UTA: UNIVERSAL TRAINING ADAPTIVE LEARNING RATE DECAY

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning rate decay (LRD) is critical for training high-performance models, yet most designs remain empirically driven. Although the Universal Budget Aware (UBA) LRD offers a theoretical framework, its reliance on SGD dynamics limits its applicability to modern Adam-type optimizers and leaves its key 'optimization difficulty parameter' opaque. To bridge this gap, we propose the Universal Training Adaptive (UTA) LRD. Unlike UBA's fixed offline trajectory anchored solely to iterations, UTA treats LRD as a closed-loop shape-and-scale problem. By continuous-time formulation, UTA is training-state adaptive with segment-wise optimal shape. This adaptivity links external iterations with internal training states, enabling a near-optimal path in both progression shape and learning rate (LR) scale, a property unattainable by other LRDs, including UBA. Empirically, we demonstrate consistent improvements across language models (LLaMA 60M to 1B) and vision models (ResNet50, ViT-B) on ImageNet with plug-and-play ease and negligible overhead. Extensive ablation studies further confirm UTA's robustness across dimensions, including peak LR, final decay ratio, its parameter φ_0, component efficacy and generalization to fine-tuning scenarios, underscoring its practical implementability.

1 INTRODUCTION

The learning rate decay (LRD) plays a fundamental role in training high-performance deep learning models (Tang et al., 2026; Tissue et al., 2025; Li et al., 2025). Its design directly governs the convergence behavior, generalization capability, and final performance of models across a wide spectrum of tasks and architectures. Notably, the influence of LRDs becomes even more pronounced in resource-constrained or time-sensitive scenarios, where the number of training iterations is limited (Hägele et al., 2024; Tang et al., 2026). These advantages underscore the practical importance of LRDs not only as a tuning parameter but as an essential lever for sustainable and scalable model development.

Research in this area has evolved along two main lines. A line of work (He et al., 2016; Smith, 2017; Smith & Topin, 2019; Li et al., 2019a; Chen et al., 2022; Pan et al., 2022a; Zhai et al., 2022; Hägele et al., 2024; Hu et al., 2024) focus on designing LRDs primarily through empirically driven approaches, effectively enhancing accuracy. Another line of research is theoretically driven, mainly seeking to establish frameworks and scaling laws for LRDs to provide theoretical explanation for phenomena (Pan et al., 2022b; Defazio et al., 2023; Tissue et al., 2025; Li et al., 2025; Luo et al., 2025; Tang et al., 2026). Notably, Tang et al. propose a theoretically grounded LRD called UBA (Tang et al., 2026) to narrow this gap. This framework adopts a min-max formulation that accounts for robustness to variations in loss landscape curvature, thereby guaranteeing performance under worst-case conditions. The UBA LRD mitigates heuristic design while maintaining strong and consistent performance across diverse tasks, architectures, model scales, and training budgets. Consequently, it removes the costly trial-and-error process typically required to select an appropriate LRD for a given network and training setup.

Despite its merits, UBA remains confined to its design framework which is analyzed in the context of Stochastic Gradient Descent (SGD). It raises questions about its generalization to adaptive optimizers such as Adam (Kingma & Ba, 2014). Moreover, although UBA establishes a connection to optimization difficulty via the local Hessian, its pivotal parameter φ is itself invariant to the internal dynamics

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.

of networks. Thus it lacks real-time feedback from the training state. This disconnect between the predetermined LRD and the dynamic training trajectory represents a key limitation in current practice. These limitations prompt a natural question:

Can we explain the pivotal parameter φ in universal budget aware LRD, further closing the gap between principle and practice?

In this work, we take a step toward this goal by extending the theoretical foundation of UBA to the Adam optimizer, aiming to combine theoretical soundness with practical, state-aware adaptability.

To achieve this, we propose the **U**niversal **T**raining LR under **A**daptive moment estimation (UTA) LRD. First, it is theoretically grounded, built upon a principled robust optimization framework. Moreover, it is dynamically training adaptive, adjusting stepwise learning rates (LRs) based on the real-time optimizer states during training. This stands in contrast to LRDs that predefine a fixed curve for LR variation. Additionally, it is inherently optimizer-agnostic. Although this paper focuses on its implementation with the Adam, the underlying framework generalizes to other optimizers.

To solve this extended optimization problem, we introduce a segment-wise continuous analysis. This approach offers three advantages. First, it yields the optimal shape of LRD curve. Second, it naturally gives rise to the concept, intrinsic time, defined as the cumulative LR effort. This quantity establishes a mapping between the external world of physical iteration count and an internal world of model optimization, serving as a more faithful measure of training progress. The continuous analysis further describes the energy evolution of optimization progress along intrinsic time, which gives the dynamic setting of the proposed LRD. Consequently, the derived LRD is near-optimal in shape and energy-calibrated in scale. Practically, the proposed LRD is plug-and-play, requiring minimal modification to existing model architectures or training pipelines, thus enabling easy adoption across diverse deep-learning frameworks.

Empirically, UTA delivers consistent improvements across both language (LLaMA-60M to 1B) and vision (ResNet50, ViT-B on ImageNet) tasks. Moreover, comprehensive ablation studies substantiate its practical advantages: (i) robustness to hyper-parameters including peak LR, final decay ratio, and φ_0; (ii) the effect of its state-aware mechanism; and (iii) effective generalization to parameter-efficient fine-tuning scenarios.

Main contributions:

- We propose UTA, a novel generalized LRD framework, which extends the principled UBA framework from SGD to adaptive optimizers, such as AdamW, thereby overcoming a key limitation of prior work that restricts its applicability.
- We introduce the concept of intrinsic time as a unified measure of training progress, establishing a mapping from external iteration counts to internal optimization dynamics. This perspective not only reveals the fundamental nature of UBA's pivotal parameter φ by identifying it as the scale of intrinsic time, but also provides a principled foundation for dynamically determining the LRs.
- We construct a theoretically grounded mechanism that combines a segment-wise analysis and an intrinsic-time-based energy evolution, coupling the optimal LRD shape with real-time optimization states.
- We conduct extensive experiments across language (LLaMA-60M to 1B) and vision (ResNet50, ViT-B on ImageNet) tasks, along with comprehensive ablations on hyper-parameter sensitivity, dynamic component efficacy, automatic LR methods, and fine-tuning transferability. These results confirm UTA's consistent performance gains, robustness, and plug-and-play deployability with negligible overhead.

The remainder of this paper is organized as follows. Section 2 is the preliminary. Section 3 presents the methodology. Section 4 reports experimental results. Section 5 concludes with a discussion and directions for future work. A comprehensive review of related work is provided in Appendix J. Theoretical properties is provided in Appendix G.

2 PRELIMINARY

Robust optimization of learning rate. Let $f(W, \xi)$ denote a neural network parameterized by parameters $W \in \mathbb{R}^N$ on the data ξ. ξ denotes the data sampling on the dataset D, and let $\mathcal{F}$ be a function class that incorporates the variation in multiple factors, such as the network structures and data sets. Let $\mathcal{L}$ be the loss function, T be a maximum number of training iterations, and t be the current learning step. UBA (Tang et al., 2026) establishes a robust optimization framework for learning rate (ROLR) under SGD, formulated as:

$$\begin{aligned} \min_{\eta_1,\eta_2,\ldots,\eta_{T-1}} \max_{f\in\mathcal{F}} \quad & \mathcal{L}(f(W_T,\xi)), \\ s.t. \quad & W_{t+1} = W_t - \eta_t \nabla \mathcal{L}(f(W_t,\xi)), \end{aligned} \tag{1}$$

where $t = 0, 1, 2, \cdots, T-1$. This min-max formulation guarantees performance in worst-case scenarios. However, this formulation constrains the feasible region to the standard gradient update process, limiting its direct compatibility with adaptive optimizers like Adam.

UBA-equivalent family. This family, as defined in Definition 2.1, encapsulates the optimal LRDs for gradient descent under a robust analysis (Tang et al., 2026). Specifically, considering the k-th segment $[T_k, T_{k+1}]$ with length ΔT_k, the optimal learning rate at the t-th step is given by the following instance of the UBA-equivalent family $\eta_t := g_{a,b}(\varphi, \theta_{t,k})$, where $\theta_{t,k} := ([2(t-T_k)-1]\pi/2(T_{k+1}-T_k)) + (k-1)\pi$, the range parameters $a = \eta_{\min}^{(k)}$ and $b = \eta_{\max}^{(k)}$ correspond to the minimal and the maximal step sizes, respectively. $(k-1)\pi$ introduces a phase shift across segments. The parameter φ controls the curvature of the LRD, related to optimization difficulty by the Hessian of local minima. However, φ is static and expertise prior-driven.

Definition 2.1 (UBA-equivalent family)**.** For a curvature parameter $\varphi > 0$, range parameters $b, a \geq 0$, a step index t and a period parameter k, the *UBA-equivalent family* is defined as

$$g_{a,b}(\varphi, \theta_{t,k}) := b + a\frac{2\left[1+\cos(\theta_{t,k})\right]}{2\varphi + (2-\varphi)\left[1+\cos(\theta_{t,k})\right]}. \tag{2}$$

3 METHODOLOGY

3.1 PROBLEM STATEMENT

We extend the ROLR equation 1 to a generalized update rule governed by optimizer-specific dynamics. The global min-max view is

$$\begin{aligned} \min_{\eta_1,\eta_2,\ldots,\eta_{T-1}} \max_{f\in\mathcal{F}} \; & \mathcal{L}\left(f(W_T,\xi)\right) \\ \text{s.t.} \quad & W_{t+1} = W_t - \eta_t \cdot P(t)\nabla\mathcal{L}(W(t),\xi), \end{aligned} \tag{3}$$

where $P(t)$ captures the optimizer-specific state dynamics, such as scaling or preconditioning, and may be chosen to represent different optimization algorithms within the same generalized framework. For instance, SGD corresponds to $P(t) = I$. Adam is recovered by defining $P(t)$ such that $P(t)\nabla\mathcal{L}(W(t)) = m_t/(\sqrt{v_t}+\epsilon)$, where m_t and v_t are the first and second moment estimates following Adam update rules. Other adaptive methods like AdaGrad or Conda (Wang et al., 2025) can be represented analogously through appropriate definitions of $P(t)$ based on their state variables. This paper primarily focuses on the case where $P(t)$ implements the Adam dynamics (See details in Appendix C).

3.2 SEGMENT-WISE CONTINUOUS ANALYSIS

There is a challenge in solving the problem equation 3 due to the exponential moving averages nature of the optimization process, where the loss landscape and gradient statistics of Adam evolve over time.

To capture this nature yet retain tractability, we analyze training dynamics through their continuous-time limit rather than discrete updates following prior work (Li et al., 2017; 2019b; 2020; 2021; 2025). This perspective offers crucial advantages. First, we can make small segment-wise approximation of Adam by preconditioned GD. In detail, we partition the problem into K segments $[T_k, T_{k+1})$, assuming local approximations for the curvature in each segment.

Segment-wise decomposition. We partition the training horizon $[0, \mathcal{T}]$ into K segments $0 = T_0 < T_1 < \cdots < T_K = \mathcal{T}$, and focus on the k-th segment $[T_k, T_{k+1}]$. For ease of theoretical analysis, we transition from the discrete interval $[T_k, T_{k+1}]$ to a continuous representation. The local dynamics could be approximately analyzed, allowing us to model Adam as a preconditioned gradient flow with bounded approximation error (will be further explained in Paragraph Optimal Shape of Problem Solution3.2). Consequently, the adaptive Adam updates can be analyzed through the simpler lens of preconditioned gradient descent within each short segment.

Definition 3.1 (intrinsic time)**.** For a segment $[T_k, T_{k+1})$, define $U_k := \int_{T_k}^{T_{k+1}} \eta(t)\,\mathrm{d}t$, where $\eta(\tau)$ denotes the continuous-time learning rate.

Intrinsic time. We introduce a key reparameterization of the training timeline via a continuous change of variables. The quantity U_k has a natural interpretation as a *time reparameterization*. Consider the mapping $t = T(\tau) := \int_{T_k}^{\tau} \eta(z)\,\mathrm{d}z$, which transforms physical time (iteration index τ) into an intrinsic time t that reflects the internal optimization progress of model. Thus, U_k represents the total intrinsic duration of the segment, providing a more accurate measure of training advancement than the raw iteration count.

Optimal shape of problem solution. We characterize the optimal *shape* of the stepsizes within each segment of Problem 3. Lemma D.1 controls the segment radius in terms of the intrinsic time U_k. Lemma D.3 shows that on such short segments Adam bias-corrected momentum $\hat{m}_t$ tracks the true gradient with a small error. These results validate a piecewise preconditioned gradient approximation and enable Theorem 3.2, which identifies the UBA-equivalent family as the optimal segment-wise stepsize shape.

Theorem 3.2 (Optimal UBA-equivalent shape on a segment)**.** *Consider the min-max problem 3 on segment k, and assume that the step sizes are constrained to lie in the interval $[\eta_{\min}^{(k)}, \eta_{\max}^{(k)}]$. There exists an optimal shape of step sizes that minimizes the worst-case loss. This optimal shape belongs to the UBA-equivalent family for some curvature parameter $\varphi_k > 0$.*

The proof can be found in Appendix D.

For segment k, Theorem 3.2 specifies that the *shape* of the optimal step sizes is a member of UBA-equivalent Family. However, the precise shape of the LRD remains unknown since the key factor φ is still free. To close this gap, in the next subsection, we determine the overall scale by introducing the intrinsic time, i.e., $U_k := \int_{T_k}^{T_{k+1}} \eta(\tau)\,\mathrm{d}\tau$, to specify φ and completing the optimal LRD.

3.3 Regularized energy evolution via intrinsic time

This section establishes a theoretical control for scales of segment-wise LRD. We first model the training dynamics within each segment as a stochastic dynamical system following (Li et al., 2020; 2025). From this stochastic differential equations (SDEs) formulation, we derive a bias-variance-style upper bound on the expected energy evolution (Theorem 3.3). Minimizing this bound naturally leads to regularized problems: the ℓ_1-regularized formulation (Corollary 3.4) under a bounded LR constraint; the ℓ_2-regularized surrogate (Corollary 3.5) under a near-constant step-size approximation. Both problems admit explicit-form solutions, which yields the optimal intrinsic time U_k^*. This solution inspires us to design the quantity that governs the scale of the UBA-shaped LR throughout the segment (see Appendix E for details).

Denote $\nabla\mathcal{L}(W_t) = \mathbb{E}_{\xi_t}\nabla\mathcal{L}(W_t, \xi_t)$, we have $\nabla\mathcal{L}(W_t, \xi_t) = \nabla\mathcal{L}(W_t) + \delta_t$, where the gradient noise δ_t satisfies $\mathbb{E}(\delta_t) = 0$. Since the segment is short, the stochastic preconditioned gradient flow can be modeled in accordance with SGD noise (Li et al., 2025) scaling by $\mathrm{d}W(t) = -\eta(t)\,P(t)\nabla\mathcal{L}(W(t))\,\mathrm{d}t + \ \eta(t)\,P(t)\Sigma^{1/2}\,\mathrm{d}B(t)$, where $W(t)$ is the weight parameter, $\Sigma \succeq 0$ is the covariance of the gradient noise and $B(t)$ is a standard Brownian motion (see Appendix E for details).

Under a quadratic expansion measurement, we define the energy function between the t-th step weights and local minimum within segment $E(t) := \frac{1}{2}\|e(t)\|^2$, $e(t) := H_k^{1/2}\left(W(t) - \bar{W}^{(k)}\right)H_k^{1/2}$, where H_k is the abbreviation for $H_f^{(k)}$, the Hessian matrix at local minima (the precise construction is detailed in Appendix E).

With the above setup, we obtain the following bias-variance-style bound expressed in terms of the intrinsic time.

Theorem 3.3. *Consider the k-th segment $[T_k, T_{k+1}]$ and the preconditioned stochastic gradient flow equation 21, the expected energy at the segment end satisfies the bias–variance bound*

$$\mathbb{E}[E(T_{k+1})] \;\le\; \mathbb{E}[E(T_k)]\, e^{-2\nu_\ell^{(k)} U_k} \;+\; \frac{1}{2}\alpha \int_{T_k}^{T_{k+1}} \eta^2(t)\, dt, \tag{4}$$

where $\alpha := \sup_{t\in[T_k,T_{k+1}]} \mathrm{tr}(H_k^{1/2} P(t) \Sigma P(t) H_k^{1/2})$ is the noise intensity bound, and $\lambda_{\min}(M(t)) \ge \nu_\ell^{(k)} > 0, M(t) := H_k^{1/2} P(t) H_k^{1/2}$ is the spectral lower bound.

The proof can be found in Appendix E.

Corollary 3.4 (ℓ_1-regularized intrinsic-time problem)**.** *Under Theorem 3.3 and the constraint $0 \le \eta(t) \le \eta_{\max}$ on $[T_k, T_{k+1}]$, the intrinsic-time selection problem reduces to the ℓ_1-regularized problem,*

$$\min_{U\in[0,\eta_{\max}\Delta T_k]} \Big\{ \mathbb{E}[E(T_k)] e^{-2\nu_\ell^{(k)} U} + \frac{1}{2}\alpha\eta_{\max} U \Big\}. \tag{5}$$

Its minimizer admits the closed form $\min\left(\max\left(0, \frac{1}{2\nu_\ell^{(k)}} \ln\left(\frac{4\nu_\ell^{(k)}\mathbb{E}[E(T_k)]}{\alpha\eta_{\max}}\right)\right), \eta_{\max}\Delta T_k\right)$.

The proof can be found in Appendix E.

Corollary 3.5 (ℓ_2 surrogate via near-constant stepsizes)**.** *Under Theorem 3.3 and continuity of UBA-shape function, the intrinsic-time selection problem can be approximated as ℓ_2-regularized problem,*

$$\min_{U_k\in[0,\eta_{\max}\Delta T_k]} \mathbb{E}[E(T_k)] e^{-2\nu_\ell^{(k)} U_k} + \frac{1}{2}\alpha \frac{U_k^2}{\Delta T_k}, \tag{6}$$

The minimizer is given by the Lambert-W (Corless et al., 1996) expression $\min\left(\max\left(0, \bar{U}\right), \eta_{\max}\Delta T_k\right)$, *where* $\bar{U} := \frac{1}{2\nu_\ell^{(k)}} W\left(\frac{2(\nu_\ell^{(k)})^2 \mathbb{E}[E(T_k)]\Delta T_k}{\alpha}\right)$.

The proof can be found in Appendix E.

A computable plug-in estimate of U_k^*. Both closed-form solutions for U_k^* in Corollaries 3.4 and 3.5 are dominated by the ratio $\frac{\nu_\ell^{(k)}\mathbb{E}[E(T_k)]}{\alpha}$. We retain this dominant part to preserve the leading-order asymptotic behavior of the theoretical solutions while relying only on quantities observable during training.

Since $\nu_\ell^{(k)}$, $\mathbb{E}[E(T_k)]$, and α are not directly observable, we adopt a plug-in estimate with measurable surrogates. The energy term is approximated by the current training loss, $\mathbb{E}[E(T_k)] \approx \mathcal{L}(W_{T_k})$, following the local quadratic definition of E. The spectral lower bound is approximated by the Adam preconditioned gradient, $\nu_\ell^{(k)} T_k \approx \frac{2E(T_k)}{\|\hat{m}_t/(\sqrt{\hat{v}_t}+\epsilon)\|^2}$. The noise intensity α is treated as a tunable hyper-parameter. A direct substitution yields an inverse-quadratic dependence on the preconditioned update norm, which becomes numerically unstable when $\|\cdot\|$ is small. We therefore replace $\|\cdot\|^{-2}$ with the surrogate $(1 + \|\cdot\|)^{-1}$, which preserves monotonicity and asymptotic behavior while remaining bounded in $(0, 1]$. This gives the stabilized plug-in estimate

$$U_t^* = 1/\left(1 + \sqrt{\varphi_0/2}\, \left\| \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \right\| \right), \tag{7}$$

where φ_0 is a hyper-parameter and $U_k^* = \inf_{t\in[T_k,T_{k+1})} 1/(1 + \sqrt{\varphi_0/2}\, \left\| \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} \right\|)$. This estimate preserves the dominant asymptotic behavior of the theoretical solutions while depending only on quantities available in standard adaptive optimizers (see Appendix E for details).

