THEORETICAL ANALYSIS ON HOW LEARNING RATE WARMUP ACCELERATES CONVERGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning rate warmup is a popular and practical technique in training large-scale deep neural networks. Despite the huge success in practice, the theoretical advantages of this strategy of gradually increasing the learning rate at the beginning of the training process have not been fully understood. To resolve this gap between theory and practice, we first propose a novel family of generalized smoothness assumptions, and validate its applicability both theoretically and empirically. Under the novel smoothness assumption, we study the convergence properties of gradient descent (GD) in both deterministic and stochastic settings. It is shown that learning rate warmup consistently accelerates GD, and GD with warmup can converge at most $\Theta(T)$ times faster than with a non-increasing learning rate schedule in some specific cases, providing insights into the benefits of this strategy from an optimization theory perspective.

1 INTRODUCTION

Mathematically, training a machine learning model can be formulated as a minimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{w}),$$

where first-order optimizers using the gradient information are normally applied to find a solution to the problem. Carefully tuning the learning rates (or step sizes) is crucial in this optimization procedure, especially when the problem scale is large. A time-varying learning rate schedule is very commonly used both in theory (e.g., for Nesterov accelerated gradient method) Malitsky & Mishchenko (2020); Teboulle & Vaisbourd (2023); Boyd & Vandenberghe (2004) and in practice (e.g., cosine schedule) He et al. (2016); Vaswani et al. (2017); Loshchilov & Hutter (2017); Touvron et al. (2023).

Learning rate warmup is a strategy commonly incorporated in those schedules during the initial phase of training deep neural networks. In this stage, the learning rate, denoted as η, is set to a value lower than its target or base level. This initially small learning rate is then gradually increased over a number of training iterations until it reaches the intended peak value. A prevalent example of this is the linear warmup strategy Goyal et al. (2017), which sets a 0 initial value and increases it linearly to the target learning rate in the initial phase. The warmup strategy has been widely observed to be powerful across many practical tasks He et al. (2016); Goyal et al. (2017); Vaswani et al. (2017).

Despite the impressive practical success of learning rate warmup, a rigorous theoretical explanation for why warmup works still remains unclear. Various studies have explored or empirically validated explanations for the benefits of learning rate warmup, including limiting the magnitude of weight updates and reducing variance Gotmare et al. (2018); Liu et al. (2020); Gilmer et al. (2022); Kalra & Barkeshli (2024); Kosson et al. (2024). Among them, Gilmer et al. (2022); Kalra & Barkeshli (2024) elaborate on the intuition that the main advantage of learning rate warmup is that small initial learning rates allow the model to safely go into smoother regions of the loss landscape, characterized by smaller local smoothness (or sharpness), i.e., the largest singular value of the Hessian, in the initial phase of training. This is beneficial since the applicable learning rate scale at a specific point $\mathbf{w}$ generally needs to be bounded by $2/L(\mathbf{w})$, where $L(\mathbf{w})$ is the local smoothness Cohen et al. (2021), which implies that first going into a smoother region enables larger learning rates in the following training process, resulting in faster convergence.

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.

The connection between learning rate warmup and local smoothness inspires us to study the benefits of the warmup strategy from an optimization perspective. To mathematically model the varying local smoothness during training, we propose a novel family of smoothness assumptions that connect local smoothness with the suboptimality gap of the loss function, i.e., $f(\mathbf{w}) - f^\star$. Note that this family of assumptions is closely related to, but different from, existing generalized smoothness assumptions that link local smoothness to the gradient norm Zhang et al. (2020b); Li et al. (2023b). We show that our novel smoothness assumption is typically weaker than existing generalized smoothness assumptions, and provide examples to show its applicability for analyzing the convergence of neural networks both empirically and theoretically. The rigorous characterization of the evolution of local smoothness during the optimization process, provided by these novel assumptions, facilitates our convergence proofs. By comparing algorithms with and without a warmup phase, we find that using warmup shows a consistent gain in accelerating convergence, yielding up to a $\Theta(T)$ improvement in the convergence rate for GD.

