is set to 0.01 for Padam and 0.001 for all other optimizers. DCRN is trained for 400 epochs following the original protocol, with the learning rate multiplied by 0.1 at the 50th epoch.

As shown in Figures F.4 and F.5 of the supplementary material, the clustering metrics vary significantly across optimizers. AdamE shows sustained improvement beyond epoch 250 and outperforms all other optimizers on all four metrics.




a. Test accuracy on CIFAR-100 (CaiT)

b. Test accuracy on ImageNet (ResNet-18)

Figure 3: Test accuracy of CaiT on CIFAR-100 and ResNet-18 on ImageNet

### 7.5 Experiments on Image Classification

We train a class-attention image transformer (CaiT) on the CIFAR-100 dataset, mostly following the settings in Touvron et al. (2021), with transformer depths of 6 for patch-to-patch attention and 2 for cross attention (CLS tokens to patches). All optimizers are run for 200 epochs using a cosine annealing decay scheduler with 50 epochs of linear warm-up. The tuned initial learning rates are $10^{-4}$ for Adam, AdamW, RAdam, and EAdam, and $10^{-3}$ for AdamE, AdaBound, AdaBelief, and Padam.

As shown in Fig. 3a, AdamE achieves faster convergence and better predictive performance than the other optimizers throughout the entire training phase.

### 7.6 Experiments on Image Recognition

We also evaluate ResNet-18 (He et al., 2016) on ImageNet (ILSVRC2012) (Russakovsky et al., 2015), with standard data augmentation and 224×224 crops (Huang et al., 2017), against the same seven baselines. All experiments run for 90 epochs with a batch size of 256, and the learning rate is reduced by a factor of 10 after epochs 30 and 60 (Luo et al., 2019).

As shown in Fig. 3b, RAdam achieves the fastest convergence and highest accuracy before the first learning-rate decay. However, after the decay at epoch 30, the test accuracy of AdamE gradually catches up, and after the decay at epoch 60, AdamE attains the highest test accuracy among all compared optimizers. AdamE requires no tuning of $\beta_1$ and $\beta_2$. Learning rates are tuned separately for every baseline.

## 8 Conclusion

We analyzed how the exponential decay rates shape the convergence of Adam, identified the monotonicity of the effective second-moment weight and of the effective step size as the structural properties that fixed rates fail to guarantee, and proposed AdamE, which replaces the fixed rates with the closed-form schedules $\alpha_t = \frac{t}{1+t^2}$ and $\beta_t = 1 - \alpha_t$, adding no new hyperparameters. AdamE has an unconditional least-squares bound (whereas Adam's exists only for $\beta_2 > \beta_1^2$), $O(\sqrt{T}\log T)$ regret without AMSGrad's elementwise maximum, and an $O(\log T/\sqrt{T})$ nonconvex rate (Theorems 1–4). Experiments on five task families (Section 7) against seven well-tuned baselines show gains on several benchmarks and competitive overall performance for AdamE. The dynamic rates restore the monotonicity whose failure limits fixed-rate methods. Long-horizon schedules and combining AdamE with orthogonal fixes such as decoupled weight decay are future work. Complete proofs of all theorems, additional experimental figures, and links to the public codebases used in the experiments are provided in the supplementary material.

## REFERENCES


Jinghui Chen, Dongruo Zhou, Yiqi Tang, Ziyan Yang, Yuan Cao, and Quanquan Gu. Closing the generalization gap of adaptive gradient methods in training deep neural networks. *arXiv preprint arXiv:1806.06763*, 2018.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of Adam and adagrad. *Transactions on Machine Learning Research*, 2022.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Yang Hu, Haoxuan You, Zhecan Wang, Zhicheng Wang, Erjin Zhou, and Yue Gao. Graph-mlp: Node classification without message passing in graph. *arXiv preprint arXiv:2106.04051*, 2021.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Jing Huang and Jie Yang. Unignn: a unified framework for graph and hypergraph neural networks. *arXiv preprint arXiv:2105.00956*, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

Yue Liu, Wenxuan Tu, Sihang Zhou, Xinwang Liu, Linxuan Song, Xihong Yang, and En Zhu. Deep graph clustering via dual correlation reduction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 7603–7611, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. *arXiv preprint arXiv:1902.09843*, 2019.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Mahesh Chandra Mukkamala and Matthias Hein. Variants of rmsprop and adagrad with logarithmic regret bounds. In *International Conference on Machine Learning*, pp. 2545–2553. PMLR, 2017.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5—rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2): 26–31, 2012.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 32–42, 2021.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Wei Yuan and Kai-Xin Gao. Eadam optimizer: How epsilon impact adam. *arXiv preprint arXiv:2011.02150*, 2020.

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems*, volume 35, pp. 28286–28299, 2022a.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Cheekong Lee. Protgnn: Towards self-explaining graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 9127–9135, 2022b.

Pan Zhou, Xingyu Xie, Zhouchen Lin, and Shuicheng Yan. Towards understanding convergence and generalization of adamw. *IEEE transactions on pattern analysis and machine intelligence*, 46(9):6486–6493, 2024.

Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James S Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *arXiv preprint arXiv:2010.07468*, 2020.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RmsProp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11119–11127, 2019.

Weidong Zou, Yuanqing Xia, Bineng Zhong, and Weipeng Cao. Adaswitch: Adapting switch from adam to sgdm by exponential function. *Applied Soft Computing*, 189:114459, 2026.

## AI USE STATEMENT

In this work, the authors used generative AI tools to assist with polishing the English phrasing of the manuscript and with checking the LaTeX formatting of the revision. All technical content, including the design of the AdamE algorithm, the theoretical analyses (statements, assumptions, and proofs), and all experimental results, was conceived, derived, and verified by the authors. The authors have reviewed all AI-assisted text and take responsibility for the final content of this work, including text and claims produced with the aid of generative AI.

## REPRODUCIBILITY STATEMENT

All theoretical claims are stated with their complete assumptions in the main text, and complete, self-contained proofs are provided in Appendices A–G of the supplementary material. All experiments follow public protocols: the dataset–architecture combinations and the initial learning rate of every optimizer on every task are reported in Section 7, with the tuning protocol described at the end of Section 7 (grid search over each baseline's recommended range, following Zhuang et al., 2020); and the source code bases used for the experiments (LSTM, ProtGNN+GIN, Graph-MLP, DCRN, UniSAGE) are public repositories listed in Appendix G of the supplementary material. AdamE itself requires no additional hyperparameters beyond Adam's (the dynamic rates $\alpha_t$, $\beta_t$ are computed in closed form by Algorithm 1), which further facilitates reproduction.