3.4 DERIVED LRD

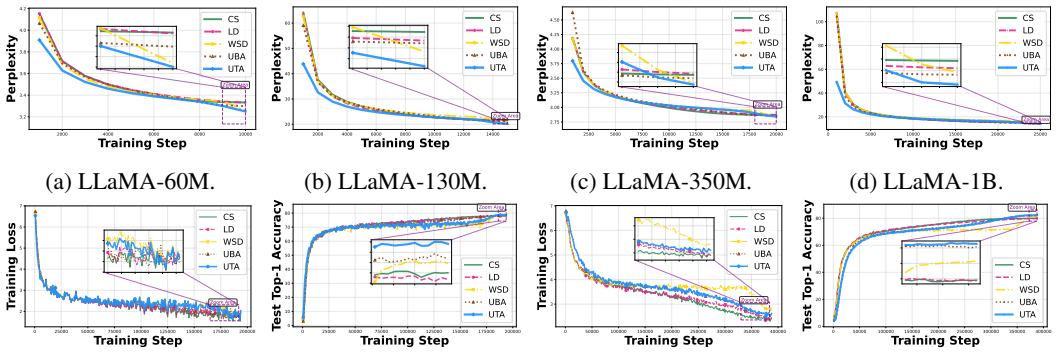
The optimal solution of problem equation 3 is given by a member of UBA-equivalent family (Definition 2.1) where φ_k is unknown. In this subsection, we complete the UTA design by *matching* the energy-optimal intrinsic time U_t^* equation 7, with the intrinsic time implied by the UBA-equivalent shape , i.e., $U_{\varphi_k} = \int_{T_k}^{T_{k+1}} g_{a,b}\left(\varphi, \theta_{t,k}\right) \mathrm{d}t$. Concretely, for the desired parameterization of the shape, such as fixing $b = \eta_{\min}^{(k)}$ and $a = \left(\eta_{\max}^{(k)} - \eta_{\min}^{(k)}\right)$, solving for φ_t under iteration-wise extension, we obtain the scalar equation $U_{\varphi_t} = U_t^*$ (the details of integral are in Appendix F). Then we have $\varphi_t = \varphi_0 \left\| \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} \right\|^2$. This yields a per-segment effective LRD that is (i) near-optimal in shape for 3, by Theorem 3.2, and (ii) energy-calibrated in scale, by subsection 3.3. Furthermore, we extend the LRD on entire horizon and keep a controlled approximation for the original dynamics (see Appendix F). The derived LRD on the entire horizon $[0, T]$ admits the explicit form

$$\eta(t) = g_{a,b}(\varphi_t, \theta(t)), \tag{8}$$

where $a = \eta_{\min}$, $b = \eta_{\max} - \eta_{\min}$, $\theta(t) := \frac{\pi \cdot t}{T}$, the curvature parameter $\varphi_t = \varphi_0 \left\| \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} \right\|^2$.

3.5 COMPLEXITY ANALYSIS

The computational overhead of UTA is negligible in practice. Each update involves merely 8-16 scalar arithmetic operations and a single norm computation. The cost is orders of magnitude lower than the FLOPs required for standard forward and backward propagation in modern neural networks.



(a) LLaMA-60M. (b) LLaMA-130M. (c) LLaMA-350M. (d) LLaMA-1B.

(e) ImageNet-ResNet50 (f) ImageNet-ResNet50. (g) ImageNet-ViT-B. (h) ImageNet-ViT-B.

Figure 1: Training curves of LLaMA models, ResNet-50 and ViT-B.

4 EXPERIMENTS

We conduct a systematic evaluation of the proposed LRD along three key axes of diversity: (i) **model architecture**, covering both convolutional (ResNet(He et al., 2016)) and attention-based (ViT(Dosovitskiy et al., 2020)) vision models, as well as language models (LLaMA); (ii) **task modality**, encompassing vision and language tasks; and (iii) **training scale**, covering the large-scale training (ImageNet and LLM pre-training from 60M to 1B). This multi-faceted evaluation is designed to rigorously assess the generality, robustness, and scalability of our method.

Baselines. Our baseline selection is informed by recent systematic evaluations of LRDs. The study (Tang et al., 2026) performs comprehensive empirical comparisons of LRDs. It shows that the UBA is a consistently top-performing LRD, and identifies cosine and linear decay as competitive approaches. Accordingly, we include **UBA** (Tang et al., 2026) as a strong latest baseline, and include **Cosine(CS)** (Loshchilov & Hutter, 2017) and **Linear decay(LD)** (Li et al., 2019a) as representative classical baselines in our evaluation. To further strengthen the comparison, we also include the **WSD** (Zhai et al., 2022; Hu et al., 2024) LRD, a widely used and empirically validated strategy in large-scale

language model pre-training, as an additional strong baseline representative of modern practices. The LRD-specific hyper-parameters and details are provided in Appendix H.1 and H.2.

Experimental details. All baselines follow **the same standard configurations** as established in prior work on LLaMA (Lialin et al., 2023; Zhao et al., 2024), ResNet (Wightman et al., 2021), and ViT (Touvron et al., 2021; Yu et al., 2022), ensuring a fair comparison. These configurations are widely adopted defaults that have been empirically validated as the baseline.

A key advantage of UTA lies in its parameter φ_0. While an optimal φ_0 can further improve model performance, we intentionally fix its value across all tasks and architectures to ensure fair evaluation (Appendix H.2).

Comprehensive experimental details, such as computational resources, configurations, implementation specifics, are elaborated in Appendix H. Our implementation strictly follows the PyTorch official API. It is inherited from the base class `torch.optim.lr_LRDr.LRDcheduler`. This design ensures native compatibility with the optimizer in `torch.optim`. Fully functional, production-ready code is publicly released (see Appendix H for link and implement details).

4.1 Performance on language task with LLaMA

We evaluate on **four scales** of the LLaMA architecture: 60M, 130M, 350M, and 1B parameters. For each model, we consider **four distinct training budgets**: 10K, 15K, 20K, and 25K training steps, covering regimes from extremely limited to moderately long training. To provide a granular view of the optimization dynamics, we report the validation perplexity at around 30%, 60%, and 100% of the total budget for each scale (e.g., for 1B model, at 8k, 16k, and final 25k steps). Performance is measured by validation perplexity, which directly reflects the model generative uncertainty and language modeling quality. Results are summarized in Tab. 1.

Table 1: Perplexity ($e^{\text{validation loss}}$) on LLaMA 60M, 130M, 350M and 1B models. Lower is better (↓).

LRD	Model	60M			130M			350M			1B		
	Step	3k	6k	final	5k	10k	final	7k	14k	final	8k	16k	final
CS		35.79	29.91	28.03	27.12	22.71	21.76	22.03	**18.24**	17.62	20.14	**16.32**	15.41
LD		35.76	30.34	27.99	26.60	22.71	21.43	22.16	18.72	17.68	19.96	16.49	15.10
WSD		34.80	30.03	26.03	27.05	23.71	20.83	22.41	19.66	17.25	20.60	17.94	14.94
UBA		35.48	30.30	27.01	27.18	23.18	21.12	22.24	18.68	17.42	19.78	16.41	14.83
UTA		**33.72**	**29.57**	**25.80**	**25.74**	**22.83**	**20.16**	**22.14**	19.58	**17.17**	**19.56**	16.97	**14.53**

UTA consistently outperforms all baselines across model scales and most training checkpoints. This is not merely a marginal improvement at the final step, but a systematic acceleration throughout the entire training process. Firstly, **UTA exhibits remarkably fast convergence in the early stages**. At the 30% checkpoint, UTA achieves the lowest perplexity across all model sizes. These early-stage gains demonstrate UTA's ability to establish a proper initial trajectory, which is critical for large-scale training where early decisions profoundly impact final performance. Most critically, **UTA achieves best final perplexity across all four model scales**, with absolute improvements from 0.23 (60M), 0.67 (130M), 0.08 (350M) and 0.30 (1B) perplexity units over the best baseline. Although the absolute final gains range from 0.23 to 0.67 perplexity, the relative improvement becomes increasingly meaningful as model size grows. Because for large-scale foundation models, even a 0.2-unit reduction in perplexity translates into a statistically significant leap in generation quality and reliability. We analyze that these consistent gains may be attributed to (i) near-optimal in *shape* by Theorem 3.2, and (ii) *energy-calibrated* in scale by subsection 3.3. It gives UTA the ability to dynamically *balance exploration and stabilization* during training.

We also illustrate the perplexity evolution in Fig. 1. All curves show a characteristic rapid initial descent followed by gradual convergence. UTA consistently outperforms throughout training, demonstrating both faster initial convergence and superior final performance, while WSD and UBA exhibit competitive performance. These patterns align with theoretical expectations in Appendix G .

Table 2: ImageNet test top-1 Acc. (%) of different LRDs on ResNet50 and ViT-B for 300 epochs. Higher values indicates better performance (↑). Best results in bold.

LRD	CS	LD	WSD	UBA	UTA
ResNet-50	$78.34_{\pm 0.024}$	$77.90_{\pm 0.028}$	$78.30_{\pm 0.024}$	$78.38_{\pm 0.014}$	$\mathbf{78.73}_{\pm 0.029}$
ViT-B	$79.92_{\pm 0.057}$	$79.96_{\pm 0.104}$	$81.14_{\pm 0.048}$	$82.05_{\pm 0.041}$	$\mathbf{82.30}_{\pm 0.037}$

Table 3: Perplexity ($e^{\text{validation loss}}$) of LLaMA-60M and LLaMA-130M under different peak LRs. Lower is better.

Model	LRD	5×10^{-4}	1×10^{-3}	3×10^{-3}	5×10^{-3}	7×10^{-3}	1×10^{-2}	Best from LR Sweep
60M	CS	32.61	28.03	26.91	26.91	26.31	26.57	26.31
	LD	32.83	27.99	26.80	26.77	26.25	26.46	26.25
	WSD	29.40	26.03	26.56	**26.34**	26.06	**26.28**	26.03
	UBA	30.65	27.01	**26.51**	26.42	26.26	26.57	26.26
	UTA	**29.28**	**25.80**	26.86	27.00	**25.81**	27.11	**25.80**
130M	CS	–	21.76	21.32	20.94	21.27	–	20.94
	LD	–	21.43	21.18	20.90	**21.24**	–	20.90
	WSD	–	20.83	21.16	**20.82**	21.39	–	20.82
	UBA	–	21.12	**20.78**	21.25	21.92	–	20.78
	UTA	–	**20.16**	21.36	21.65	22.39	–	**20.16**

Moreover, we emphasize that the observed performance gains are not simply due to an increased average LR. In our experiments, the average LR of the WSD baseline was notably higher than that of our UTA, yet UTA still achieves better perplexity. The advantage stems from the optimal selection of intrinsic time in UTA, which promotes more effective learning and reduces the accumulation of noise during training, by Theorem G.2. This leads to improved generalization, ultimately resulting in superior model performance across different scales.

4.2 Performance on vision task ImageNet

We further validate the generalization of UTA on the ImageNet classification benchmark using ResNet-50 and ViT-B architectures. This selection tests generalization across model depth, width, and fundamental architectural biases. Fig. 1e, 1f, 1g and 1h depict training curves of ResNet-50 and ViT-B on ImageNet under different LRDs. For both architectures, UTA consistently achieves the highest final accuracy. UBA demonstrates competitive performance, while WSD and CS show similar but suboptimal convergence. Tab. 2 reports the test top-1 accuracy. UTA consistently achieves the highest accuracy across both models. Specifically, it improves ResNet-50 accuracy to 78.73%, outperforming the strongest baseline UBA (78.38%) by 0.35%. For ViT-B, UTA reaches 82.30%, surpassing the competitive UBA LRD (82.05%) by 0.25%. These gains are statistically meaningful given the scale and complexity of ImageNet. These consistent outperformance demonstrates that UTA generalizes beyond language modeling to vision tasks. It offers a robust and architecture-agnostic advantage in large-scale, real-world training scenarios, generalizing beyond language modeling to vision tasks.

4.3 Ablation study

Peak LR sensitivity analysis. To address the concern that a fixed peak LR may disadvantage baselines with different optimal LRs, we conduct a LR sweep on LLaMA-60M and 130M (Tab. 3). As expected, optimal peak LRs vary across LRDs. Nevertheless, under each LRD's optimal LR, **UTA consistently outperforms all baselines on both scales**. On 60M, UTA achieves a best perplexity of **25.80**, surpassing the best baseline UBA (26.26) by 0.46 and WSD (26.03) by 0.23. On 130M, UTA's advantage is more pronounced, achieving **20.16**, which outperforms WSD (20.82) by 0.66 and UBA (20.78) by 0.62. These margins, while modest in absolute terms, are meaningful in the context of language modeling validation loss.

Parameter analysis of φ_0. We analyze UTA's sensitivity to φ_0 (equation 7) on LLaMA-60M and 130M (Tab.s 7–8). UTA exhibits robustness. On 60M, it achieves perplexity below 26.10 for 7 out of 10 tested values, with variation of only 0.24 across a 10-fold range; on 130M, performance remains nearly flat (20.63–20.67) across $\varphi_0 \in [0.15, 0.5]$. Performance improves modestly as φ_0 decreases, consistent with our theory that smaller φ_0 yields larger intrinsic steps. Crucially, UTA demonstrates a broader effective range than UBA, which degrades for $\varphi_0 < 0.5$. We attribute this to UTA's adaptive mechanism, which dynamically compensates for suboptimal φ_0 choices, a capability absent in UBA. These results confirm that UTA is both performant and easy to use.

The effect of final LR in decay. Regarding the LRD implementations, some practices use decay to 10% of the peak LR, while others use 0%. Thus, we ablate the final LR ratio (10% vs. 0%) on LLaMA-60M and 130M (Tab. 5). Results show that decaying to 0% underperforms 10% across all methods and scales. Specifically, on 130M, LD degrades from 21.43 to 21.89, and UTA from 20.16 to 20.68. This suggests that aggressive decay prematurely saturates optimization under our moderate training budgets. Importantly, UTA remains the top performer under both 0% and 10% decay, surpassing the best baseline by 0.16 (60M) and 0.32 (130M). This confirms that UTA is robust to the choice of final LR, and that the default 10% decay setting does not unfairly advantage UTA.

The role of dynamic state-Aware mechanism. To validate the component efficacy of the dynamic mechanism in our UTA LRD, we conduct an ablation study comparing the full version (Ours-Full) against a static variant (Ours-Static). We conduct the ablation study on the CIFAR100-ResNet34, ImageNet-ResNet50, LLaMA60M and 130M. In the static variant, the dynamic part is removed. All other hyper-parameters and training settings remain identical. As shown in Tab. 9, UTA (Full) consistently outperforms its static counterpart. Moreover, the improvement is pronounced, suggesting that the dynamic state adaptation mechanism is a crucial component for LRD performance. More details and analyses can be found in Appendix I.3 and I.3.

Comparison with automatic LR methods. We additionally compare UTA against automatic LR methods, such as D-Adaptation (Defazio & Mishchenko, 2023) and Prodigy (Mishchenko & Defazio, 2023), which adjust LRs dynamically without predefined schedule shapes, in contrast to our main baselines. Results on LLaMA-60M, 130M, and 350M reported in Tab. 10 in Appendix. UTA matches D-Adaptation on 60M and outperforms it on 130M and 350M by 0.49 and 0.47, respectively. Against Prodigy, UTA achieves consistent improvements of 1.57, 1.10, and 0.80 across the three scales. We attribute UTA's advantage to its shape design, which explicitly accounts for the training horizon, unlike methods that rely solely on gradient statistics.

Parameter-efficient fine-tuning scenario. To demonstrate that our LRD generalizes beyond full-parameter training, we evaluate it in a parameter-efficient fine-tuning (PEFT) scenario. We apply UTA to fine-tune a ViT-B model on CIFAR-100 and OxfordPets , using LoRA(Hu et al., 2022) ($r = 16$) as the PEFT method. Tab. 12 demonstrates UTA effectiveness in PEFT settings. UTA outperforms UBA on CIFAR100 and OxfordPets. These results confirm that UTA generalizes effectively to parameter-constrained optimization (see details in Appendix I.3).

5 Conclusion

We introduced **UTA**, a novel LRD framework that bridges the gap between theoretically principled shape design and practical, state-aware adaptation. Methodologically, UTA extends the ROLR from SGD to adaptive optimizers, particularly AdamW, overcoming a key limitation of prior work. Through the concept of intrinsic time and energy-calibrated adaptation, UTA automatically determines the LR in response to real-time optimization states. Experiments on vision and language models confirm that UTA consistently outperforms existing LRDs with minimal overhead. Comprehensive ablations confirm its robustness to hyper-parameters (peak LR, final decay ratio, φ_0), the necessity of its dynamic mechanism, and its superiority over automatic LR methods. Overall, UTA not only delivers a plug-and-play LRD solution but also advances the broader understanding of state dependent optimization. In the future, potential extensions include adapting the framework to optimizers such as Conda (Wang et al., 2025) and SOAP (Vyas et al., 2025). Moreover, the optimal amplitude of LR fluctuations remains an open problem. A rigorous analysis could help choosing optimal fluctuation range, further improve model performance.

AI USE STATEMENT

In this work, we used generative AI tools solely for suggested-disclosure tasks to aid writing. These tools are used improve readability, suggesting alternative phrasing, and refining the overall presentation of our work. We have not used generative AI tools for any required-disclosure tasks as defined by the ICLR policy. Specifically, we did not use them to generate synthetic data, develop theoretical models or mathematical claims, proofs, design experiments or methodology, implement methods, or interpret results. We have reviewed all AI-assisted work. All AI-suggested text edits were critically reviewed, revised as needed, and checked for accuracy and consistency by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details. This statement should not be more than 1 page.

REPRODUCIBILITY STATEMENT

We have taken the following steps to ensure the reproducibility of our work.

Our implementation strictly follows the PyTorch official API, i.e. it is inherited from the base class

`torch.optim.lr_LRDr.LRDcheduler.`

It maintains full compatibility with all

`torch.optim`

optimizers. The complete production-ready code is publicly available.

For all results. The complete source code is provided as anonymous material and is also available at `https://anonymous.4open.science/r/UTA-LRD-7B39/UTA.py`.

All computational resources, training details, configurations, implementation specifics, model architectures and hyper-parameters are elaborated in Appendix H and Appendix H.3.

For theoretical results. All assumptions are clearly stated in Appendix B. Complete proofs of all claims are provided in Appendix D, Appendix E and Appendix F.

For experimental results. The datasets used are publicly available, described in Appendix H.5. The evaluation metrics, baselines, model architectures, hyper-parameters and comparison protocols are detailed in Section H.1 and Appendix H.3. Ablation studies and sensitivity analyses supporting our main claims are provided in Appendix I.3.

We believe the above materials are sufficient for reproducing all main results reported in this paper. Any remaining implementation details not covered here are available upon reasonable request to the authors.

REFERENCES

John Chen, Cameron Wolfe, and Tasos Kyrillidis. Rex: Revisiting budgeted training with an improved schedule. *Proceedings of Machine Learning and Systems*, 4:64–76, 2022.

Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth. On the lambert w function. *Advances in Computational mathematics*, 5(1):329–359, 1996.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by d-adaptation. In *International conference on machine learning*, pp. 7449–7479. PMLR, 2023.