Our main contributions are summarized as follows:

1. We propose a novel family of generalized smoothness assumptions, connecting the local smoothness with the suboptimality gap. We prove that this novel family of assumptions is strictly weaker than the existing generalized (ρ, L_0, L_ρ)-smoothness with respect to the gradient norm Zhang et al. (2020b); Li et al. (2023a) for $\rho < 2$. Experimental validation on typical deep learning models, along with several neural network examples, demonstrates the applicability of our generalized smoothness assumptions to practical deep neural network training tasks, especially in the early phase of training.
2. Based on our generalized smoothness assumptions, we theoretically prove that using a warm-up learning rate schedule can accelerate the convergence of gradient descent (GD) and stochastic gradient descent (SGD) methods, thereby bridging the gap between theory and practice in training neural networks. Specifically, it is shown that under a specific theoretically derived learning rate warmup schedule, showing a similar behavior to real practical warmup schedules, GD can achieve $\Theta(T)$ times faster convergence rates compared to directly using non-increasing learning rates. For SGD, we apply the ABC inequality Khaled & Richtárik (2023) as the noise assumption, which is general and implies further benefits of doing warmup in accelerating convergence in a noisy setting.

Due to the page limit, we defer the related work section to Appendix A.

2 A Family of Novel Generalized Smoothness Assumptions

We first review the existing smoothness assumptions. For a twice continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, the standard L-smooth assumption assumes that the spectral norm of the loss Hessian is uniformly bounded, i.e.,

$$\left\|\nabla^2 f(\mathbf{w})\right\| \leq L, \quad \forall \mathbf{w} \in \mathbb{R}^d.$$

Although the L-smoothness assumption is widely used in optimization theory, it fails to capture the local smoothness of the loss function at different points, and even some simple and common functions, such as the exponential function, do not satisfy this assumption Zhang et al. (2020b).

To generalize L-smoothness, Li et al. (2023a) proposed the (ρ, L_0, L_ρ)-smoothness:

$$\left\|\nabla^2 f(\mathbf{w})\right\| \leq L_0 + L_\rho \left\|\nabla f(\mathbf{w})\right\|^\rho, \quad \forall \mathbf{w} \in \mathbb{R}^d.$$

When $\rho = 1$, it reduces to the (L_0, L_1)-smoothness Zhang et al. (2020b). The (ρ, L_0, L_ρ)-smoothness assumes that local smoothness is bounded by an increasing polynomial function of the gradient norm and is considered to be more consistent with the deep neural networks than the L-smooth assumption based on some empirical verifications Zhang et al. (2020b). We are particularly interested in the $0 \leq \rho < 2$ case, where local smoothness is bounded by a sub-quadratic function of the gradient norm. This is because Li et al. (2023a) showed that GD may diverge for (ρ, L_0, L_ρ) functions with $\rho \geq 2$.

Although (ρ, L_0, L_ρ)-smoothness has been empirically validated as an effective assumption for characterizing the loss landscape of deep neural networks, there are still some limitations. Firstly, there exist simple examples showing that neural networks do not satisfy the (ρ, L_0, L_ρ)-smoothness with $0 \leq \rho < 2$ Patel et al. (2022). We discuss some examples in Section 2.2 in detail. This

implies that, based on the results in Li et al. (2023a), fundamental first-order optimizers like GD can diverge even under some simple examples, which is inconsistent with real practice. Moreover, for nonconvex functions, the gradient norm is not necessarily monotonically decreasing during the optimization process of GD. Since the primary role of warm-up is to suppress sharpness to allow the network to tolerate larger learning rates, the (ρ, L_0, L_ρ)-smoothness is inadequate for characterizing the decreasing trend of sharpness during the warm-up phase, regardless of whether training starts with a sharpness reduction or a progressive sharpening phase Kalra & Barkeshli (2024); Gilmer et al. (2022). The limitations raise a need for a novel family of generalized assumptions.