Aaron Defazio, Ashok Cutkosky, Harsh Mehta, and Konstantin Mishchenko. Optimal linear decay learning rate schedules and further refinements. *arXiv preprint arXiv:2310.07831*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alex Hägele, Elie Bakouch, Atli Kosson, Leandro Von Werra, Martin Jaggi, et al. Scaling laws and compute-optimal training beyond fixed training durations. *Advances in Neural Information Processing Systems*, 37:76232–76264, 2024.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

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

Atli Kosson, Bettina Messmer, and Martin Jaggi. Rotational equilibrium: How weight decay balances learning across neural networks. *arXiv preprint arXiv:2305.17212*, 2023.

Binghui Li, Fengling Chen, Zixun Huang, Lean Wang, and Lei Wu. Functional scaling laws in kernel regression: Loss dynamics and learning rate schedules. *Advances in Neural Information Processing Systems*, 2025.

Mengtian Li, Ersin Yumer, and Deva Ramanan. Budgeted training: Rethinking deep neural network training under resource constraints. *arXiv preprint arXiv:1905.04753*, 2019a.

Qianxiao Li, Cheng Tai, et al. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pp. 2101–2110. PMLR, 2017.

Qianxiao Li, Cheng Tai, et al. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019b.

Zhiyuan Li, Kaifeng Lyu, and Sanjeev Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33:14544–14555, 2020.

Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). *Advances in Neural Information Processing Systems*, 34:12712–12725, 2021.

Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. Stack more layers differently: High-rank training through low-rank updates. 2023.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Kairong Luo, Haodong Wen, Shengding Hu, Zhenbo Sun, Zhiyuan Liu, Maosong Sun, Kaifeng Lyu, and Wenguang Chen. A multi-power law for loss curve prediction across learning rate schedules. In *International Conference on Learning Representations*, 2025.

Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv preprint arXiv:2306.06101*, 2023.

Rui Pan, Shizhe Diao, Jianlin Chen, and Tong Zhang. Extremebert: A toolkit for accelerating pretraining of customized bert. *arXiv preprint arXiv:2211.17201*, 2022a.

Rui Pan, Haishan Ye, and Tong Zhang. Eigencurve: Optimal learning rate schedule for SGD on quadratic objectives with skewed hessian spectrums. In *International Conference on Learning Representations*, 2022b.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Fabian Schaipp, Alexander Hägele, Adrien Taylor, Umut Simsekli, and Francis Bach. The surprising agreement between convex optimization theory and learning-rate scheduling for large model training. *arXiv preprint arXiv:2501.18965*, 2025.

Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision*, pp. 464–472. IEEE, 2017.

Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pp. 369–386. SPIE, 2019.

Anda Tang, Yiming Dong, Yutao Zeng, Zhouchen Lin, et al. Stepsize anything: A unified learning rate schedule for budgeted-iteration training. *Advances in Neural Information Processing Systems*, 38:150160–150209, 2026.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Howe Tissue, Venus Wang, and Lu Wang. Scaling law with learning rate annealing. *Advances in Neural Information Processing Systems*, 2025.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam for language modeling. In *International Conference on Learning Representations*, volume 2025, pp. 93423–93444, 2025.

Junjie Wang, Pan Zhou, Yiming Dong, Huan Li, Jia Li, Xun Zhou, Qicheng Lao, Cong Fang, and Zhouchen Lin. Conda: Column-normalized adam for training large language models faster. *arXiv preprint arXiv:2509.24218*, 2025.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective. *arXiv preprint arXiv:2410.05192*, 2024.

Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021.

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10819–10829, 2022.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12104–12113, 2022.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient LLM training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024.

Pan Zhou, Xingyu Xie, Zhouchen Lin, Kim-Chuan Toh, and Shuicheng Yan. Win: Weight-decay-integrated nesterov acceleration for faster network training. *Journal of Machine Learning Research*, 25(83):1–74, 2024.

CONTENTS

1 Introduction **1**

2 Preliminary **3**

3 Methodology **3**

3.1 Problem statement 3

3.2 Segment-wise continuous analysis 3

3.3 Regularized energy evolution via intrinsic time 4

3.4 Derived LRD 6

3.5 Complexity analysis 6

4 Experiments **6**

4.1 Performance on language task with LLaMA 7

4.2 Performance on vision task ImageNet 8

4.3 Ablation study 8

5 Conclusion **9**

A Notation **16**

B Assumption **18**

C Problem statement **18**

D Optimal shape of problem solution **19**

E Regularized energy evolution via intrinsic time **21**

F Derived LRD **26**

G Theoretical analysis **30**

G.1 Functional scaling laws for LRDs 30

G.2 Analytical framework for our LRD 30

H Experimental implementation details **33**

H.1 Baselines 33

H.2 LRD setting 34

H.3 Model architectures and hyper-parameters 34

H.4 Hardware and software environment 35

H.5 Datasets 35

H.6 Code and reproducibility 35

I Experimental results **36**

I.1 Performance on language model-LLaMA 36

I.2 Performance on ImageNet 36

I.3 Ablation study 38

I.4 Discussion of UTA characteristics 42

J Related work **42**

APPENDIX

IMPACT STATEMENT

In this paper, researchers introduce a Universal Training Adaptive learning rate decay called UTA.

Positive impacts UTA offers several significant benefits to the machine learning community and broader ecosystem.

- **Enhanced Training Efficiency and Accessibility.** UTA provides a leading performance across diverse architectures and scales without requiring LRD-specific tuning. This reduces the computational resources and expert knowledge needed for effective training, potentially lowering the barrier to entry for researchers and practitioners with limited resources. By minimizing the need for extensive hyper-parameter searches, which can consume thousands of GPU-hours for large models, UTA contributes to more sustainable and equitable AI research.
- **Environmental Sustainability.** The improved optimization efficiency of UTA can reduce total training time and trail-and-error. For large-scale model training, even marginal reductions in required iterations translate to substantial energy savings, thereby decreasing the carbon footprint associated with AI development. As computational demands of foundation models continue to escalate, efficient optimization strategies like UTA become increasingly critical for environmentally responsible research.
- **Scientific and Industrial Advancement.** By providing a principled, theoretically-grounded alternative to heuristic LRDs, UTA advances our fundamental understanding of optimization dynamics. The functional scaling law perspective offers new insights into how learning rate trajectories affect final performance. Practically, the reliability of UTA makes it suitable for deployment in production systems where consistent performance is paramount, potentially accelerating the development and deployment of AI applications across sectors.

Potential Risks and Mitigations

- **Homogenization of Training Practices.** If widely adopted, our method could lead to reduced diversity in optimization approaches, potentially stifling innovation in optimization strategies. To mitigate this, we emphasize that UTA complements rather than replaces existing methods, such as complementing advanced adaptive optimizer. Moreover, we encourage continued exploration of diverse optimization techniques. Our theoretical framework may actually inspire new research directions in learning rate design.
- **Computational Disparities.** While UTA reduces tuning requirements, it still assumes access to sufficient computational resources for training. This could perpetuate existing disparities between well-resourced and under-resourced institutions. We address this by providing open-source implementations and focusing on efficiency gains that benefit all users proportionally. The reduced tuning overhead may actually help smaller teams by freeing limited resources for other critical tasks.
- **Verification and Reproducibility.** As with any new optimization method, incorrect implementation or misunderstanding of UTA theoretical frameworks could lead to sub-optimal results. We mitigate this by providing clear theoretical analyses, documentation, reference implementations, and comprehensive empirical validation.

Conclusion Overall, UTA represents a net positive contribution to the field. Its primary impacts, such as increased training efficiency, reduced environmental costs, and improved accessibility, align with the community's growing emphasis on sustainable and reproducible AI research. By sharing our work openly and encouraging rigorous validation, we aim to maximize its positive impacts while engaging the community in ongoing discussions about responsible optimization research.

A NOTATION

For ease of reference, we summarize the main notation used throughout the paper.

Optimization and training.

Symbol	Meaning
$W \in \mathbb{R}^N$	model parameters (a single vector)
t	current training step (physical iteration index)
T	total number of training iterations
$\mathcal{T}$	training horizon, $[0, \mathcal{T}]$
$\mathcal{L}(\cdot)$	training loss function
ξ	data sample; ξ_t denotes the sample at step t
D	training dataset
$\mathcal{F}$	function class (network structures, datasets, etc.)
$f(W, \xi)$	neural network parameterized by W on data ξ
$\nabla\mathcal{L}(W_t)$	full gradient at step t
$\nabla\mathcal{L}(W_t, \xi_t)$	stochastic gradient at step t
δ_t	gradient noise, $\delta_t := \nabla\mathcal{L}(W_t, \xi_t) - \nabla\mathcal{L}(W_t)$

Optimizer and learning rate.

Symbol	Meaning
η_t	learning rate at step t
$\eta(\cdot)$	continuous-time learning rate
$\eta_{\min}, \eta_{\max}$	minimum and maximum learning rates
$\eta_{\min}^{(k)}, \eta_{\max}^{(k)}$	range parameters on segment k
$\phi(\cdot)$	learning rate decay (LRD) as a function of physical time
$P(t)$	optimizer-specific preconditioner (e.g., $P(t) = I$ for SGD)
m_t, v_t	first and second moment estimates in Adam
$\hat{m}_t, \hat{v}_t$	bias-corrected moments in Adam
β_1, β_2	Adam momentum coefficients
ϵ	small constant for numerical stability
λ	weight decay coefficient
B	batch size

Segments and intrinsic time.

Symbol	Meaning
K	number of segments
$[T_k, T_{k+1}]$	the k-th segment
ΔT_k	segment length, $\Delta T_k := T_{k+1} - T_k$
k	segment index; under iteration-wise extension, k coincides with t
U_k	intrinsic time of segment k, $U_k := \int_{T_k}^{T_{k+1}} \eta(t)\,\mathrm{d}t$
U_t^*	energy-optimal intrinsic time at step t
$\bar{T}(\tau)$	intrinsic-time reparameterization, $\bar{T}(\tau) := \int_{T_k}^{\tau} \eta(z)\,\mathrm{d}z$
$\bar{T}^{-1}(t)$	inverse intrinsic-time map
φ	curvature parameter of the UBA-equivalent family
φ_k, φ_t	curvature parameter on segment k / at step t
φ_0	hyper-parameter in the UTA plug-in estimate
ψ	normalized intrinsic time, $\psi := U_k^*/\Delta T_k$
$\theta_{t,k}$	phase variable of the UBA-equivalent family
$\theta(t)$	phase variable in the global UTA LRD

Energy evolution and curvature.

Symbol	Meaning
$\bar{W}^{(k)}$	local minimum in segment k
H_k	Hessian at the local minimum, $H_k := H_f^{(k)}$
$E(t)$	energy function, $E(t) := \frac{1}{2}\|e(t)\|^2$
$e(t)$	transformed deviation, $e(t) := H_k^{1/2}(W(t) - \bar{W}^{(k)})$
$\nu_\ell^{(k)}$	spectral lower bound of $H_k^{1/2} P(t) H_k^{1/2}$
α	noise intensity bound
Σ	covariance of the gradient noise
$B(t)$	standard Brownian motion
$\mu(t)$	integrating factor in the SDE analysis
ε	target accuracy in the plug-in estimate

Functional scaling law (FSL).

Symbol	Meaning
M	model size (number of parameters)
β	model capacity exponent
s	task difficulty exponent
$\mathcal{E}_t$	expected loss at intrinsic training time t
σ^2	label noise variance
$\mathcal{K}(\cdot)$	forgetting kernel
$e(t)$	learning term, $e(t) := (1+t)^{-s}$
$\gamma(\cdot)$	noise intensity as a function of intrinsic time
$T_{\max}$	total number of training steps in the FSL analysis
T_1, T_2	durations of the stable and decay stages in the UTA FSL analysis
a, b	auxiliary constants, $a := \frac{1}{2}(\eta_{\max} + \eta_{\min})$, $b := \frac{1}{2}(\eta_{\max} - \eta_{\min})$
$\hat{a}, \hat{b}$	scaled constants in the noise-term estimate
$\eqsim$	equivalence up to a constant factor

Baselines and models.

Symbol	Meaning
UBA	Universal Budget Aware LRD
UTA	Universal Training Adaptive LRD
CS	cosine decay LRD
LD	linear decay LRD
WSD	warmup–stable–decay LRD
LLaMA-60M/130M/350M/1B	model scales used in the language experiments
ResNet-50	convolutional vision backbone
ViT-B	vision transformer backbone
LoRA	low-rank adaptation for PEFT
PEFT	parameter-efficient fine-tuning

B ASSUMPTION

We adopt assumptions including local quadratic(inherited from the framework (Tang et al., 2026)) and continuous-time (standard for analyzing optimizer dynamics (Li et al., 2017; 2025)) . These assumptions provide intuition and design principles that motivate UTA in Section 3.

We also know that deep learning training landscapes are often highly non-convex and dynamic, so these assumptions may not fully reflect real-world behavior. Importantly, UTA's robust optimization formulation mitigates approximation errors by optimizing for worst-case scenarios. This ensures the method remains effective even when assumptions are partially violated, as supported by our empirical results.

C PROBLEM STATEMENT

Let $f(W, \xi)$ denote neural network parameterized by parameters $W \in \mathbb{R}^N$ on the data ξ. ξ denote the data sampling on the dataset D, and let $\mathcal{F}$ be a function class. Let $\mathcal{L}$ be the loss function, T be a maximum number of training iterations, and t be the current learning step.

We seek learning rates $\{\eta_t\}_{t=0}^{T-1}$ that minimize loss of worst-case among the training process under a generalized update rule governed by optimizer-specific dynamics. The global min-max view is

$$\begin{aligned} &\min_{\eta_1,\eta_2,\ldots,\eta_{T-1}} \max_{f\in\mathcal{F}} \mathcal{L}(f(W_T,\xi)) \\ &\text{s.t.} \quad W_{t+1} = W_t - \eta_t \cdot P(t)\nabla\mathcal{L}(W(t)), \end{aligned} \tag{9}$$

where $P(t)$ captures the optimizer-specific state dynamics, such as scaling or preconditioning, and may be chosen to represent different optimization algorithms within the same generalized framework. For instance, SGD corresponds to $P(t) = I$. Adam is recovered by defining $P(t)$ such that $P(t)\nabla\mathcal{L}(W(t)) = m_t/(\sqrt{v_t} + \epsilon)$, where m_t and v_t are the first and second moment estimates following Adam update rules. Other adaptive methods like RMSProp or AdaGrad can be represented analogously through appropriate definitions of $P(t)$ based on their state variables. This paper primarily focuses on the case where $P(t)$ implements the Adam dynamics.

This paper specializes the framework to *the Adam optimizer* for two compelling reasons. First, the unique $P(t)$ of Adam encapsulates a complex yet representative interplay of momentum and adaptive scaling mechanisms. It combines first-moment momentum with second-moment adaptive scaling, represents a complex and widely-used class of optimization dynamics. Second, as the default optimizer in much of current deep learning practice, its performance is of paramount practical importance. Analyzing Adam thus serves as a highly relevant and sufficiently comprehensive case study for our theory. Then the optimization problem is as follows,

$$\begin{aligned} &\min_{\eta_1,\eta_2,\ldots,\eta_{T-1}} \max_{f\in\mathcal{F}} \mathcal{L}(f(W_T,\xi)) \\ &\text{s.t.} \quad W_{t+1} = W_t - \eta_t \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}, \end{aligned} \tag{10}$$

where $m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla\mathcal{L}(W_t)$, $v_t = \beta_2 v_{t-1} + (1-\beta_2)\nabla\mathcal{L}(W_t)^{\odot 2}$, and $\hat{m}_t = m_t/(1-\beta_1^t)$, $\hat{v}_t = v_t/(1-\beta_2^t)$ denote the bias-corrected moments, $\odot$ is the element-wise operation.

D Optimal shape of problem solution

Lemma D.1. *Under the generalized gradient flow $\dot{W}(\tau) = -\eta(\tau)P(\tau)\nabla\mathcal{L}(W(\tau))$, where $P(\tau)$ denotes precondition. For all $t \in [T_k, T_{k+1}]$,*

$$\|W(t) - W(T_k)\| \leq U_k \cdot C_P^{(k)} \cdot G_k \leq \eta_{\max,k}\Delta T_k \cdot C_P^{(k)} \cdot G_k, \tag{11}$$

where $G_k := \sup_{\tau\in[T_k,T_{k+1})}\|\nabla\mathcal{L}(W(\tau))\|$, $C_P^{(k)} := \sup_{\tau\in[T_k,T_{k+1})}\|P(\tau)\|$, $\eta_{\max,k} := \sup_{t\in[T_k,T_{k+1}]}\eta(t)$.

This Lemma shows that segment radius can be controlled by segment length.

Proof. By integrating the flow $\dot{W}(\tau) = -\eta(\tau)P(\tau)\nabla\mathcal{L}(W(\tau))$, we have

$$W(t) - W(T_k) = \int_{T_k}^{t} \dot{W}(s)\,\mathrm{d}s, \tag{12}$$

so

$$\|W(t) - W(T_k)\| \leq \int_{T_k}^{t} \eta(\tau)\|P(\tau)\nabla\mathcal{L}(W(\tau))\|\,\mathrm{d}\tau \leq U_k \cdot C_P^{(k)} \cdot G_k, \qquad t \in [T_k, T_{k+1}), \tag{13}$$

The segment-radius bound then follows from the triangle inequality: $\|W(t) - \bar{W}^{(k)}\| \leq \|W(t) - W(T_k)\| + \|W(T_k) - \bar{W}^{(k)}\| \leq V_k\Delta T_k + U_k \cdot C_P^{(k)} \cdot G_k$. □

Assumption D.2 (Locally Lipschitz gradients)**.** On the segment $[T_k, T_{k+1})$, the gradient is Lipschitz with constant $L_g^{(k)}$, i.e., $\forall\, W, W' \in \{W(\tau) : \tau \in [T_k, T_{k+1})\}$, $\|\nabla\mathcal{L}(W) - \nabla\mathcal{L}(W')\| \leq L_g^{(k)}\|W - W'\|$.

Combining Assumption D.2 with equation 11, we obtain the segment-wise gradient drift bound

$$\begin{aligned} &\sup_{\tau\in[T_k,T_{k+1})} \|\nabla\mathcal{L}(W(\tau)) - \nabla\mathcal{L}(W(T_k))\| \\ \leq &L_g^{(k)} \sup_{\tau\in[T_k,T_{k+1})} \|W(\tau) - W(T_k)\| \leq L_g^{(k)} G_k C_P^{(k)} U_k. \end{aligned} \tag{14}$$

Since $U_k \leq \Delta T_k \cdot \max_{t\in[T_k,T_{k+1})} \eta_t$, if the segment is sufficiently small, the difference of gradients can be bounded.