2.1 A Novel Family of Generalized Smoothness

We consider the following (ρ, K_0, K_ρ)-smoothness, which relates the local smoothness with the function suboptimality gap.

Definition 1. *We say a twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is (ρ, K_0, K_ρ) smooth if*

$$\left\|\nabla^2 f(\mathbf{w})\right\| \leq K_0 + K_\rho \left(f(\mathbf{w}) - f^\star\right)^\rho \tag{1}$$

for $K_0, K_\rho \geq 0$ and $\rho > 0$, where we assume $f^\star = \inf_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) > -\infty$.

Note that the lower bound f^* is standard in nonconvex analysis, which should also be satisfied by neural networks. When $K_\rho = 0$, our generalized smoothness reduces to the classical L-smoothness. It is not hard to see that the (ρ, K_0, K_ρ)-smoothness is strictly weaker than the L-smoothness since exponential functions are (ρ, K_0, K_ρ)-smooth but not L-smooth. Moreover, we can prove that the (ρ, K_0, K_ρ)-smoothness family is also weaker than the (ρ, L_0, L_ρ)-smoothness for $0 \leq \rho < 2$.

Lemma 1. *If a function $f : \mathbb{R}^d \to \mathbb{R}$ is (ρ, L_0, L_ρ)-smooth with $0 \leq \rho < 2$, then it is (α, K_0, K_α)-smooth with $\alpha = \frac{\rho}{2-\rho}$.*

Based on Lemma 1, properties of (ρ, L_0, L_ρ)-smoothness for $0 \leq \rho < 2$ as well as its applicability to deep neural networks can be inherited by (ρ, K_0, K_ρ)-smoothness. Moreover, the following simple example shows that (ρ, K_0, K_ρ)-smoothness is strictly weaker than (ρ, L_0, L_ρ)-smoothness with $0 \leq \rho < 2$ and cannot be covered by $\rho \geq 2$.

Example 1. *The following function*

$$f(x) = \begin{cases} 2x + x\sin x, & x \in [0, +\infty), \\ 2(\mathrm{e}^x - 1), & x \in (-\infty, 0). \end{cases}$$

is $(1, K_0, K_1)$-smooth but not (ρ, L_0, L_ρ)-smooth for any $\rho > 0$.

This example also illustrates that compared to (ρ, L_0, L_ρ)-smoothness, (ρ, K_0, K_ρ)-smoothness is better at capturing the properties of functions with multiple stationary points or local minima, which is a common case in deep neural network training. We also show that (ρ, K_0, K_ρ)-smoothness is closed under finite sums and affine transformation in Appendix E.4, which are common operations in neural network, whereas (ρ, L_0, L_ρ)-smoothness does not possess this property.

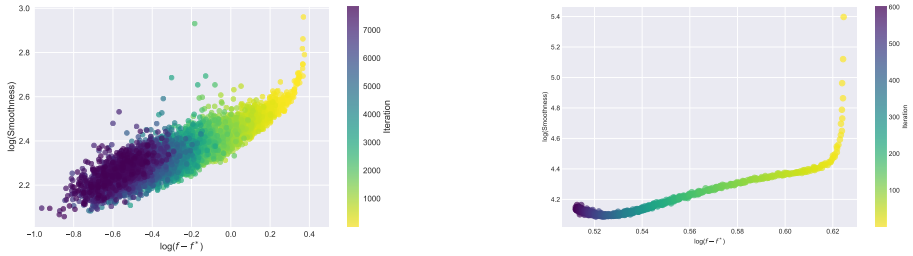
2.2 Generalized Smoothness in Neural Networks

We have shown that (ρ, K_0, K_ρ)-smoothness is a more general assumption than (ρ, L_0, L_ρ)-smoothness for $0 \leq \rho < 2$. Next, we demonstrate that (ρ, K_0, K_ρ)-smoothness is more applicable to deep neural networks. We adopt the two examples in Patel et al. (2022), where for binary classification tasks, simple feed forward network and recurrent neural network both fail to satisfy the (ρ, L_0, L_ρ)-smoothness with $0 \leq \rho < 2$, but satisfy the (ρ, K_0, K_ρ)-smoothness.