Lemma D.3. *On segment $[T_k, T_{k+1})$ with $\Delta T_k := T_{k+1} - T_k$, let $\hat{m}_t$ be the bias-corrected first moment in Adam and assume:(1) the gradient along the trajectory is $L_g^{(k)}$-Lipschitz in time;(2) the gradient norm is bounded by G_k, i.e., $\|\nabla\mathcal{L}_t\| \leq G_k$ for all $t \in [T_k, T_{k+1})$.*

Then the bias-corrected momentum $\hat{m}_t$ tracks the true gradient $\nabla\mathcal{L}_t$ with a segment-wise error bound

$$\|\nabla\mathcal{L}_t - \hat{m}_t\| \leq L_g^{(k)} C_P^{(k)} G_k \cdot \Delta T_k \cdot \max_{t\in[T_k,T_{k+1})} \eta_t \frac{\beta_1}{1-\beta_1}. \tag{15}$$

Proof. It is noted that the first-order moment $\hat{m}_t$ is a convex combination of recent gradients as follows,

$$\hat{m}_t = \sum_{i=0}^{t-1} \frac{(1-\beta_1)\beta_1^i}{1-\beta_1^t} \nabla\mathcal{L}_(t-i), \quad \sum_{i=0}^{t-1} \frac{(1-\beta_1)\beta_1^i}{1-\beta_1^t} = 1. \tag{16}$$

Combining with equation 14, we have

$$\begin{aligned}
\|\nabla\mathcal{L}_t - \hat{m}_t\| &= \left\|\nabla\mathcal{L}_t - \sum_{i=0}^{t-1}\frac{(1-\beta_1)\beta_1^i}{1-\beta_1^t}\nabla\mathcal{L}_(t-i)\right\| \\
&= \left\|\sum_{i=0}^{t-1}\frac{(1-\beta_1)\beta_1^i}{1-\beta_1^t}(\nabla\mathcal{L}_t - \nabla\mathcal{L}_(t-i))\right\| \\
&\le \sum_{i=0}^{t-1}\frac{(1-\beta_1)\beta_1^i}{1-\beta_1^t}\cdot L_g^{(k)}G_k\cdot\Delta T_k\cdot\max_{t\in[T_k,T_{k+1})}\eta_t\cdot i \\
&= L_g^{(k)}C_P^{(k)}G_k\cdot\Delta T_k\cdot\max_{t\in[T_k,T_{k+1})}\eta_t\cdot\frac{(1-\beta_1)}{1-\beta_1^t}\frac{\beta_1 - t\beta_1^t + (t-1)\beta_1^{t+1}}{(1-\beta_1)^2} \\
&\le L_g^{(k)}C_P^{(k)}G_k\cdot\Delta T_k\cdot\max_{t\in[T_k,T_{k+1})}\eta_t\frac{\beta_1}{1-\beta_1}.
\end{aligned} \tag{17}$$

It yields the claim. Hence on a short segment, $\hat{m}_t \approx \nabla\mathcal{L}_t$. □

Theorem 3.2 [Optimal UBA-equivalent shape on a segment] *Consider the min-max problem 10 on segment k, and assume that the step sizes are constrained to lie in the interval* $[\eta_{\min}^{(k)}, \eta_{\max}^{(k)}]$. *There exists an optimal shape of step sizes that minimizes the worst-case loss. This optimal shape belongs to the UBA-equivalent family for some curvature parameter* $\varphi_k > 0$.

Proof. By Lemma D.3, it makes precise the intuition that on a short and smooth segment the Adam momentum is close to the current gradient.

The bound

$$\varepsilon_g^{(k)} := L_g^{(k)}C_P^{(k)}G_k\cdot\Delta T_k\max_{t\in[T_k,T_{k+1})}\eta_t\frac{\beta_1}{1-\beta_1}. \tag{18}$$

depends linearly on the segment length ΔT_k and on the maximal stepsize on the segment. Thus, by choosing ΔT_k small and avoiding excessively large η_t, we can make $\varepsilon_g^{(k)}$ small. In other words, on sufficiently short segments we may treat the Adam momentum $\hat{m}_t$ as a perturbation of the true gradient $\nabla\mathcal{L}_t$, with a perturbation norm controlled by $\varepsilon_g^{(k)}$.

Thus, by Lemma D.3 and following the local quadratic modeling in prior work (Tang et al., 2026), we can approximate the constraint in UBA framework(Tang et al., 2026) as follows,

$$W_{t+1} - \bar{W}^{(k)} \approx (I - \eta_t P_t H_k)(W_t - \bar{W}^{(k)}),$$

where $P_t := \frac{1}{\sqrt{\hat{v}_t}+\epsilon}$. Denote $P_t H_k := \bar{H}_k$, by iteratively applying this equation, we obtain

$$W_{T_{k+1}} - \bar{W}^{(k)} = \left(I - \eta_{T_{k+1}-1}\bar{H}_k\right)\cdots\left(I - \eta_{T_k+1}\bar{H}_k\right)\left(W_{T_k+1} - \bar{W}^{(k)}\right). \tag{19}$$

Since the matrix $\bar{H}_k \in \mathbb{R}^{N\times N}$ possesses N eigenvalues $\lambda_1^{(k)}(f(\xi)), \lambda_2^{(k)}(f(\xi)), \cdots, \lambda_N^{(k)}(f(\xi))$ abbreviated as $\lambda_i^{(k)}$, which satisfy $\lambda_l^{(k)} \le \lambda_i^{(k)}(f(\xi)) \le \lambda_u^{(k)}$ for all i. Here, $\lambda_l^{(k)}$ and $\lambda_u^{(k)}$ denote the bounds on the non-zero eigenvalues of the $\bar{H}_k$ of optimum in k-th segment. Moreover, there exist N linearly independent eigenvectors $u_1^{(k)}(f(\xi)), u_2^{(k)}(f(\xi)), \cdots, u_N^{(k)}(f(\xi))$ abbreviated as $u_i^{(k)}$, which can form the basis for the N dimension vector space. Then we have $\bar{H}_k u_i^{(k)} = \lambda_i^{(k)} u_i^{(k)}$ $(i = 1, 2, \cdots, N)$ and the initial deviation from the optimal solution can be expressed as $W_{T_k} - \bar{W}^{(k)} = \sum_{i=1}^{N} s_i^{(k)} u_i^{(k)}$ where $s_i^{(k)}$ are the coefficients corresponding to the eigenvector components. Thus, we

obtain

$$\begin{aligned}\left\|W_{T_{k+1}}-\bar{W}^{(k)}\right\|_2^2 &= \sum_{i=1}^{N}(s_i^{(k)})^2\|u_i^{(k)}\|_2^2\left[\prod_{t=1+T_k}^{T_{k+1}}\left(1-\eta_t\lambda_i^{(k)}\right)\right]^2 \\ &\le \sum_{i=1}^{N}(s_i^{(k)})^2\|u_i^{(k)}\|_2^2 \max_{\lambda_l^{(k)}\le\lambda_i^{(k)}\le\lambda_u^{(k)}}\left[\prod_{t=1+T_k}^{T_{k+1}}\left(1-\eta_t\lambda_i^{(k)}\right)\right]^2 \\ &= \max_{\lambda_l^{(k)}\le\lambda_i^{(k)}\le\lambda_u^{(k)}}\left[\prod_{t=1+T_k}^{T_{k+1}}\left(1-\eta_t\lambda_i^{(k)}\right)\right]^2\left\|W_{T_k}-\bar{W}^{(k)}\right\|_2^2\end{aligned} \tag{20}$$

We observe that this formulation shares the same optimization form as the UBA framework(Tang et al., 2026). Thus, in each short segment, the problem equation 10 reduces to UBA problem. The solution of problem equation 10 belongs to the UBA-equivalent family. □

E REGULARIZED ENERGY EVOLUTION VIA INTRINSIC TIME

Denote $\mathbb{E}_{\xi_t}\nabla\mathcal{L}(W_t,\xi_t)=\nabla\mathcal{L}(W_t)$, we have $\nabla\mathcal{L}(W_t,\xi_t)=\nabla\mathcal{L}(W_t)+\delta_t$, where the gradient noise δ_t satisfies $\mathbb{E}(\delta_t)=0$. Since the segment is short, the stochastic preconditioned gradient flow be modeled in accordance with SGD noise (Li et al., 2025) scaling by

$$\mathrm{d}W(t) = -\eta(t)\,P(t)\nabla\mathcal{L}(W(t))\,\mathrm{d}t \;+\; \eta(t)\,P(t)\Sigma^{1/2}\,\mathrm{d}B(t), \tag{21}$$

where $W(t)$ is the weight parameter, $\eta(t)\ge 0$, $P(t)\succ 0$ is diagonal, and $\Sigma\succeq 0$ is the covariance of the gradient noise and $B(t)$ is a standard Brownian motion.

According to the second-order necessary condition for the strict minimum $\bar{W}^{(k)}$, this approximation is achieved through the positive semi-definite Hessian matrix $H_f^{(k)}(\xi)\in\mathbb{R}^{N\times N}$, $H_k\succeq 0$. By quadratic expansion, the loss function can be formulated as $\mathcal{L}(f(\bar{W}^{(k)},\xi))+\nabla\mathcal{L}(f(\bar{W}^{(k)},\xi))\left(W_{T_{k+1}}-\bar{W}^{(k)}\right)+\frac{1}{2}\left(W_{T_{k+1}}-\bar{W}^{(k)}\right)^{\top}H_f^{(k)}(\xi)\left(W_{T_{k+1}}-\bar{W}^{(k)}\right)\quad(k=1,2,\cdots,K)$. Since the neural networks possess sufficient capacity to fit the data, the loss at the optimal point can be made negligible. Thus the term $\mathcal{L}(f(\bar{W}^{(k)},\xi))$ can be reasonably neglected. In addition, the first-order condition $\nabla\mathcal{L}(f(\bar{W}^{(k)},\xi))=0$ at the optimal point $\bar{W}^{(k)}$. Thus the loss function can be measured by $\frac{1}{2}\left(W_{T_{k+1}}-\bar{W}^{(k)}\right)^{\top}H_f^{(k)}(\xi)\left(W_{T_{k+1}}-\bar{W}^{(k)}\right)\quad(k=1,2,\cdots,K)$

Under this quadratic expansion measurement, we define the energy function between the t-th step weights and local minimum within segment

$$E(t):=\frac{1}{2}\|e(t)\|^2,\qquad e(t):=H_k^{1/2}\left(W(t)-\bar{W}^{(k)}\right). \tag{22}$$

with the above assumption, we have

Theorem 3.3 [Bias-variance-style upper bound via intrinsic time] *Consider the k-th segment $[T_k,T_{k+1}]$ and the preconditioned stochastic gradient flow equation 21, the expected energy at the segment end satisfies the bias–variance bound*

$$\mathbb{E}[E(T_{k+1})] \;\le\; \underbrace{\mathbb{E}[E(T_k)]\,e^{-2\nu_\ell^{(k)}U_k}}_{\textit{bias (contraction)}} \;+\; \underbrace{\frac{1}{2}\alpha\int_{T_k}^{T_{k+1}}\eta(t)^2\,dt}_{\textit{variance (noise accumulation)}}\;. \tag{23}$$

where $\alpha := \sup_{t\in[T_k,T_{k+1}]}\mathrm{tr}(H_k^{1/2}P(t)\Sigma P(t)H_k^{1/2})$ is the noise intensity bound, and $\lambda_{\min}(M(t))\ge\nu_\ell^{(k)}>0, M(t):=H_k^{1/2}P(t)H_k^{1/2}$ is the spectral lower bound.

Proof. According to equation 21 and local quadratic approximation $\nabla\mathcal{L}(W(t)) = H_k\left(W(t) - \bar{W}^{(k)}\right)$, we have

$$\begin{aligned}
\mathrm{d}e(t) &= H_k^{1/2}\mathrm{d}W(t) \\
&= -\eta(t)\, H_k^{1/2}\, P(t) H_k \left(W(t) - \bar{W}^{(k)}\right) \mathrm{d}t \;+\; \eta(t)\, H_k^{1/2}\, P(t)\Sigma^{1/2}\, \mathrm{d}B(t) \\
&= -\eta(t)\, H_k^{1/2}\, P(t) H_k^{1/2} e(t)\, \mathrm{d}t \;+\; \eta(t)\, H_k^{1/2}\, P(t)\Sigma^{1/2}\, \mathrm{d}B(t).
\end{aligned} \tag{24}$$

Consider $E(t)$ as a function of $e(t)$and apply the Itô formula:

$$\mathrm{d}E(t) = \nabla_e E(t)^\top \mathrm{d}e(t) + \frac{1}{2}\mathrm{tr}\left[\mathrm{d}e(t)\mathrm{d}e(t)^\top \nabla_e^2 E(t)\right], \tag{25}$$

where $\nabla_e E(t) = e(t)$ and $\nabla_e^2 E(t) = I$.

Substituting equation 24 into equation 25, the first term is

$$\begin{aligned}
\nabla_e E(t)^\top \mathrm{d}e(t) &= e(t)^\top \left[-\eta(t)\, H_k^{1/2}\, P(t) H_k^{1/2} e(t)\, \mathrm{d}t \;+\; \eta(t)\, H_k^{1/2}\, P(t)\Sigma^{1/2}\, \mathrm{d}B(t)\right] \\
&= -\eta(t) e(t)^\top\, H_k^{1/2}\, P(t) H_k^{1/2} e(t)\mathrm{d}t + \eta(t) e(t)^\top H_k^{1/2} P(t)\Sigma^{1/2}\mathrm{d}B(t).
\end{aligned} \tag{26}$$

For the part in second term

$$\mathrm{d}e(t)\mathrm{d}e(t)^\top = \left[-\eta(t)\, H_k^{1/2}\, P(t) H_k^{1/2} e(t)\, \mathrm{d}t \;+\; \eta(t)\, H_k^{1/2}\, P(t)\Sigma^{1/2}\, \mathrm{d}B(t)\right] \tag{27}$$

$$\cdot \left[-\eta(t)\, H_k^{1/2}\, P(t) H_k^{1/2} e(t)\, \mathrm{d}t \;+\; \eta(t)\, H_k^{1/2}\, P(t)\Sigma^{1/2}\, \mathrm{d}B(t)\right]^\top \tag{28}$$

$$= \left[\eta(t)\, H_k^{1/2}\, P(t)\Sigma^{1/2}\, \mathrm{d}B(t)\right]\left[\eta(t)\, H_k^{1/2}\, P(t)\Sigma^{1/2}\, \mathrm{d}B(t)\right]^\top \tag{29}$$

$$= \eta^2(t) H_k^{1/2} P(t)\Sigma^{1/2}\mathrm{d}B(t)\mathrm{d}B(t)^\top \Sigma^{1/2\top} P(t)^\top \left(H_k^{1/2}\right)^\top \tag{30}$$

$$= \eta^2(t) H_k^{1/2} P(t)\Sigma P(t)^\top H_k^{1/2}\mathrm{d}t. \tag{31}$$

The second equality is because multiplication rules in stochastic differential form. In addition, The second equality is because $\mathrm{d}B(t)\mathrm{d}B(t)^\top = I\mathrm{d}t$. Hence

$$\frac{1}{2}\mathrm{tr}\left[\mathrm{d}e(t)\mathrm{d}e(t)^\top \nabla_e^2 E(t)\right] = \frac{1}{2}\eta^2(t)\mathrm{tr}\left[P(t)\Sigma P(t)^\top H_k\right]\mathrm{d}t. \tag{32}$$

Putting together

$$\begin{aligned}
\mathrm{d}E(t) = \Big[&-\eta(t) e(t)^\top\, H_k^{1/2}\, P(t) H_k^{1/2} e(t) \\
&+\frac{1}{2}\eta^2(t)\mathrm{tr}\left(P(t)\Sigma P(t)^\top H_k\right)\Big]\mathrm{d}t \\
&+ \eta(t) e(t)^\top H_k^{1/2} P(t)\Sigma^{1/2}\mathrm{d}B(t).
\end{aligned} \tag{33}$$

Taking expectations, the martingale term has zero mean, yielding

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[E(t)] = &-\eta(t)\,\mathbb{E}\left[e(t)^\top H_k^{1/2}\, P(t) H_k^{1/2} e(t)\right] \\
&+ \frac{1}{2}\eta(t)^2\mathrm{tr}\left(P(t)\Sigma P(t)^\top H_k\right).
\end{aligned} \tag{34}$$

Since $P(t)$ is symmetric positive definite and $H_k^{1/2}$ is symmetric positive semi-definite, there exists eigenvalue lower bound of $\nu_\ell^{(k)} > 0$ of $H_k^{1/2}\, P(t) H_k^{1/2}$ such that

$$e(t)^\top H_k^{1/2}\, P(t) H_k^{1/2} e(t) \;\geq\; \nu_\ell^{(k)} \|e(t)\|^2, \qquad \forall t \in [T_k, T_{k+1}). \tag{35}$$

Therefore

$$\frac{d}{dt}\mathbb{E}[E(t)] \leq -2\eta(t)\nu_\ell^{(k)}\,\mathbb{E}[E(t)] + \frac{1}{2}\alpha\eta(t)^2, \tag{36}$$

where we denote $\alpha = \mathrm{tr}\left(P(t)\Sigma P(t)^\top H_k\right)$.

Solving the linear inequality via an integrating factor $\mu(t) := \exp\left(\int_{T_k}^{t} \eta(\tau)\nu_\ell^{(k)}\mathrm{d}\tau\right), \quad t \in [T_k, T_{k+1}]$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu(t) = \nu_\ell^{(k)}\eta(t)\mu(t). \tag{37}$$

Multiplying $\mu(t)$ to equation 36, i.e., $\mu(t)\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[E(t)] + 2\eta(t)\nu_\ell^{(k)}\mu(t)\mathbb{E}[E(t)] \le \frac{1}{2}\alpha\eta(t)^2\mu(t)$,we have

$$\frac{\mathrm{d}}{\mathrm{d}t}[\mu(t)\mathbb{E}[E(t)]] \le \frac{1}{2}\alpha\eta(t)^2\mu(t). \tag{38}$$

Integrate from T_k to T_{k+1}, we have

$$\mu(T_{k+1})\mathbb{E}[E(T_{k+1})] - \mu(T_k)\mathbb{E}[E(T_k)] \le \frac{1}{2}\alpha\int_{T_k}^{T_{k+1}} \eta(t)^2\mu(t)\mathrm{d}t. \tag{39}$$

Since $\mu(T_k) = 1$ and $U_k = \int_{T_k}^{T_{k+1}} \eta(\tau)\,\mathrm{d}\tau$, we have

$$\mathbb{E}[E(T_{k+1})] \le \mathbb{E}[E(T_k)]\, e^{-2\nu_\ell^{(k)}U_k} + \frac{1}{2}\alpha\, e^{-\nu_\ell^{(k)}U_k}\int_{T_k}^{T_{k+1}} \eta(t)^2\mu(t)\mathrm{d}t. \tag{40}$$

In typical network training, the learning rate satisfies $\eta(t) > 0$. Consequently, we have $\mu(t) := \exp\left(\int_{T_k}^{t} \eta(\tau)\nu_\ell^{(k)}\mathrm{d}\tau\right) \le \exp\left(\nu_\ell^{(k)}U_k\right)$, which is equation 23. □

Corollary 3.4 [ℓ_1-regularized intrinsic-time problem] *Under Theorem 3.3 and the constraint* $0 \le \eta(t) \le \eta_{\max}$ *on* $[T_k, T_{k+1}]$*, the intrinsic-time selection problem reduces to the* ℓ_1*-regularized problem,*

$$\min_{U\in[0,\eta_{\max}\Delta T_k]} \left\{\mathbb{E}[E(T_k)]e^{-2\nu_\ell^{(k)}U} + \frac{1}{2}\alpha\eta_{\max}U\right\}. \tag{41}$$

The unconstrained minimizer admits the closed form

$$U_k^* = \frac{1}{2\nu_\ell^{(k)}}\ln\left(\frac{4\nu_\ell^{(k)}\mathbb{E}[E(T_k)]}{\alpha\eta_{\max}}\right). \tag{42}$$

The constrained minimizer is followed by clamping to $[0, \eta_{\max}\Delta T_k]$.

Proof. Since $0 \le \eta(t) \le \eta_{\max}$, we have $\eta(t)^2 \le \eta_{\max}\eta(t)$, hence

$$\int \eta(t)^2 dt \le \eta_{\max}\int \eta(t)\mathrm{d}t = \eta_{\max}U_k. \tag{43}$$

Substituting equation 43 into equation 23, we arrive at the simplified bound

$$\mathbb{E}[E(T_{k+1})] \le \mathbb{E}[E(T_k)]\, e^{-2\nu_\ell^{(k)}U_k} + \frac{1}{2}\alpha\eta_{\max}U_k. \tag{44}$$

Consequently, selecting U_k by minimizing the upper bound equation 44 reduces to the ℓ_1-regularized problem

$$\min_{U\in[0,\eta_{\max}\Delta T_k]} \left\{\mathbb{E}[E(T_k)]e^{-2\nu_\ell^{(k)}U} + \frac{1}{2}\alpha\eta_{\max}U\right\}. \tag{45}$$

Differentiating the objective in equation 45 gives the condition $2\nu_\ell^{(k)}\mathbb{E}[E(T_k)]e^{-2\nu_\ell^{(k)}U} = \frac{1}{2}\alpha\eta_{\max}$. Then it gives the explicit form optimal intrinsic time,

$$\hat{U_k^*} = \frac{1}{2\nu_\ell^{(k)}}\ln\left(\frac{4\nu_\ell^{(k)}\mathbb{E}[E(T_k)]}{\eta_{\max}\alpha}\right). \tag{46}$$

Note that $U_k \in [0, \eta_{\max}(T_{k+1} - T_k)]$, therefore the actual U_k^* is obtained by clamping $\hat{U_k^*}$. It rearranges to equation 42. □

Corollary 3.5 [ℓ_2 surrogate via near-constant stepsizes] *Under Theorem 3.3 and continuity of UBA-shape function, the intrinsic-time selection problem can be approximated as ℓ_2-regularized problem,*

$$\min_{U_k \in [0, \eta_{\max}\Delta T_k]} \mathbb{E}[E(T_k)] e^{-2\nu_\ell^{(k)} U_k} + \frac{1}{2}\alpha \frac{U_k^2}{\Delta T_k}, \tag{47}$$

whose minimizer is given by the Lambert-W expression

$$U_k^* = \frac{1}{2\nu_\ell^{(k)}} W\left(\frac{2(\nu_\ell^{(k)})^2 \mathbb{E}[E(T_k)]\Delta T_k}{\alpha} \right), \tag{48}$$

followed by clamping to $[0, \eta_{\max}\Delta T_k]$.

Proof. Given the continuity of UBA-shape function, the stepsize on a short segment k is slowly varying so that $\eta(t) \approx \bar{\eta}_k := U_k/\Delta T_k$. Then

$$\int_{T_k}^{T_{k+1}} \eta(t)^2 \mathrm{d}t \approx \frac{U_k^2}{\Delta T_k}. \tag{49}$$

Under Theorem 3.3, substituting into equation 23 yields the ℓ_2-regularized objective

$$\min_{U_k \in [0, \eta_{\max}\Delta T_k]} \mathbb{E}[E(T_k)] e^{-2\nu_\ell^{(k)} U_k} + \frac{1}{2}\alpha \frac{U_k^2}{\Delta T_k}, \tag{50}$$

Differentiating equation 50 with respect to U_k gives

$$J_k(U_k) = -2\nu_\ell^{(k)} \mathbb{E}[E(T_k)]\, e^{\left(-2\nu_\ell^{(k)} U_k\right)} + \frac{\alpha}{\Delta T_k} U_k. \tag{51}$$

Setting $J_k(U_k) = 0$ yields

$$\frac{\alpha}{\Delta T_k} U_k = 2\nu_\ell^{(k)} \mathbb{E}[E(T_k)]\, e^{\left(-2\nu_\ell^{(k)} U_k\right)}. \tag{52}$$

Multiplying both sides of equation 52 by $2\nu_\ell^{(k)} e^{\left(2\nu_\ell^{(k)} U_k\right)}$ leads to

$$2\nu_\ell^{(k)} \frac{\alpha}{\Delta T_k} U_k\, e^{\left(2\nu_\ell^{(k)} U_k\right)} = 2\nu_\ell^{(k)} \nu_\ell^{(k)} \mathbb{E}[E(T_k)]. \tag{53}$$

The above equation becomes

$$\left(2\nu_\ell^{(k)} U_k\right) e^{\left(2\nu_\ell^{(k)} U_k\right)} = \frac{2\left(\nu_\ell^{(k)}\right)^2 \mathbb{E}[E(T_k)]\Delta T_k}{\alpha}. \tag{54}$$

By definition of the Lambert W function, $2\nu_\ell^{(k)} U_k = W\left(\frac{2\left(\nu_\ell^{(k)}\right)^2 \mathbb{E}[E(T_k)]\Delta T_k}{\alpha} \right)$, and thus

$$U_k^* = \frac{1}{2\nu_\ell^{(k)}} W\left(\frac{2(\nu_\ell^{(k)})^2 \mathbb{E}[E(T_k)]\Delta T_k}{\alpha} \right), \tag{55}$$

which is equation 48. □

A computable plug-in estimate of U_k^*. Both closed-form solutions in equation 42 and equation 48 depend on the quantities $\nu_\ell^{(k)}$, $\mathbb{E}[E(T_k)]$, and α. Direct estimation of these quantities is computationally expensive. We therefore derive a plug-in estimate based on measurable surrogates.

For the ℓ_1 solution equation 42, the logarithmic term is dominated by the ratio $\nu_\ell^{(k)} \mathbb{E}[E(T_k)]/\alpha$. For the ℓ_2 solution equation 48, using the first-order approximation $W(x) \approx x$ for small x, we obtain $U_k^* \approx \frac{\nu_\ell^{(k)} \mathbb{E}[E(T_k)]\Delta T_k}{\alpha}$. Both solutions share the common dominant factor

$$\frac{\nu_\ell^{(k)} \mathbb{E}[E(T_k)]}{\alpha}. \tag{56}$$

Recall that the energy function is defined as the local quadratic surrogate of the true loss $\mathcal{L}$:

$$E(t) := \frac{1}{2}\|H_k^{1/2}(W(t) - \bar{W}^{(k)})\|^2 \approx \mathcal{L}(W(t)) - \mathcal{L}(\bar{W}^{(k)}) \approx \mathcal{L}(W(t)), \tag{57}$$

where the last approximation follows from the assumption that the loss at the local minimum $\mathcal{L}(\bar{W}^{(k)})$ is negligible. Under this identification, E and $\mathcal{L}$ are used interchangeably in the following derivations.

For the symmetric positive semidefinite Hessian $H_k = \nabla^2\mathcal{L}(\bar{W}^{(k)})$, the Rayleigh–Ritz variational formula characterizes its smallest positive eigenvalue as

$$\nu_\ell^{(k)} = \lambda_{\min}^+(H_k) = \min_{v \perp \ker(H_k)} \frac{v^T H_k v}{\|v\|^2}. \tag{58}$$

Consequently, for any update direction ΔW,

$$\Delta W^T H_k \Delta W \geq \nu_\ell^{(k)} \|\Delta W\|^2. \tag{59}$$

Let the actual parameter update be $\Delta W_k = W_{k+1} - W_k = d_k$, where $d_k := \hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$ is the normalized Adam update direction. Using the quadratic expansion of E around $\bar{W}^{(k)}$, the single-step energy decrease satisfies

$$\Delta E_k := E(W_k) - E(W_{k+1}) \approx \frac{1}{2}\nu_\ell^{(k)} \|\Delta W_k\|^2. \tag{60}$$

Substituting $\Delta W_k = d_k$ yields

$$\nu_\ell^{(k)} \approx \frac{2\Delta E_k}{\|d_k\|^2}, \tag{61}$$

where $\approx$ denotes less-than or approximately less-than.

Summing the single-step decrease over $k = 0, \ldots, T_k - 1$ gives

$$\nu_\ell^{(k)} \sum_{i=0}^{T_k-1} \|\Delta W_i\|^2 \approx \sum_{i=0}^{T_k-1} \Delta E_i = |E(T_k) - E(0)|\,. \tag{62}$$

Assuming the curvature $\nu_\ell^{(k)}$ varies slowly within the segment and that the update directions satisfy $\|\Delta W_i\| \approx \|d_k\|$ for i near T_k, we obtain

$$\nu_\ell^{(k)} T_k \|d_k\|^2 \approx |E(T_k) - E(0)|\,. \tag{63}$$

For large T_k, the initial energy $E(0)$ is negligible compared with $E(T_k)$, so

$$\nu_\ell^{(k)} E(T_k) \approx \frac{|E(T_k) - E(0)|\, E(T_k)}{T_k \|d_k\|^2}. \tag{64}$$

Since $|E(T_k) - E(0)|\, E(T_k)$ is related to loss function and the loss decrease, which is a small amount. Let T_k denote the number of steps required to reach a target accuracy ε, so that $T_k \approx (|E(T_k) - E(0)|\, E(T_k))\,/\varepsilon$. Substituting this relation gives

$$\frac{\nu_\ell^{(k)} E(T_k)}{\alpha} \approx \frac{1}{\alpha} \cdot \frac{\varepsilon}{\|d_k\|^2}. \tag{65}$$

The quadratic form obtained above, i.e.,

$$\frac{\nu_\ell^{(k)} E(T_k)}{\alpha} \approx \frac{1}{\alpha} \cdot \frac{\varepsilon}{\|d_k\|^2}, \tag{66}$$

provides a theoretically characterization of the dominant factor. However, when employed as a plug-in estimate for the intrinsic-time scale U_k^* in the UBA framework, the inverse quadratic dependence on $\|d_k\|$ may lead to numerical instability when $\|d_k\|$ becomes small, as the estimate can grow arbitrarily large without saturation.

To address this issue while preserving the qualitative behavior of the original estimate, we replace the quadratic inverse $\|d_k\|^{-2}$ with the reciprocal-linear surrogate $(1 + \|d_k\|)^{-1}$. This substitution is justified by the following considerations: 1)*Monotonicity preservation:* Both forms are strictly decreasing functions of $\|d_k\|$, ensuring that larger updates correspond to smaller intrinsic-time scales. 2) *Asymptotic consistency:* 3) *Numerical stability:* The reciprocal-linear form is naturally bounded in $(0, 1]$, preventing unbounded growth when $\|d_k\|$ approaches zero, which is crucial for robust estimation in practice. Thus, we obtain

$$\frac{\nu_\ell^{(k)} E(T_k)}{\alpha} \approx \frac{1}{1 + C\|d_k\|}, \tag{67}$$

where $C := \sqrt{\varepsilon/(\alpha)}$ absorbs the constants.

We set C as a hyper-parameter, for the consistency of the initial φ_0 This yields the stabilized plug-in estimate

$$U_t^* = \frac{1}{1 + \sqrt{\varphi_0/2}\left\|\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}\right\|} \tag{68}$$

for the intrinsic-time energy scale.

F Derived LRD

Since the optimal solution of problem equation 3 within each short segment is given by a member of UBA-equivalent family (Definition equation 2.1), U_k is a deterministic function of the segment-specific shape parameters such as $\eta_{\min}^{(k)}, \eta_{\max}^{(k)}$ and φ_k.

To finalize the design, we calibrate the scale of the UTA LRD by *matching* two forms of intrinsic time:

- The energy-optimal intrinsic time U_k^* equation 7 from regularized energy evolution analysis (Subsection 3.3).
- The intrinsic time implied by the UBA-equivalent shape in Definition equation 2.1, i.e., $U_{\varphi_k} = \int_{T_k}^{T_{k+1}} g_{a,b}\left(\varphi, \theta_{t,k}\right) \mathrm{d}t$.

Concretely, for desired parameterization of the shape, such as fixing $b = \eta_{\min}^{(k)}$ and $a = \left(\eta_{\max}^{(k)} - \eta_{\min}^{(k)}\right)$, solving for φ_k, we solve the scalar equation

$$U_{\varphi_k} = U_k^*, \tag{69}$$

which uniquely determines the remaining free parameter on the segment. This yields a per-segment effective LRD that is (i) near-optimal in shape for problem equation 3, by Theorem 3.2, and (ii) energy-calibrated in scale, by subsection 3.3.

Next, we solve the matching equation equation 69 and derive the explicit form of the parameter φ_k. Without loss of generality, set $\eta_{\max}^{(k)} = 1$ and $\eta_{\min}^{(k)} = 0$, we proceed as follows:

Theorem F.1. *For a given energy-optimal intrinsic time U_k^*, there exists an explicit solution of φ_k to matching $U_{\varphi_k} = U_k^*$, expressed analytically as*

$$\varphi_k = \frac{2\left(1 - \psi\right)^2}{\psi^2}, \tag{70}$$

where

$$\psi = \frac{U_k^*}{\Delta T_k}. \tag{71}$$

Proof. Without loss of generality, let $\eta_{\max}^{(k)} = 1$ and $\eta_{\min}^{(k)} = 0$. By definition of $\theta_k(t)$ and $\tau := t - T_k$, we have

$$\frac{d\theta_k}{d\tau} = \frac{\pi}{\Delta T_k}, \qquad d\tau = \frac{\Delta T_k}{\pi}\, d\theta_k.$$

Thus

$$\int_{T_k}^{T_{k+1}} \eta_k(t)\,\mathrm{d}t = \int_0^{\Delta T_k} \eta_k(T_k+\tau)\,\mathrm{d}\tau = \frac{\Delta T_k}{\pi}\int_0^{\pi} g_k(\theta)\,\mathrm{d}\theta, \tag{72}$$

where

$$g_k(\theta) := \frac{2(1+\cos(\theta+(k-1)\pi))}{2\varphi_k + (2-\varphi_k)(1+\cos(\theta+(k-1)\pi))}. \tag{73}$$

Using $\cos(\theta+(k-1)\pi) = (-1)^{k-1}\cos\theta$ and the substitution $\theta' = \pi - \theta$ (so that $\cos\theta' = -\cos\theta$ and the integration limits remain $[0,\pi]$), one checks that

$$\int_0^{\pi} g_k(\theta)\,\mathrm{d}\theta = \int_0^{\pi} g_1(\theta)\,\mathrm{d}\theta =: I(\varphi_k), \tag{74}$$

so the accumulated step U_k is independent of k and it suffices to compute

$$I(\varphi) = \int_0^{\pi} \frac{2(1+\cos\theta)}{2\varphi + (2-\varphi)(1+\cos\theta)}\,\mathrm{d}\theta. \tag{75}$$

Let $a := \varphi + 2$ and $b := 2 - \varphi$, so that

$$I(\varphi) = \int_0^{\pi} \frac{2(1+\cos\theta)}{a + b\cos\theta}\,\mathrm{d}\theta. \tag{76}$$

We decompose

$$\begin{aligned}\int_0^{\pi} \frac{1+\cos\theta}{a+b\cos\theta}d\theta &= \int_0^{\pi}\frac{1}{a+b\cos\theta}\mathrm{d}\theta + \int_0^{\pi}\frac{\cos\theta}{a+b\cos\theta}\mathrm{d}\theta \\ &:= I_1 + I_2.\end{aligned} \tag{77}$$

$$\begin{aligned}\int_0^{\pi}\frac{\mathrm{d}\theta}{a+b\cos\theta} &= \frac{2}{\sqrt{a^2-b^2}}\arctan\left(\sqrt{\frac{a-b}{a+b}}\tan\frac{\theta}{2}\right)\Big|_0^{\pi}, \\ &= \frac{\pi}{\sqrt{a^2-b^2}}.\end{aligned} \tag{78}$$

Since

$$I_2 = \frac{\pi}{b} - \frac{a}{b}I_1, \tag{79}$$

and using the standard integral $\int_0^{\pi}\mathrm{d}\theta/(a+b\cos\theta) = \pi/\sqrt{a^2-b^2}$ for $a > |b|$, we have

$$\begin{aligned}\int_0^{\pi}\frac{1+\cos\theta}{a+b\cos\theta}\mathrm{d}\theta &= \frac{\pi}{\sqrt{a^2-b^2}} + \left(\frac{\pi}{b} - \frac{a\pi}{b\sqrt{a^2-b^2}}\right), \\ &= \pi\left(\frac{2-\sqrt{2\varphi}}{4-2\varphi}\right).\end{aligned} \tag{80}$$

Since $a^2 - b^2 = 8\varphi$, we obtain

$$I(\varphi) = 2\left[\frac{\pi}{b} + \left(1-\frac{a}{b}\right)\frac{\pi}{\sqrt{a^2-b^2}}\right] = \pi\,\frac{2-\sqrt{2\varphi}}{2-\varphi}. \tag{81}$$

Substituting this into equation 72 yields

$$U_k = \Delta T_k\,\frac{2-\sqrt{2\varphi_k}}{2-\varphi_k}. \tag{82}$$

Substituting the optimal U_k^* in each segment into equation 82 gives

$$\varphi_k = \frac{2\,(1-\psi)^2}{\psi^2}, \tag{83}$$

where

$$\psi = \frac{U_k^*}{\Delta T_k}. \tag{84}$$

□

Therefore, substituting the derived φ_k from Theorem F.1 into the UBA-equivalent family (Definition equation 2.1), match the iteration-wise extended formula φ_t(under iteration-wise framework, $\Delta T_k = 1$) with the scale by equation 68, we obtain the optimal step size at $t \in [T_k, T_{k+1}]$ on segment k as follows:

$$\eta_k(t) = \frac{2(1+\cos\theta_k(t))}{2\varphi_k + (2-\varphi_k)(1+\cos\theta_k(t))}, \tag{85}$$

where

$$\theta_k(t) := \frac{\pi(t-T_k)}{\Delta T_k} + (k-1)\pi,$$

and the curvature parameter is obtained by substituting the plug-in estimate $U_t^* = \frac{1}{1+\sqrt{\varphi_0/2}\left\|\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}\right\|}$ into the matching solution $\varphi_t = \frac{2(1-U_t^*)^2}{(U_t^*)^2}$ from Theorem F.1, which simplifies to the closed form

$$\varphi_t = \varphi_0 \left\| \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} \right\|^2. \tag{86}$$

Global LRD by segment-wise UTA. The above derivation specifies an optimal *segment-wise* shape $\eta_k(t)$ under a locally approximation on $[T_k, T_{k+1}]$. We extend it to a learning-rate LRD on the entire horizon $[0, T]$ by concatenation:

$$\eta(t) := \eta_k(t), \qquad t \in [T_k, T_{k+1}], \quad k = 1, \dots, K. \tag{87}$$

This yields a well-defined global LRD whose shape is min-max optimal *within each segment* and whose scale is calibrated by matching the intrinsic time U_k^*.

Proposition F.2. *Consider the segment dynamics on $[T_k, T_{k+1})$, the end-of-segment state satisfies*

$$\|W_{T_{k+1}} - \bar{W}^{(k)}\| \le \rho_k \|W_{T_k} - \bar{W}^{(k)}\| + \left(\varepsilon_\varsigma^{(k)} + \eta_{\max}\varepsilon_M^{(k)}\|W_{T_k} - \bar{W}^{(k)}\|\right)\Delta T_k \rho_k, \tag{88}$$

where $\eta_{\max} := \max_{t\in[T_k,T_{k+1})}\eta_t$, $\rho_k := \|\prod_{t=T_k}^{T_{k+1}-1}(I-\eta_t\bar{H}_k)\|$, $\bar{W}^{(k)}$ is the local minimum in k-th segment. ς_t is an additive perturbation collecting by (i) local quadratic residuals and (ii) momentum or preconditioner approximation errors. Throughout the segment we assume, $\|M_t - \bar{H}_k\| \le \varepsilon_M^{(k)}$ and $\|\varsigma_t\| \le \varepsilon_\varsigma^{(k)}$, where M_t denotes the preconditioned Hessian estimate used at each step.

Proof. In the ideal segment-wise steady model, the dynamic can be estimated by segment-wise $\bar{H}_k$ for each segment, as equation 19,

$$W_{T_{k+1}} - \bar{W}^{(k)} = \left(I - \eta_{T_{k+1}-1}\bar{H}_k\right)\cdots\left(I - \eta_{T_k+1}\bar{H}_k\right)\left(W_{T_k+1} - \bar{W}^{(k)}\right). \tag{89}$$

In a real system, the true segment dynamics on $[T_k, T_{k+1})$ is

$$W_{t+1} - \bar{W}^{(k)} = (I - \eta_t M_t)\left(W_t - \bar{W}^{(k)}\right) + \varsigma_t, \qquad t \in [T_k, T_{k+1}), \tag{90}$$

where $M_t \succeq 0$ and $M_t = \bar{H}_k + \Delta M_t$, and ς_t is an additive perturbation collecting by (i) local quadratic residuals and (ii) momentum or preconditioner approximation errors.

Substitute $M_t = \bar{H}_k + \Delta M_t$ into equation 90:

$$W_{t+1} - \bar{W}^{(k)} = (I - \eta_t \bar{H}_k)\left(W_t - \bar{W}^{(k)}\right) + \hat{\varsigma}_t, \qquad t \in [T_k, T_{k+1}), \tag{91}$$

where $\hat{\varsigma}_t := \varsigma_t - \eta_t \Delta M_t \left(W_t - \bar{W}^{(k)}\right)$.