Example 2 (Example 1, Patel et al. (2022))**.** *Consider the following simple multi-layer feed forward network for binary classification:*

$$\begin{aligned} z_i &= \sigma\left(w_i z_{i-1}\right), \quad i = 1, 2, 3 \\ \hat{y} &= \varphi\left(w_4 z_3\right), \end{aligned}$$

where z_0 is the input feature, σ is the activation function and φ is the sigmoid function. Given a sample point (z_0, y), we aim to predict y. Let $f(\mathbf{w})$ be the cross entropy loss plus a ridge penalty. Then for some simple distribution, $f(\mathbf{w})$ is (ρ, K_0, K_ρ)-smooth for $\rho \geq 3$ but not (ρ, L_0, L_ρ)-smooth for any $0 \leq \rho < 2$.



(a) ResNet (b) NanoGPT

Figure 1: Local smoothness vs. function suboptimality gap on training (a) ResNet18 on CIFAR-10 (b) NanoGPT on Tiny Shakespeare character dataset. Both x and y axes are in log scale and the color bar indicates the iteration number. We use $f^* = 0$ in the plots.

Example 3 (Example 2, Patel et al. (2022))**.** *Consider the following simple recurrent neural network for binary classification:*

$$h_i = \sigma\left(w_1 h_{i-1} + w_2 z_i\right), \quad i = 0, 1, 2, 3$$
$$\hat{y} = \varphi(w_3 h_3),$$

where σ is the activation function and φ is the sigmoid function. Given a sample point (z_0, z_1, z_2, z_3, y), we sequentially observe $z_0, \ldots, z_3$ and aim to predict y. Let $f(\mathbf{w})$ be the cross entropy loss plus a a ridge penalty. Then for some simple distribution, $f(\mathbf{w})$ is (ρ, K_0, K_ρ)-smooth for $\rho \geq 3$ but not (ρ, L_0, L_ρ)-smooth for any $0 \leq \rho < 2$.

We elaborate on these two examples in detail in Appendix D. In both cases, the loss function of neural networks either do not satisfy the (ρ, L_0, L_ρ)-smoothness or only satisfy the case with $\rho \geq 2$. For the latter, GD with constant learning rates cannot guarantee convergence without additional assumptions. In contrast, in Section 3, we show that GD can converge for (ρ, K_0, K_ρ)-smooth functions for any $\rho \geq 0$, highlighting the advantage of the (ρ, K_0, K_ρ)-smoothness assumption.

2.3 Empirical Validation of the Assumption

To empirically investigate the posited relationship between local smoothness and the loss suboptimality gap within neural networks, we perform a numerical validation on Convolutional Neural Networks (CNNs) and Transformers. As direct Hessian computation is often intractable, we approximate local smoothness following Zhang et al. (2020b); Crawshaw et al. (2022). As shown in the log-log plots, a polynomial dependence of the local smoothness on the function suboptimality gap is generally clear, showing the applicability of Assumption 2. Also, as one can observe from the plots, the local smoothness can be extremely large at the beginning of the training process, which also provides evidence for the importance of using small learning rates in the initial phase of training. The experiment details are in Appendix C.1.

2.4 Properties of the Novel Generalized Smoothness

Similar to (ρ, L_0, L_ρ)-smoothness Li et al. (2023a), Definition 1 indicates that ∇f is locally Lipschitz continuous. Therefore, by careful analysis through integration, we are able to obtain the following locally Lipschitz continuous property of ∇f. We first define constants C_1, C_2 which only depend on K_0, K_ρ and ρ:

$$C_1 = \frac{1}{\left(2+\sqrt{2}\right)\sqrt{3^\rho K_\rho}}, \quad C_2 = \frac{1}{2\sqrt{3}+\sqrt{6}} \frac{K_0^{\frac{1}{2\rho}-\frac{1}{2}}}{K_\rho^{\frac{1}{2\rho}}}.$$