Iterating equation 91 over the segment gives

$$W_{T_{k+1}} - \bar{W}^{(k)} = A_k\left(W_{T_k} - \bar{W}^{(k)}\right) + \sum_{t=T_k}^{T_{k+1}-1}\left(\prod_{s=t+1}^{T_{k+1}-1}(I - \eta_s\bar{H}_k)\right)\hat{\varsigma}_t, \tag{92}$$

where A_k is defined as

$$A_k := \prod_{t=T_k}^{T_{k+1}-1}(I - \eta_t\bar{H}_k), \qquad \rho_k := \|A_k\|. \tag{93}$$

.

Taking norms in equation 92 yields

$$\|W_{T_{k+1}} - \bar{W}^{(k)}\| \leq \|A_k\| \cdot \|W_{T_k} - \bar{W}^{(k)}\| + \sum_{t=T_k}^{T_{k+1}-1} \left\| \prod_{s=t+1}^{T_{k+1}-1} (I - \eta_s \bar{H}_k) \right\| \cdot \|\hat{\varsigma}_t\|. \tag{94}$$

By definition $\|A_k\| = \rho_k$, it gives,

$$\|W_{T_{k+1}} - \bar{W}^{(k)}\| \leq \rho_k \|W_{T_k} - \bar{W}^{(k)}\| + \Delta T_k \rho_k \|\hat{\varsigma}_t\|. \tag{95}$$

By definition $\hat{\varsigma}_t := \varsigma_t - \eta_t \Delta M_t \left(W_t - \bar{W}^{(k)}\right)$ and triangle inequality,

$$\|\hat{\varsigma}_t\| \leq \|\varsigma_t\| + \eta_t \|\Delta M_t\| \cdot \|W_t - \bar{W}^{(k)}\|.$$

Assume that for $t \in [T_k, T_{k+1})$, $\|\Delta M_t\| \leq \varepsilon_M^{(k)}$ and $\|\varsigma_t\| \leq \varepsilon_\varsigma^{(k)}$. We have

$$\|\hat{\varsigma}_t\| \leq \varepsilon_\varsigma^{(k)} + \eta_{\max} \varepsilon_M^{(k)} R_k. \tag{96}$$

where $R_k := \max_{t \in [T_k, T_{k+1}]} \|W_t - \bar{W}^{(k)}\|$ is the segment radius.

Substituting this into equation 95 gives

$$\|W_{T_{k+1}} - \bar{W}^{(k)}\| \leq \rho_k \|W_{T_k} - \bar{W}^{(k)}\| + \left(\varepsilon_\varsigma^{(k)} + \eta_{\max} \varepsilon_M^{(k)} \|W_{T_k} - \bar{W}^{(k)}\|\right) \Delta T_k \rho_k.$$

This shows the deviation from the segment prediction is controlled by $(\varepsilon_\varsigma^{(k)}, \varepsilon_M^{(k)})$ and the segment length ΔT_k. □

Thus, concatenating segment-wise near-optimal LRDs (which minimize ρ_k) yields a global LRD whose deviation from the prediction vanishes as $\Delta T_k \to 0$ and $(\varepsilon_\varsigma^{(k)}, \varepsilon_M^{(k)}) \to 0$.

Corollary F.3. *Under the conditions of Proposition F.2, the concatenated LRD $\eta(t) = \eta_k(t)$ for $t \in [T_k, T_{k+1})$ yields the global bound*

$$\|u_T\| \leq \left(\prod_{k=1}^{K} \rho_k\right) \|u_0\| + \sum_{i=1}^{K} \left(\Delta_i \prod_{j=i+1}^{K} \rho_j\right), \tag{97}$$

where $\rho_k := \|\prod_{t=T_k}^{T_{k+1}-1} (I - \eta_t \bar{H}_k)\|$ and $\Delta_k := \left(\varepsilon_\varsigma^{(k)} + \eta_{\max} \varepsilon_M^{(k)} \|W_{T_k} - \bar{W}^{(k)}\|\right) \Delta T_k \rho_k$ is the segment-wise accumulated perturbation term in Proposition F.2. In particular, when Δ_k is small (short segments), minimizing ρ_k segment-wise yields a near-optimal global contraction.

Justification of concatenation. Proposition F.2 and Corollary F.3 further show that under mild non-stationarity, the concatenated LRD remains near-optimal, with deviation controlled by the segment length, curvature and preconditioner drift. The segment length ΔT_k controls the approximation error. Shorter segments yield smaller trajectory radius R_k and smaller curvature drift. Hence, the piecewise construction becomes more accurate as $\max_k \Delta T_k$ decreases. Consequently, to achieve the best possible approximation within our framework, we set the segment length to its limit value, $\Delta T_k = 1$. This implies that each segment corresponds to a single optimization step, thereby equating the segment index k with the time index t (so that, for instance, φ_k becomes φ_t). The resulting final UTA LRD is therefore defined per time step as follows.

Derived UTA LRD. Although the parameters $(\eta_{\min}^{(k)}, \eta_{\max}^{(k)}, \varphi_k)$ may vary across segments, the LRD on every segment belongs to the same UBA-equivalent function family. Therefore, the resulting global LRD equation 87 is a *piecewise UBA* LRD on $[0, T]$. To avoid discontinuities at segment boundaries, we set the starting endpoint of the next segment to match the ending value of the previous segment, $\eta_{\min}^{(k)} := \eta_{\max}^{(k+1)}$ for decay LRDs (analogously for increasing LRDs). This keeps $\eta(t)$ continuous on $[0, T]$ while preserving the segment-wise optimal shape.

The derived LRD on the entire horizon $[0, T]$ admits the closed form

$$\eta(t) = g_{a,b}(\varphi, \theta(t)) = \eta_{\min} + (\eta_{\max} - \eta_{\min}) \frac{2(1+\cos\theta(t))}{2\varphi_t + (2-\varphi_t)(1+\cos\theta(t))}, \tag{98}$$

where $a = \eta_{\min}$, $b = \eta_{\max} - \eta_{\min}$, $\theta(t) := \frac{\pi(t)}{T}$, the curvature parameter φ_t is determined by the exact intrinsic-time matching condition

$$\varphi_t = \varphi_0 \left\| \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \right\|^2. \tag{99}$$

G Theoretical analysis

G.1 Functional scaling laws for LRDs

To theoretically validate the effectiveness of the UTA, we analyze it under the framework of the Functional Scaling Law(FSL) (Li et al., 2025). This FSL theory characterizes how the entire loss trajectory scales with training resources under given LRD. Formally, for a model with capacity exponent β and task difficulty s, the expected loss $\mathcal{E}_t$ at intrinsic training time t follows:

$$\mathcal{E}_t \eqsim \underbrace{M^{-s\beta}}_{\substack{\text{model} \\ \text{error}}} + \underbrace{e(t)}_{\substack{\text{model} \\ \text{learning}}} + \underbrace{\int_0^t \mathcal{K}(t-z)[e(z)+\sigma^2]\gamma(z)\mathrm{d}z}_{\text{noise accumulation}}, \tag{100}$$

where M denotes model size, σ^2 is label noise, $\mathcal{K}(\cdot)$ is a forgetting kernel that decays as $\mathcal{K}(t) := (1+t)^{-\left(2-\frac{1}{\beta}\right)}$, $e(t) := (1+t)^{-s}$ denotes the learning term, and $\gamma(z)$ encapsulates the joint effect of learning rate and batch size LRD. In addition, the notation $\eqsim$ denotes equivalence up to a constant factor.

The FSL reveals that different learning rate decays fundamentally affect training efficiency through two competing mechanisms: (i) accumulating sufficient *intrinsic time* $t = \int_0^\tau \eta(r)\mathrm{d}r$ for signal learning, and (ii) controlling noise injection via the LRD-dependent term $\gamma(z)$. This framework has enabled rigorous comparison of LRD strategies, establishing that WSD achieves strictly better scaling efficiency than both constant and exponential decay LRDs (Li et al., 2025).

G.2 Analytical framework for our LRD

Given the widespread adoption of cosine LRD and the fact that the cosine function forms a core component of our proposed LRD approach, we first analyze the theoretical effect of the cosine LRD.

Within the functional scaling law framework, the model error term is determined solely by model capacity. Therefore, the differences between LRDs manifest in the model learning term and noise accumulation term. Analysis reveals that: (i) Compared to cosine LRD, UTA achieves smaller values in both terms, provides superior control of expected loss $\mathcal{E}_t$ at intrinsic training time. (ii) Compared to WSD FSL (Li et al., 2025), UTA exhibits equivalent scaling behavior since its model learning and noise accumulation terms are of the same order.

Our analysis proceeds through the following sequence of steps. First, we formulate the learning rate decay in physical time $\phi(\tau)$. Next, we compute the transformation to the intrinsic time $\bar{T}(\tau) = \int_0^\tau \phi(z)\mathrm{d}z$ and examine its properties, thereby circumventing the need for an explicit inverse function. We then express the noise intensity $\gamma(z)$ within the intrinsic time coordinate. Finally, we substitute these results into the FSL formulation to derive both the signal learning term $e(t)$ and the noise accumulation term $\int_0^t \mathcal{K}(t-z)[e(z)+\sigma^2]\gamma(z)\mathrm{d}z$.

Since cosine function is a part of our LRD function, we first analyze the theoretical efficiency of cosine learning rate decay.

Theorem G.1 (Cosine LRD under FSL)**.** *Under the FSL assumptions and in the hard-learning regime $s \le 1 - 1/\beta$, consider the cosine-annealed learning-rate LRD*

$$\phi_{\cos}(\tau) = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})(1 + \cos(\pi\tau/T_{\max})), \qquad \tau \in [0, T_{\max}],$$

where $T_{\max}$ is the total training step number. Let $\bar{T}(\tau) = \int_0^\tau \phi_{\cos}(r)\, dr$ be the intrinsic time and $\bar{T}(T_{\max})$ the final intrinsic time. Then the FSL risk satisfies the form

$$\mathcal{E}_t \eqsim M^{-s\beta} + (1 + aT_{\max})^{-s} + \frac{\sigma^2}{B}\left[(a - bT_{\max}) + b\,\frac{\min\{M, (aT_{\max})^{1/\beta}\}}{aT_{\max}}\right], \tag{101}$$

where B denotes the batch size, $a = \frac{1}{2}(\eta_{\max} + \eta_{\min})$, $b = \frac{1}{2}(\eta_{\max} - \eta_{\min})$.

Proof. By direct integration,

$$\bar{T}(\tau) = \int_0^\tau \phi_{\cos}(r)\,\mathrm{d}r = a\tau + b\frac{T_{\max}}{\pi}\sin(\pi\tau/T_{\max}),$$

with $a = \frac{1}{2}(\eta_{\max} + \eta_{\min})$, $b = \frac{1}{2}(\eta_{\max} - \eta_{\min})$. Hence the final intrinsic time is $\bar{T}(T_{\max}) = aT_{\max}$. Since $\sin(\cdot)$ is bounded, from the pointwise bounds $\eta_{\min} \le \phi_{\cos}(\tau) \le \eta_{\max}$ we obtain

$$\eta_{\min}\tau \le \bar{T}(\tau) \le \eta_{\max}\tau, \qquad 0 \le \tau \le T, \tag{102}$$

which guarantees monotonicity and gives the corresponding inverse bounds

$$\frac{t}{\eta_{\max}} \le \bar{T}^{-1}(t) \le \frac{t}{\eta_{\min}}. \tag{103}$$

Then, under the Theorem 4.4 and Assumption 5.1 in FSL rules (Li et al., 2025), the FSL can be simplified to

$$\mathcal{E}_t \eqsim M^{-s\beta} + e(t) + \frac{\sigma^2}{B}\mathcal{N}(\phi),$$

where $e(t) := (1+t)^{-s}$, $\mathcal{N}[\gamma] := \int_0^t \mathcal{K}_M(t-z)\phi(\bar{T}^{-1}(t))\,\mathrm{d}z$. It gives the signal term as

$$e(t) = (1 + aT_{\max})^{-s}.$$

Moreover, the forgetting kernel can be approximated as

$$\mathcal{K}_M(t) \eqsim \sum_{j=1}^{M} j^{-2\beta}e^{-2j^{-\beta}t} \eqsim \int_1^M x^{-2\beta}e^{-2x^{-\beta}t}\mathrm{d}x \eqsim \int_{M^{-\beta}}^1 u^{1-1/\beta}e^{-2ut}\mathrm{d}u.$$

In addition, the noise intensity γ within the intrinsic time coordinate is $\gamma(t) = \frac{\phi_{\cos}(\bar{T}^{-1}(t))}{B}$, and $a + b(1 - \frac{\pi t}{T_{\max}}) \le \phi(\bar{T}^{-1}(t)) \le a + b(\pi - \frac{\pi t}{T_{\max}})$.

It gives $N(\phi)$ as

$$\begin{aligned}
\int_0^T \mathcal{K}_M(T-t)\phi_{\cos}(t)\,dt &\eqsim \int_0^T \int_{M^{-\beta}}^1 u^{1-1/\beta}e^{-2u(T-t)}\mathrm{d}u\left(\hat{a} - \hat{b}t\right)\mathrm{d}t \\
&= \int_{M^{-\beta}}^1 u^{1-1/\beta}e^{-2uT}\left(\int_0^T e^{2ut}(\hat{a} - \hat{b}t)\,\mathrm{d}t\right)\mathrm{d}u \\
&= \int_{M^{-\beta}}^1 u^{1-1/\beta}e^{-2uT}\left[\frac{\hat{a}}{2u}(e^{2uT} - 1) - \frac{\hat{b}}{2u}\left(Te^{2uT} - \frac{e^{2uT} - 1}{2u}\right)\right]\mathrm{d}u \\
&= \int_{M^{-\beta}}^1 \left[\frac{\hat{a}}{2u^{1/\beta}} - \frac{\hat{a}e^{-2uT}}{2u^{1/\beta}} - \frac{\hat{b}T}{2u^{1/\beta}} + \frac{\hat{b}(1 - e^{-2uT})}{4u^{1+1/\beta}}\right]du \\
&= \int_{M^{-\beta}}^1 \left[\frac{\hat{a} - \hat{b}T}{2u^{1/\beta}} - \frac{(\hat{a} - \hat{b}T + \hat{b}T)e^{-2uT}}{2u^{1/\beta}} + \frac{\hat{b}(1 - e^{-2uT})}{4u^{1+1/\beta}}\right]\mathrm{d}u \\
&= (\hat{a} - \hat{b}T)\int_{M^{-\beta}}^1 \frac{1 - e^{-2uT}}{2u^{1/\beta}}\mathrm{d}u + \hat{b}T\int_{M^{-\beta}}^1 \left(-\frac{e^{-2uT}}{2u^{1/\beta}} + \frac{1 - e^{-2uT}}{4Tu^{1+1/\beta}}\right)\mathrm{d}u,
\end{aligned}$$

where $\hat{a} := \pi\eta_{\max} \eqsim (a+b)\pi$, $\hat{b} = \frac{1}{2}(\eta_{\max} - \eta_{\min}) = b\frac{\pi}{T_{\max}}$ and total intrinsic time $\bar{T}(T_{\max}) = aT_{\max}$.

Then, according to Lemma F.5 and Lemma F.6 in FSL (Li et al., 2025), the noise term can be estimated as

$$N(\phi) \eqsim \left[\eta_{\max} - \frac{1}{2}\left(\eta_{\max}^2 - \eta_{\min}^2\right)\right]\pi + \frac{1}{2}\left(\eta_{\max} - \eta_{\min}\right)\pi \frac{\min\{M, (aT_{\max})^{1/\beta}\}}{aT_{\max}},$$

which gives

$$\mathcal{E}_t \eqsim M^{-s\beta} + (1 + aT_{\max})^{-s} + \frac{\sigma^2}{B}\left[\left(a + b - \frac{ab}{2}\right)\pi + b\pi \frac{\min\{M, (aT_{\max})^{1/\beta}\}}{aT_{\max}}\right].$$

Combining these contributions yields equation 101. □

Theorem G.2. *Under the FSL assumptions and in the hard-learning regime $s \le 1 - 1/\beta$. Let $\bar{T}(\tau) = \int_0^\tau \phi(r)\,dr$ be the intrinsic time and $\bar{T}(T_{\max})$ the final intrinsic time. When a sufficiently small φ_k is sufficiently small that causes the LRD to maintain a consistently high learning rate over a prolonged duration, analogous to WSD-like period, the FSL risk satisfies the form*

$$\mathcal{E}_t \eqsim M^{-s\beta} + \left[\frac{2 - \sqrt{2\varphi}}{2 - \varphi}(T_1 + T_2)\right]^{-s} + \frac{\sigma^2}{B}\left[\left(a + b - \frac{ab}{2}\right)\pi + b\pi \frac{\min\{M, \left(c\frac{2-\sqrt{2\varphi}}{2-\varphi}T_2\right)^{1/\beta}\}}{aT_2}\right].$$

where B denotes the batch size.

Proof. According to equation 72, it yields the integration of our proposed LRD in k-th segment,

$$U_k = \eta_{\min}^{(k)}\Delta T_k + \left(\eta_{\max}^{(k)} - \eta_{\min}^{(k)}\right)\Delta T_k \frac{2 - \sqrt{2\varphi_k}}{2 - \varphi_k}.$$

As $\frac{2-\sqrt{2\varphi_k}}{2-\varphi_k}$ decreases from 1 to 0, a sufficiently small φ_k causes the LRD to maintain a consistently high learning rate over a prolonged duration, analogous to the stable period observed in WSD (Hu et al., 2024). Assume that when $\varphi_k \le \mathrm{Z}$, the learning rate enters a stable phase, i.e.,$\mathbb{E}(\eta_t) = \frac{2-\sqrt{2\varphi}}{2-\varphi}\eta_{\max}$, where $r \le 1$ is a constant close to 1; when $\varphi_k > \mathrm{Z}$, the learning rate decay enters a phase of significant decay. Let $T_1 := |T_{\varphi_k \le \mathrm{Z}}|$ denote the duration of stable stage, $T_2 := T_{\max} - T_{\varphi_k > \mathrm{Z}}$ denote the duration of this pronounced decay stage.

For T_1 stage, since $\mathbb{E}(\eta_t) = \frac{2-\sqrt{2\varphi}}{2-\varphi}\eta_{\max}$, according to FSL of constant LRD,

$$\mathcal{E}_t \eqsim M^{-s\beta} + \left(\frac{2 - \sqrt{2\varphi}}{2 - \varphi}\eta_{\max}T_1\right)^{-s} + \frac{\frac{2-\sqrt{2\varphi}}{2-\varphi}\eta_{\max}\sigma^2}{B}.$$

For T_2 stage, we analyze the case where φ_k is invariant with k. Denote $\varphi_k \equiv \varphi$ and $K_\tau = \{K : \tau \in (T_K, T_{K+1}]\}$. If $\eta_{\max}^{(k)} = \eta_{\min}^{(k)} + \epsilon$ where ϵ is a very small constant, and $\Delta T_k = 1$, it gives

$$\begin{aligned}
\bar{T}(\tau) =& \sum_{k=1}^{K_\tau} \eta_{\min}^{(k)}\Delta T_k + \left(\eta_{\max}^{(k)} - \eta_{\min}^{(k)}\right)\Delta T_k \frac{2 - \sqrt{2\varphi}}{2 - \varphi} \\
=& \frac{\tau}{2}\left[\left(\eta_{\min}^{(1)} + \eta_{\min}^{(K_\tau)}\right) + \left(\eta_{\max}^{(1)} + \eta_{\max}^{(K_\tau)} - \eta_{\min}^{(1)} + \eta_{\min}^{(K_\tau)}\right)\frac{2 - \sqrt{2\varphi}}{2 - \varphi}\right] \\
\eqsim& c\frac{2 - \sqrt{2\varphi}}{2 - \varphi}\tau,
\end{aligned}$$

where c is a constant and $c \le \eta_{\max}$. Thus we have

$$\tau = \bar{T}^{-1}(t) = \frac{2 - \varphi}{c\left(2 - \sqrt{2\varphi}\right)}t.$$

Furthermore, the total intrinsic time is

$$\bar{T}(T_{\max}) \eqsim c\frac{2 - \sqrt{2\varphi}}{2 - \varphi}T_{\max}.$$

In addition, since

$$\phi = \eta_{\min}^{(k)} + (\eta_{\max}^{(k)} - \eta_{\min}^{(k)}) \frac{2(1+\cos\theta_k(t))}{2\varphi_k + (2-\varphi_k)(1+\cos\theta_k(t))}, \tag{104}$$

where $\theta_k(t) := \frac{\pi(t-T_k)}{\Delta T_k} + (k-1)\pi$, we have the LRD-adjusted function in intrinsic time is given by

$$\phi\left(\bar{T}^{-1}(t)\right) = \eta_{\min}^{(k)} + (\eta_{\max}^{(k)} - \eta_{\min}^{(k)}) \frac{2(1+\cos\hat{\theta}_k(t))}{2\varphi_k + (2-\varphi_k)(1+\cos\hat{\theta}_k(t))}.$$

where $\hat{\theta}_k(t) := \frac{\pi\left(\frac{2-\varphi}{c(2-\sqrt{2\varphi})}(t-T_k)\right)}{\Delta T_k} + (k-1)\pi$.

Similar to the proof in Theorem G.1, the noise intensity γ within the intrinsic time coordinate is $\gamma(t) = \frac{\phi_{\cos}(\bar{T}^{-1}(t))}{B}$, and $a + b(1 - \frac{2-\varphi}{c(2-\sqrt{2\varphi})}\frac{t\pi}{\Delta T_k}) \leq \phi(\bar{T}^{-1}(t)) \leq a + b(\pi - \frac{2-\varphi}{c(2-\sqrt{2\varphi})}\frac{t\pi}{\Delta T_k})$. The noise term can be estimated as

$$N(\phi) \eqsim \left[\eta_{\max} - \frac{1}{2}\left(\eta_{\max}^2 - \eta_{\min}^2\right)\right]\pi + \frac{1}{2}(\eta_{\max} - \eta_{\min})\,\pi \frac{\min\{M, (c\frac{2-\sqrt{2\varphi}}{2-\varphi}T_2)^{1/\beta}\}}{aT_2},$$

which gives

$$\mathcal{E}_t \eqsim M^{-s\beta} + (1 + c\frac{2-\sqrt{2\varphi}}{2-\varphi}T_2)^{-s} + \frac{\sigma^2}{B}\left[\left(a + b - \frac{ab}{2}\right)\pi + b\pi\,\frac{\min\{M, \left(c\frac{2-\sqrt{2\varphi}}{2-\varphi}T_2\right)^{1/\beta}\}}{aT_{\max}}\right].$$

Following the Theorem 5.4 in FSL (Li et al., 2025), the FSL risk of entire training dynamic under our proposed LRD is as follow

$$\mathcal{E}_t \eqsim M^{-s\beta} + \left[\frac{2-\sqrt{2\varphi}}{2-\varphi}(T_1 + T_2)\right]^{-s} + \frac{\sigma^2}{B}\left[\left(a + b - \frac{ab}{2}\right)\pi + b\pi\,\frac{\min\{M, \left(c\frac{2-\sqrt{2\varphi}}{2-\varphi}T_2\right)^{1/\beta}\}}{aT_2}\right].$$

□

H Experimental Implementation Details

H.1 Baselines

We detail the baselines as follows,

- **Cosine LRD**, originally proposed for SGD with restarts (Loshchilov & Hutter, 2017), has become a de facto standard in modern deep learning and large scale model training due to its smooth transition and robust performance (Loshchilov & Hutter, 2017; Schaipp et al., 2025). The LRD follows

  $$\eta_t = \eta_{min} + \frac{1}{2}(\eta_0 - \eta_{min})(1 + \cos(\frac{t\pi}{T})),$$

  where T is total iterations, η_0 and $\eta_{\min}$ are the initial and final learning rates. Its widespread adoption in modern deep learning and large language models establishes it as an essential classical baseline.
- **Linear decay LRD** represents the simplest form of LRDd annealing, providing a transparent and easily interpretable comparison point. Following the implementation in (Li et al., 2019a), we define

  $$\eta_t = \eta_0 \times (1 - t/T),$$

  where T is total iterations, after an optional warmup phase. Despite its simplicity, linear decay remains competitive in resource-constrained scenarios (Li et al., 2019a) and serves as an important baseline for assessing whether more complex LRDs provide meaningful advantages over straightforward linear reduction.

- **WSD LRD** exemplifies modern empirical practices in large-scale model training, particularly for language models (Zhai et al., 2022; Hu et al., 2024). The LRD comprises three distinct phases, (1)Warmup phase: linear ramp from η_{start} to η_{peak} over T_{warmup}; (2) Stable phase: maintain η_{peak} for T_{stable} steps. (3) Decay phase: Linear or exponential decay to $\eta_{\min}$ (we use the WSD LRDr with 10% steps in the decay phase following previous works (Hu et al., 2024; Hägele et al., 2024)). Its empirical success in large language models makes WSD a critical baseline for evaluating LRDs in realistic large-scale scenarios.
- **UBA LRD** is a theoretically principled and empirically strong latest baseline. It derives from a min-max problem on optimization dynamics. The LRD takes the form:

$$\eta_t = (\eta_{\max} - \eta_{\min}) \frac{2\left[1 + \cos\left(\frac{[2(t-T_k)-1]\pi}{2(T_{k+1}-T_k)}\right)\right]}{2\varphi + (2-\varphi)\left[1 + \cos\left(\frac{[2(t-T_k)-1]\pi}{2(T_{k+1}-T_k)}\right)\right]} + \eta_{\min},$$

  where φ is a hyper-parameter. We use the default setting (Tang et al., 2026), fixing the $\varphi = 0.5$. Recent systematic evaluations (Tang et al., 2026) rank UBA among the top-performing LRDs across diverse architectures and tasks, making it a particularly strong benchmark for theoretical LRDs.

H.2 LRD SETTING

For our UTA LRD, we fix the unique hyper-parameter $\varphi_0 = 0.15$. For UBA LRD, we use the default setting (Tang et al., 2026), fixing the $\varphi = 0.5$. For WSD LRD, we use 10% steps in the decay phase following the setting in previous works (Hu et al., 2024; Hägele et al., 2024).

H.3 MODEL ARCHITECTURES AND HYPER-PARAMETERS

Settings on LLaMA Following Lialin et al. (2023) and Zhao et al. (2024), we pretrain vanilla LLaMA models from scratch on the C4 dataset (Raffel et al., 2020). We consider four model scales with 60M, 130M, 350M, and 1B parameters. For each scale, we evaluate four training budgets of 10K, 15K, 20K, and 25K steps. The corresponding architectural hyperparameters and token budgets are summarized in Tab. 4. Across all experiments, we set the batch size to 512 and the maximum sequence length to 1024, and train the models using AdamW with $\beta_1/\beta_2 = 0.9/0.95$. We use a peak learning rate of 1e-3 for the 60M–350M models, and 5e-4 for the 1B model. For a fair comparison across LRDs, we apply a learning rate warmup over the first 10% of training steps for all experiments, and enforce a minimum learning rate equal to 10% of the peak learning rate for all LRDrs. Since UTA has the phase drift mechanism that keeps ascending phase at early stage, we also preserve its original warmup-like behavior.

Table 4: LLaMA architecture hyper-parameters for evaluation. Data amounts are specified in tokens.

Params	Hidden Dim	Intermediate Dim	Heads	Layers	Training Steps	Data amount (Tokens)
60M	512	1376	8	8	10k	5.2B
130M	768	2048	12	12	15k	7.9B
350M	1024	2736	16	24	20k	10.5B
1 B	2048	5461	24	32	25k	13.1B

Settings on ResNet. For ResNet, we adopt the A2 training recipe introduced by the work Wightman et al. (2021), which employs stronger data augmentation and has been shown to substantially improve ResNet performance. The augmentation pipeline includes random cropping, random horizontal flipping, Mixup ($\alpha = 0.1$), CutMix ($\alpha = 1.0$) applied with 50% probability, and RandAugment ($M = 7$, $N = 2$, MSTD $= 0.5$). For all experiments, We train the model 300 epochs with 2048 batch size. For training loss, we use a binary cross-entropy loss. We use AdamW with a weight decay of 0.05 and $\beta_1/\beta_2 = 0.9/0.999$, and set the peak learning rate to 1e-3. Following Win (Zhou et al., 2024), we incorporate a 5-epoch warmup phase for all baseline LRDs, starting from an initial learning rate of 1e-8. For LRDs that natively include an ascending phase (e.g., the WSD), we preserve their original warmup-like behavior during that period. Since UTA has the phase drift mechanism that keeps ascending phase at early stage, we also preserve its original warmup-like behavior. Additionally, we apply a 10-epoch cooldown phase at the end of training, with a minimum learning rate of 1e-5.

Settings on ViT. For ViT, we follow the widely-used official training setting of ViTs described in the work (Touvron et al., 2021; Yu et al., 2022). For this setting, data augmentation consists of random crop, horizontal flipping with probability, Mixup with parameter 0.8 (Zhang et al., 2017), CutMix with parameter 1.0 and probability 0.5 (Yun et al., 2019), RandAugment (Cubuk et al., 2020) with M = 9, N = 2 and MSTD = 0.5, and Random Erasing with parameter p = 0.25. For all experiments, We train the models under 300 epochs with 1024 batch size. For training loss, we use cross entropy loss. We use the AdamW optimizer with a weight decay of 0.1 and $\beta_1/\beta_2 = 0.9/0.999$. The peak learning rate is set to 1e-3. Additionally, We adopt the same warmup and cooldown settings for baselines as in the ResNet experiments.

H.4 Hardware and software environment

All experiments are conducted on hardware consisting of NVIDIA H800.

H.5 Datasets

C4. C4 dataset is a colossal, cleaned version of Common Crawl's web crawl corpus, which is mainly intended to pre-train language models and word representations (Raffel et al., 2020). To best simulate the practical pre-training scenario, we train without data repetition over a sufficiently large amount of data.

ImageNet. To assess the scalability of our LRD on more demanding and realistic visual recognition tasks, we perform extensive experiments on the ImageNet dataset. Widely recognized as the de-facto benchmark for large-scale image classification, ImageNet comprises approximately 1.28 million training images and 50,000 validation images spanning 1,000 diverse classes, and is accompanied by an official, standardized data split.

CIFAR100. CIFAR-100 is a widely recognized benchmark for evaluating the performance and efficiency of deep learning models. The dataset consists of 60,000 color images, each with a resolution of 32×32 pixels, categorized into 100 fine-grained classes. These classes are further organized into 20 super-classes, with each class containing 600 images. Every image is annotated with both a fine label, indicating its specific class, and a coarse label, indicating its super-class).

H.6 Code and reproducibility

Our implementation strictly follows the PyTorch official API, i.e. it is inherited from the base class

`torch.optim.lr_LRDr.LRDcheduler.`

It maintains full compatibility with all

`torch.optim`

optimizers. The complete production-ready code is publicly available.

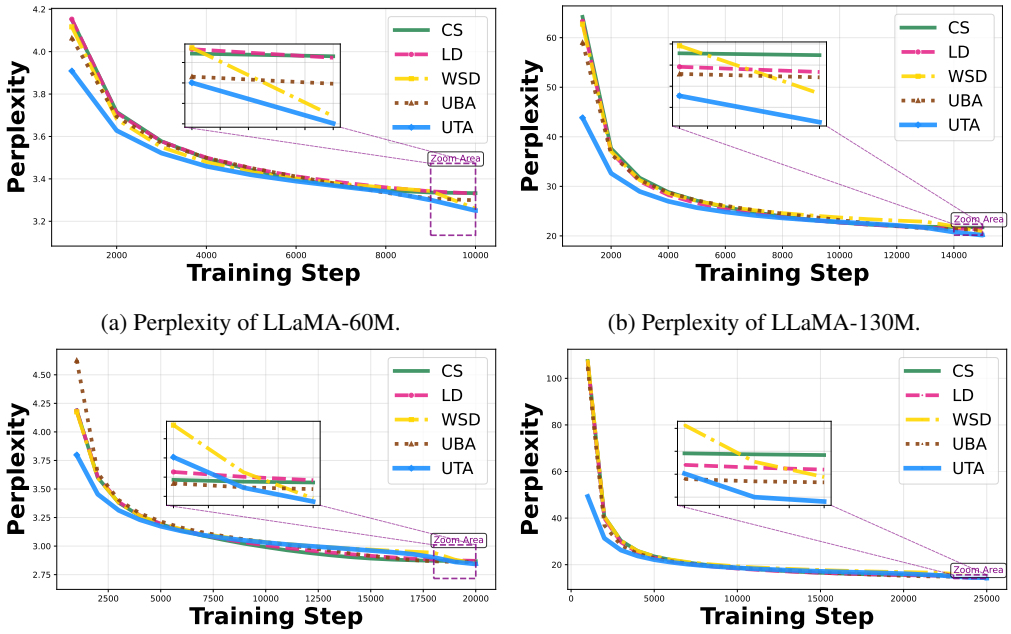
The code of UBA LRD is available at https://anonymous.4open.science/r/UTA-LRD-7B39/UTA.py.

We conduct an extensive evaluation comprising **40-60 experimental runs**, **requiring approximately 5000 H800 GPU-hours**. It is a substantial commitment for a single-study paper (context: DeepSeek-V3 consumed 3M H800 GPU-hours). This computational scale ensures statistically reliable conclusions and validates the robustness and scalability of our findings. The systematic, large-scale validation on a modern Transformer architecture provides strong evidence for UTA's practical utility and impact in real-world training scenarios.

I EXPERIMENTAL RESULTS

I.1 PERFORMANCE ON LANGUAGE MODEL-LLAMA

We evaluate on four scales of the LLaMA architecture: 60M, 130M, 350M, and 1B parameters. Performance is measured by perplexity on a held-out validation set, which directly reflects the model's generative uncertainty and language modeling quality.



(a) Perplexity of LLaMA-60M.

(b) Perplexity of LLaMA-130M.

(c) Perplexity of LLaMA-350M.

(d) Perplexity of LLaMA-350M.

Figure 2: Perplexity curves of LLaMA models.

Table 5: Perplexity ($e^{\text{validation loss}}$) on LLaMA models. Lower is better (↓). 'D-10' and 'D-0' denote final learning rates decaying to 10% and 0% of the peak rate, respectively.

LRD	LLaMA Model Size			
	60M		130M	
	D-10	D-0	D-10	D-0
CS	28.03	28.27	21.76	21.99
LD	27.99	28.21	21.43	21.89
WSD	26.03	26.93	20.83	21.18
UBA	27.01	27.23	21.12	21.00
UTA (Ours)	**25.80**	**26.77**	**20.16**	**20.68**

I.2 PERFORMANCE ON IMAGENET

We further validate the generalization of UTA on the ImageNet classification benchmark using ResNet-50 and ViT-B architectures. This selection tests generalization across model depth, width, and fundamental architectural biases. Tab. 6 reports the test top-1 accuracy. UTA consistently achieves the highest accuracy across both models. Specifically, it improves ResNet-50 accuracy to 78.73%, outperforming the strongest baseline UBA (78.38%) by 0.35%. For ViT-B, UTA reaches 82.30%, surpassing the competitive UBA LRD (82.05%) by 0.25%. These gains are statistically meaningful given the scale and complexity of ImageNet These consistent outperformance demonstrate that UTA generalizes beyond language modeling to vision tasks. It has robust and architecture-agnostic

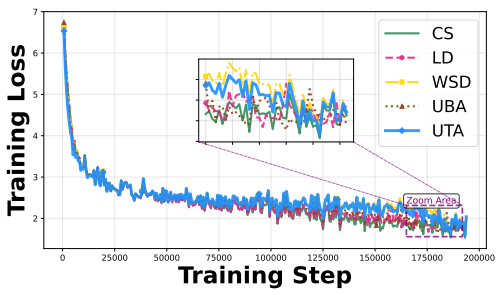
advantage in large-scale, real-world training scenarios. and generalizes beyond language modeling to vision tasks.

Table 6: ImageNet test top-1 Acc. (%) of different LRDs on ResNet50 and ViT-B under 300 epochs. Higher values indicates better performance (↑). Best results in bold.

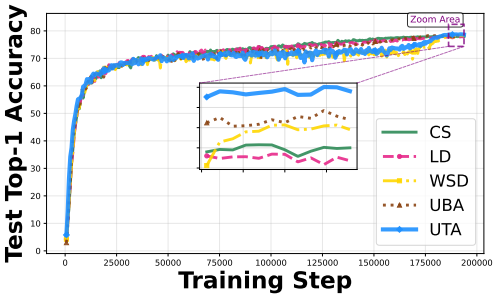
LRD	ResNet-50	ViT-B
CS	$78.34_{\pm 0.024}$	$79.92_{\pm 0.057}$
LD	$77.90_{\pm 0.028}$	$79.96_{\pm 0.104}$
WSD	$78.30_{\pm 0.024}$	$81.14_{\pm 0.048}$
UBA	$78.38_{\pm 0.014}$	$82.05_{\pm 0.041}$
UTA (Ours)	$\mathbf{78.73}_{\pm 0.029}$	$\mathbf{82.30}_{\pm 0.037}$

We also plot the training loss and validation accuracy across training steps for LRDs on ImageNet using ResNet-50 and ViT-B models in Fig. 3. A key and consistent observation is that while our proposed UTA LRDr does not always achieve the lowest training loss, it attains the highest final validation accuracy. Specifically, the training loss curves in Fig. 3(a) and (c) show that, unlike empirically tuned LRDs, such as CS and LD, that aggressively push training loss to very low values, UTA maintains a steady descent accompanied by moderate fluctuations. Its training loss is never the absolute minimum. This is not a sign of inadequate optimization, but a direct consequence of the robust-optimization design of UTA equation 3. Derived from a min-max framework that minimizes the learning rate impact under worst-case conditions, UTA implicitly introduces a regularization effect that prevents over-fitting to the training data. Consequently, although training loss is not minimized as aggressively, UTA consistently surpasses all baselines on validation accuracy as training progresses, demonstrating enhanced generalization rooted in its robust formulation.

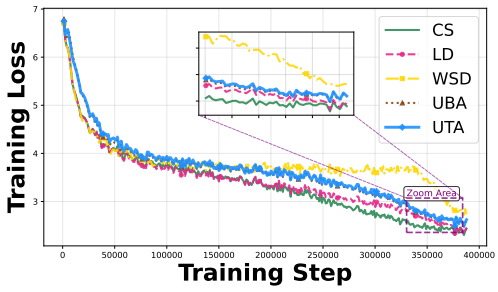
Moreover, the learning rate trajectory of UTA exhibits noticeable fine-grained fluctuations, more pronounced than those of static or predefined LRDs. These fluctuations are not noise. They reflect a state-aware, active adaptation mechanism that responds to the internal dynamics of training. By dynamically adjusting the learning rate according to local landscape stability, UTA achieves a more robust optimization trajectory, ultimately leading to superior generalization performance.



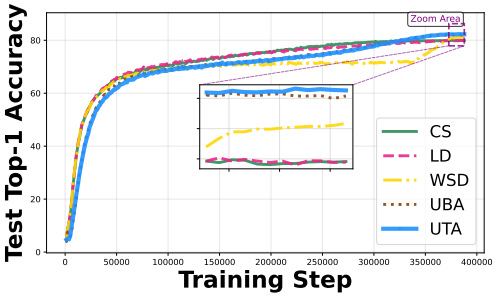
(a) Training Loss on ImageNet-ResNet50



(b) Validation Accuracy on ImageNet-ResNet50.



(c) Training Loss on ImageNet-ViT-B.