Lemma 2. *Suppose f is (ρ, K_0, K_ρ)-smooth. Let $\Delta = f(\mathbf{x}) - f^\star$,*

$$L(\Delta) := 2K_0 + K_\rho\left(2\Delta\right)^\rho \text{ and } r(\Delta) := \min\left\{C_1 \Delta^{-\frac{\rho-1}{2}}, C_2\right\}.$$

Then for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ satisfying $\|\mathbf{y} - \mathbf{x}\| \leq r(\Delta)$, we have

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \leq L(\Delta)\|\mathbf{y} - \mathbf{x}\|$$

and

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L(\Delta)}{2} \|\mathbf{y} - \mathbf{x}\|^2 .$$

Lemma 2 is useful for our convergence analysis. As long as the consecutive iterates $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|$ are small enough, we can obtain a descent lemma and proceed with the standard analysis techniques.

3 Theory of GD

In this section, we analyze GD for (ρ, K_0, K_ρ)-smooth functions:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t).$$

We consider two learning rate settings: constant learning rate and increasing learning rate, i.e., $\eta_t \leq \eta_{t+1}, t = 0, \ldots, T-1$. The increasing learning rate strategy can be viewed as a specific type of learning rate warmup, which will be validated empirically in Section 3.2. We show that the increasing learning rate leads to a faster convergence rate compared to the constant learning rate. We first list the assumptions we require for convergence analysis.

Assumption 1. *We assume $f(\mathbf{w}_0) - f^\star < \infty$, where $f^\star = \inf_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w})$.*

Assumption 2. *$f(\mathbf{w})$ is (ρ, K_0, K_ρ)-smooth.*

We use $\Delta_t \triangleq f(\mathbf{w}_t) - f^\star$ for simplicity throughout the paper.

3.1 Upper Bounds

We first present the results under the general nonconvex scheme.

Theorem 1. *Suppose Assumptions 1 and 2 hold. $\{\mathbf{w}_t\}$ is generated by GD. Let the learning rate $\eta_t = \frac{1}{4\sqrt{2}+4} \min\left\{\frac{1}{K_0}, \frac{1}{3^\rho K_\rho} \Delta_t^{-\rho}\right\}$. Then it holds that $\Delta_t \geq \Delta_{t+1}$ for all $t \in [T]$, and*

$$\min_{t<T} \|\nabla f(\mathbf{w}_t)\|^2 \leq \frac{2\left(f(\mathbf{w}_0) - f^\star\right)}{\sum_{t=0}^{T-1} \eta_t} = \mathcal{O}\left(\frac{K_0 \Delta_0}{T} + \frac{K_\rho \Delta_0 \sum_{t=0}^{T-1} \Delta_t^\rho}{T^2}\right).$$

Moreover, if we use a constant learning rate $\eta = \frac{1}{4\sqrt{2}+4} \min\left\{\frac{1}{K_0}, \frac{1}{3^\rho K_\rho} \Delta_0^{-\rho}\right\}$, then we have $\Delta_t \geq \Delta_{t+1}$ for all $t \in [T]$, and

$$\min_{t<T} \|\nabla f(\mathbf{w}_t)\|^2 \leq \frac{2\left(f(\mathbf{w}_0) - f^\star\right)}{\eta T} = \mathcal{O}\left(\frac{K_0 \Delta_0 + K_\rho \Delta_0^{\rho+1}}{T}\right).$$

We can conduct a simple comparison between the two results in Theorem 1. Since the function gap Δ_t is monotonically decreasing during the optimization process, the learning rate schedule $\{\eta_t\}$ is monotonically increasing, and thus can be viewed as a specific adaptive strategy of learning rate warmup. By $\sum_{t=0}^{T-1} \Delta_t^\rho \leq T \Delta_0^\rho$, we know that the convergence rate with learning rate warmup is better than that with a constant learning rate, showing an acceleration effect when K_ρ is significant, i.e., the local smoothness is varying and highly dependent on the suboptimality. This can likely happen since K_0 can be quite small, as it doesn't need to globally bound the Hessian norm as in the case of L-smoothness. Moreover, to provide more insights into how significant this gap can be, we further analyze the convex convergence of GD as presented in Theorem 2.