(d) Validation Accuracy on ImageNet-ViT-B.

Figure 3: Training loss and validation accuracy on ImageNet using ResNet50 and ViT-B.

Table 7: Cross-φ_0 ablation studies: Perplexity of LLaMA-60M using UTA under different φ_0.

Network	UTA									
	φ_0=0.01	φ_0=0.05	φ_0=0.1	φ_0=0.15	φ_0=0.2	φ_0=0.25	φ_0=0.5	φ_0=1	φ_0=2.5	φ_0=5
LLaMA-60M	26.18	25.94	25.85	25.81	25.80	25.80	25.92	26.04	26.39	26.78

Table 8: Cross-φ_0 ablation studies: Perplexity of LLaMA-130M using UTA under different φ_0.

Network	UTA				
	φ_0=0.15	φ_0=0.2	φ_0=0.25	φ_0=0.5	φ_0=2.5
LLaMA-130M	20.63	20.65	20.66	20.67	20.99

I.3 Ablation study

Peak LR sensitivity analysis. To make a fair comparison, our main experiments used a fixed peak LR across all LRDs. Although these configurations are widely adopted defaults that have been empirically validated by prior works, to prevent peak LR from disadvantaging certain baselines if their optimal peak LRs differ, we conduct a systematic learning rate sweep on LLaMA-60M and 130M models. For 60M, we sweep peak LRs from 5×10^{-4} to 1×10^{-2}; for 130M, from 1×10^{-3} to 7×10^{-3}. All LRDs are trained under the same step budgets as in the main experiments, and we report perplexity ($e^{\text{validation loss}}$) for direct comparison. Results are summarized in Tab. 3.

First, **Optimal peak LRs differ across LRDs.** As expected, we observe that different LRDs indeed favor different peak learning rates. On LLaMA-60M, CS, LD and UBA achieve their best at 7×10^{-3}, WSD at 1×10^{-3}, while UTA achieves its best jointly at 1×10^{-3} and 7×10^{-3} with a perplexity of 25.80. On LLaMA-130M, CS and LD peak at 5×10^{-3}, WSD also at 5×10^{-3}, whereas UBA and UTA achieve their best at 3×10^{-3} and 1×10^{-3}, respectively. This confirms that a fixed LR across all LRDs is not ideal and underscores the necessity of per-LRD tuning for fair comparison.

Second, **UTA achieves the best overall performance under its optimal LR.** When comparing the best result per LRD (reported in the "Best from LR Sweep" column), UTA consistently outperforms all baselines on both model scales. On 60M, UTA reaches a best perplexity of **25.80**, surpassing the best baseline WSD (26.03) by 0.23 and UBA (26.26) by 0.46. On 130M, UTA achieves **20.16**, outperforming WSD (20.82) by 0.66 and UBA (20.78) by 0.62. These margins, while modest in absolute terms, are consistent and non-trivial in the context of language modeling validation loss.

Parameter analysis of φ_0 In this subsection, we perform a sensitivity analysis of the hyper-parameter $\varphi_0 \in \{0.01, 0.05, 0.1, 0.15, 0.20.25, 0.5, 1.0, 2.5, 5\}$ in equation 7 , which controls the variation speed of learning rate. To efficiently identify general parameter trends at low cost, then verify their scalability and robustness on larger models. We further examine $\varphi_0 \in \{0.1, 0.15, 0.20.25, 2.5\}$ on a larger 130M model. We narrow the range of φ_0 to focus computational resources on the most promising intervals, filtering out ineffective values identified in the initial broad search. The results on LLaMA-60M listed in Tab. 7. The results on LLaMA-130M listed in Tab. 8).

Our sensitivity analysis reveals three key findings. **First, UTA exhibits remarkable robustness to φ_0 over a wide range.** On LLaMA-60M, UTA achieves perplexity below 26.0 for **7 out of 10** tested φ_0 values (0.05 through 0.5), and the performance variation across this broad range is remarkably small, spanning only 25.80 to 25.92. Even at the extremes ($\varphi_0 = 0.01$ and $\varphi_0 = 5$), the degradation remains modest (26.18 and 26.78, respectively). On LLaMA-130M, UTA maintains perplexity between 20.63 and 20.67 across $\varphi_0 \in [0.15, 0.5]$, which is a variation of merely 0.04, and only degrades to 20.99 at $\varphi_0 = 2.5$. This **flat response surface** indicates that UTA's performance is largely insensitive to the choice of φ_0, a desirable property for practical deployment where extensive hyper-parameter tuning is often computationally prohibitive.

Second, **the performance trend aligns with our theoretical interpretation.** As φ_0 decreases from 0.5 to 0.15, performance improves slightly (from 25.92 to 25.80 on 60M, and from 20.67 to 20.63 on 130M). This is consistent with our continuous-time analysis: a smaller φ_0 corresponds to a larger intrinsic step U^*, enabling more thorough exploration of the loss landscape. However, the

gains saturate beyond $\varphi_0 = 0.15$, suggesting that once the intrinsic step is sufficiently large, further expansion yields diminishing returns—a phenomenon that aligns with the diminishing sensitivity predicted by our theoretical framework.

Third, and most importantly, **UTA demonstrates significantly superior robustness compared to the baseline UBA.** While both methods share a similar functional form of the learning rate decay, UBA exhibits a much narrower effective range: its performance declines sharply for $\varphi_0 < 0.5$, and further reduction beyond this point brings little improvement or even degradation. In contrast, UTA maintains strong performance across a **10-fold wider range** of φ_0 (from 0.05 to 0.5 on 60M), with a performance spread of only 0.24, compared to UBA's much larger degradation outside its narrow optimal region. We attribute this advantage to UTA's learning rate adaptation mechanism, which dynamically adjusts the intrinsic time scale in response to optimization geometry, allowing the optimizer to fully exploit the expanded exploration capacity afforded by larger U^* while automatically compensating when φ_0 is suboptimal. This adaptive coupling between the learning rate and the preconditioned gradient norm, which is absent in UBA, ensures that increased intrinsic time translates effectively into improved convergence and generalization, regardless of the specific φ_0 choice.

Overall, this sensitivity analysis confirms that UTA is not only performant but also **robust and easy to use**: it achieves near-optimal performance across a broad spectrum of φ_0 values, substantially reducing the burden of hyper-parameter tuning compared to both UBA and other schedule-based methods.

The effect of final LR in decay Regarding the LRD implementations, some practices use decay to 10% of the peak LR, some use 0%. To investigate whether this choice affects performance. We conduct this ablation on LLaMA-60M and 130M models, comparing both settings for all LRD methods. Results are reported in Tab. 5 in Appendix.

We observe that decaying to 0% consistently yields worse perplexity than decaying to 10% across nearly all methods and model scales. For instance, on LLaMA-130M, linear decay degrades from 21.43 to 21.89 (0.46), and even our UTA increases from 20.16 to 20.68 (0.52) when using 0% decay. Aggressively decaying the learning rate to zero prematurely reduces the model's ability to adapt, as the optimizer enters a near-stagnant regime before the model has fully converged.

Importantly, regardless of the decay ratio, UTA consistently outperforms all baselines. Under the 0% setting, UTA achieves 26.77 on 60M and 20.68 on 130M, still surpassing the best baselines (WSD at 26.93 and UBA at 21.00) by margins of 0.16 and 0.32, respectively. This confirms that our main conclusions are robust to the choice of final learning rate, and that the default 10% decay setting does not unfairly advantage UTA.

The role of dynamic state-aware mechanism To validate the contribution of the dynamic mechanism in our UTA LRD, we conduct an ablation study comparing the full version (Ours-Full) against a static variant (Ours-Static). We conduct the ablation study on the CIFAR-100 dataset with a ResNet-34 architecture. The CIFAR-100 dataset with ResNet-34 provides a balanced testbed. It is sufficiently complex to reveal LRD differences yet computationally tractable for ablation studies. We intentionally keep all other factors identical to isolate the dynamic mechanism's effect. In the static variant, the dynamic part is removed. All other hyper-parameters and training settings remain identical.

Table 9: Ablation Study: Impact of Dynamic State-Aware Mechanism.

	Ours-Static	Ours-Full	Improvement
CIFAR100-ResNet34(↑)	76.86	77.81	0.95
ImageNet-ResNet50(↑)	78.23	78.55	0.32
LLaMA60M(↓)	27.38	26.78	0.60
LLaMA130M(↓)	21.02	20.15	0.87

Comparison with automatic LR methods. While our primary baselines focus on LRDs, another line of LR research is automatic LR methods such as D-Adaptation (Defazio & Mishchenko, 2023) and Prodigy (Mishchenko & Defazio, 2023). These methods dynamically adjust the learning rate

during training *without predefined schedules*. Although these methods are not the main focus of our work, we conduct a comparison to position UTA against them.

Specifically, we evaluate D-Adaptation and Prodigy on LLaMA-60M, 130M, and 350M under a fixed 10K-step training budget, following their default hyperparameter configurations. Results are reported in Tab. 10. As shown, UTA consistently outperforms both automatic schedulers across all three model scales, achieving average perplexity reductions of 1.2 and 0.9 over D-Adaptation and Prodigy, respectively. We hypothesize that UTA's advantage stems from its schedule-aware shape design, which explicitly accounts for the remaining training horizon, whereas automatic methods rely solely on gradient statistics and may lack such long-term planning capability.

Given the clear superiority of UTA and the high computational cost of scaling automatic methods to larger models, we exclude the 1B model from this comparison and focus our main experiments on carefully designed schedule baselines.

Table 10: Perplexity ($e^{\text{validation loss}}$) comparison between automatic LR methods (D-Adaptation, Prodigy) and our UTA on LLaMA models. Lower is better (↓).

Method	60M	130M	350M
D-Adaptation	25.80	20.65	17.64
Prodigy	27.37	21.26	17.97
UTA (Ours)	**25.80**	**20.16**	**17.17**

Curvature alignment between UTA and UBA. We observe that UTA belongs to UBA family, and the effective scale of UTA hyperparameter is modulated by the RMS norm of the preconditioned update (approximately 0.2) (Liu et al., 2025; Kosson et al., 2023), whereas UBA's hyperparameter φ acts directly. To isolate and evaluate the contribution of UTA's state-adaptive mechanism—independent of hyperparameter scaling—we conduct a parameter-alignment experiment. This ensures that the underlying trend of the learning-rate curve is comparable between UTA and UBA, allowing us to directly attribute performance differences to UTA's dynamic fluctuations.

Specifically, we align the effective learning-rate scales by setting φ for UBA and φ_0 for UTA such that $\varphi \approx \varphi_0 \times 0.2$. We then train models under identical settings and compare both the macroscopic learning rate trajectories and the final performance. The results are listed in Tab. 11 and plotted in Fig. 4.

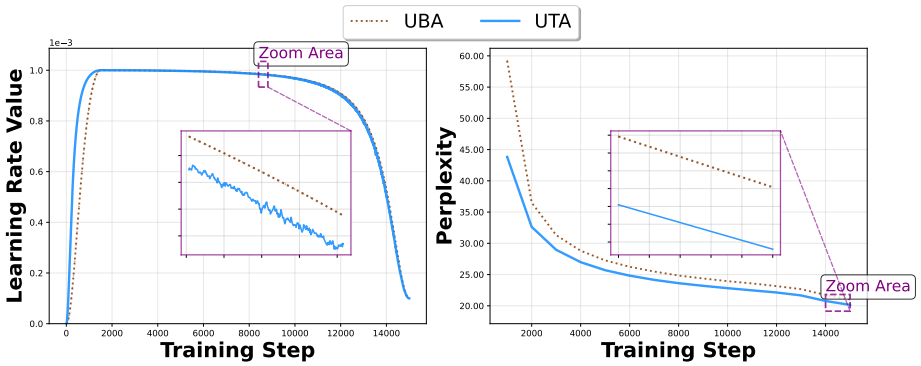
The results confirm that after parameter alignment, UTA and UBA produce LRD with nearly identical macroscopic shapes. However, UTA consistently achieves superior final accuracy. This performance gain can be directly attributed to the fine-grained, state-dependent fluctuations unique to UTA, which are absent in the static UBA curve. These fluctuations act as a form of online regularization, allowing the optimizer to navigate the loss landscape more robustly.

Table 11: Curvature alignment between UTA and UBA.

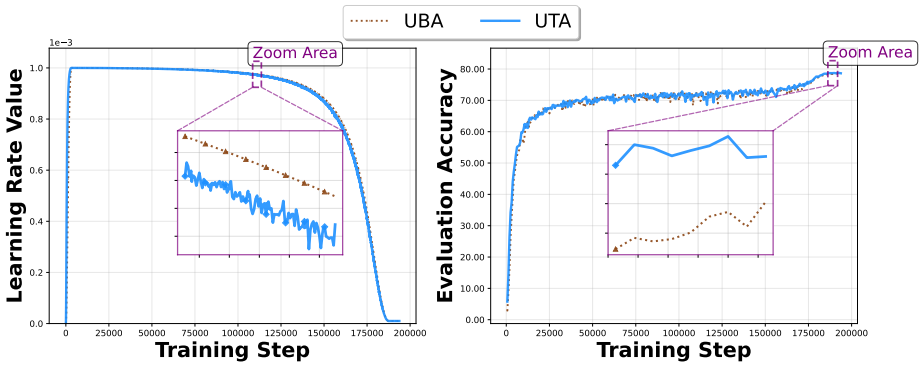
	UBA($\varphi = 0.03$)	UTA($\varphi_0 = 0.15$)	Improvement
ImageNet-ResNet50(↑)	78.47	78.64	0.17
LLaMA130M(↓)	21.08	20.15	0.93
LLaMA1B(↓)	14.81	14.53	0.28

Generalization to parameter-efficient fine-tuning To demonstrate the versatility of UTA beyond full-parameter optimization, we evaluate it in parameter-efficient fine-tuning (PEFT) scenarios. PEFT presents distinct challenges: only a small subset of parameters (e.g., LoRA adapters) are updated, creating a different optimization landscape where learning-rate scheduling may have amplified effects.

Experimental Setup. We fine-tune a pre-trained ViT-B model on two datasets: (1) **CIFAR-100** (generic object classification) and (2) **OxfordPets** (fine-grained pet breed classification). We employ **LoRA** with rank $r = 16$ applied to the query, key and value projections in attention layers, and dense



(a) LLaMA-130M.



(b) ImageNet-ResNet50.

Figure 4: Curvature alignment between UTA and UBA.

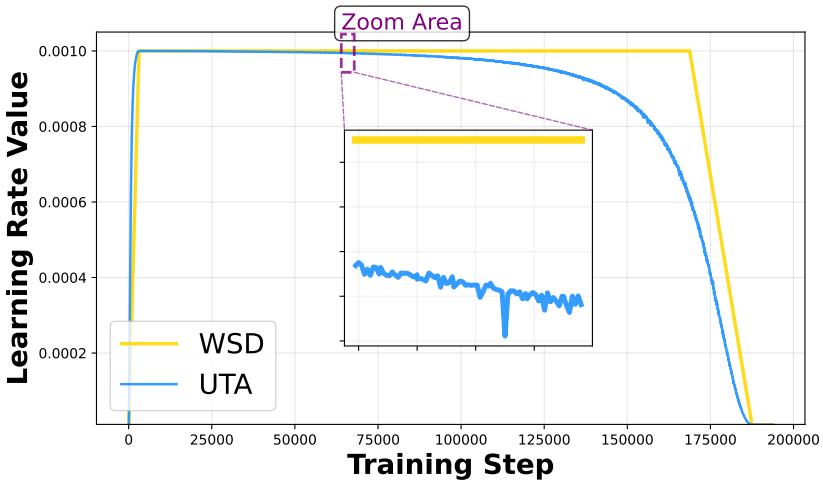


Figure 5: The learning rate curve in UTA LRD across training steps.

layer. We compare UTA against the baseline UBA LRD. All methods use the same total training budget (10 epochs), batch size, and optimizer AdamW setting such as 0.01 weight decay.

Table 12: UTA for LoRA with ViT-B.

Network	Dataset	CS	LD	WSD	UBA	UTA
ViT-B	CIFAR100	91.93	92.45	92.62	92.58	**92.81**
	OxfordPets	92.59	92.67	92.86	93.05	**93.43**

I.4 Discussion of UTA characteristics

To illustrate the characteristics of UTA in practice, we plot its learning rate trajectory during the training of ImageNet-ResNet50 in Fig. 5.

Three key observations emerge.

(i) **Overall Trajectory.** The general shape of UTA closely resembles that of the widely adopted WSD. It exhibits a prolonged stable-like phase followed by decay. Notably, however, the duration of the stable phase and and the decay rate are automatically determined by UTA, rather than being preset by hyperparameters.

(ii) Although the average learning rate of WSD is higher than that of UTA, UTA consistently achieves better performance. This confirms that the performance gains are not merely due to a higher average learning rate, but stem from UTA being near-optimal in both shape and scale, a property grounded in the preceding methodological and theoretical analysis.

(iii) **Training Adaptive Fluctuations.** Unlike most conventional LRDs that follow a static, predefined curve, UTA introduces fine-grained fluctuations in the learning rate. These adjustments are driven by real-time feedback from the training state, reflecting the optimizer's internal dynamics. This responsiveness demonstrates UTA's strong adaptive capability in practical training scenarios.

J Related work

The LRD plays a crucial role in deep learning, and prior research has investigated a wide range of approaches. Step LRD is a typical instance that reduces the learning rate by a decay factor after a predefined number of epochs (He et al., 2016). A common enhancement is the warmup phase, which gradually raises the learning rate early in training to stabilize optimization (He et al., 2016). Then Loshchilov et al. (Loshchilov & Hutter, 2017) propose the cosine LRD, which is widely regarded

as the de-facto standard choice in large-scale model training (Loshchilov & Hutter, 2017; Schaipp et al., 2025). To improve the performance under budgeted training scenarios, some LRDs include the cyclical LRD (Smith, 2017), alternative LRD under budgeted training setting (Li et al., 2019a) and REX LRD (Chen et al., 2022), yet there is no consensus on the optimal choice.

Recently, research in this area has evolved along two main lines. A line of work focus on designing LRD primarily through empirically driven approaches, and effectively enhance accuracy. Pan et al. (Pan et al., 2022a) introduce elastic step decay LRD to accelerate the pre-training for Berts. As an alternative on large model training, Warmup-Stable-Decay (WSD) LRD (Zhai et al., 2022; Hu et al., 2024) uses an extended constant learning rate phase before late-stage decay. It has become popular in LLM pre-training (Hägele et al., 2024; Hu et al., 2024) and is deployed in training industrial-scale LLMs such as DeepSeek-V3 (Liu et al., 2024) and Kimi-K2 (Team et al., 2025). Yet its mechanism remains unclear, recent works (Wen et al., 2024; Schaipp et al., 2025; Li et al., 2025) offer some insights, showing that the advantage of WSD can be explained and quantified through scaling efficiency comparisons with other LRDs.

Another line of research is more theoretically driven, seeking to establish principled frameworks and scaling laws for designing or analyzing LRD. Pan et al. (Pan et al., 2022b) propose Eigencurve, and demonstrate that it has minimax-optimal convergence on skewness Hessian quadratic objectives. Defazio et al. (Defazio et al., 2023) establish optimal linear decay LRD and refinements through rigorous mathematical derivation. Some studies (Tissue et al., 2025; Li et al., 2025; Luo et al., 2025) establish scaling law with LRD, providing theoretical explanation for empirical results observed in this area. Recently, Tang et al. propose a theoretically grounded LRD called UBA LRD (Tang et al., 2026), derives from a constructed UBA framework. This framework adopts a min-max formulation that explicitly accounts for robustness to variations in loss landscape curvature, thereby guaranteeing performance under worst-case conditions. The UBA LRD eliminates the need for heuristic design while maintaining strong and consistent performance across diverse tasks, architectures, model scales, and training budgets. Consequently, it removes the costly trial-and-error process typically required to select an appropriate LRD for a given network and training setup.