Theorem 2. *Suppose Assumptions 1 and 2 hold. Further assume that f is convex and level-bounded. Define $\mathbf{w}^\star = \arg\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w})$ and $D_0 = \|\mathbf{w}_0 - \mathbf{w}^\star\|$. If we use the learning rate schedule $\eta_t = \frac{1}{8\sqrt{2}+8} \min\left\{\frac{1}{K_0}, \frac{1}{3^\rho K_\rho} \Delta_t^{-\rho}\right\}$, then we have $\Delta_{t+1} \leq \Delta_t$,*

$$f(\mathbf{w}_{T-1}) - f(\mathbf{w}^\star) \leq \mathcal{O}\left(\frac{D_0^2 K_0}{T} + \left(\frac{D_0^2 K_\rho}{T}\right)^{\frac{1}{1-\rho}}\right)$$

for $\rho < 1$ and

$$f(\mathbf{w}_{T-1}) - f(\mathbf{w}^\star) \leq \mathcal{O}\left(\frac{2 D_0^2 K_0}{T}\right), T \geq \mathcal{O}\left(K_\rho D_0^2 \Delta_0^{\rho-1}\right)$$

for $\rho \geq 1$. Moreover, if we use the constant learning rate $\eta_t = \eta = \frac{1}{8\sqrt{2}+8} \min\left\{\frac{1}{K_0}, \frac{1}{3^\rho K_\rho}\Delta_0^{-\rho}\right\}$, then we have $\Delta_{t+1} \leq \Delta_t$, and

$$f(\mathbf{w}_{T-1}) - f(\mathbf{w}^\star) \leq \mathcal{O}\left(\frac{D_0^2 K_0}{T} + \frac{D_0^2 K_\rho \Delta_0^\rho}{T}\right).$$

From Theorem 2, to obtain an ϵ-optimal solution $f(\mathbf{w}) - f(\mathbf{w}^\star) \leq \epsilon$, the required iteration number is $\mathcal{O}\left(\frac{K_0 D_0^2}{\epsilon} + \frac{K_\rho D_0^2 \Delta_0^{\max\{0,\rho-1\}}}{\epsilon^{\max\{0,1-\rho\}}}\right)$ for the warm-up schedule and $\mathcal{O}\left(\frac{K_0 D_0^2}{\epsilon} + \frac{K_\rho D_0^2 \Delta_0^\rho}{\epsilon}\right)$ for the constant learning rate. Therefore, we can clearly see that the upper bound of convergence rate of GD with the specific warm-up schedule is strictly better than that of GD with a constant learning rate if $K_\rho > 0$. Specifically, if $\rho \geq 1$, the convergence rate of GD with the warmup learning rate schedule is $\mathcal{O}\left(\frac{K_0 D_0^2}{\epsilon} + K_\rho D_0^2 \Delta_0^{\rho-1}\right)$, an acceleration of $\Theta(\Delta_0 T)$ compared to a constant learning rate.

Remark 1. *We remark that the primary objective of our work is to provide a theoretical explanation for the benefits of learning rate warm-up from an optimization perspective, rather than proposing a warm-up strategy that outperforms linear warm-up. Although we choose a specific warm-up schedule in Theorems 1 and 2 for the theoretical analysis, it is worth noting that any increasing learning rate schedule satisfying some conditions like (11) in the proofs can improve the convergence rate over GD with constant learning rate.*

We also come across the following simple but intuitive example to illustrate that the difference in K_ρ terms can be significant.

Example 4. *Consider a specific 2-dimensional function $f(x, y) = h(x) + g(y)$, with*

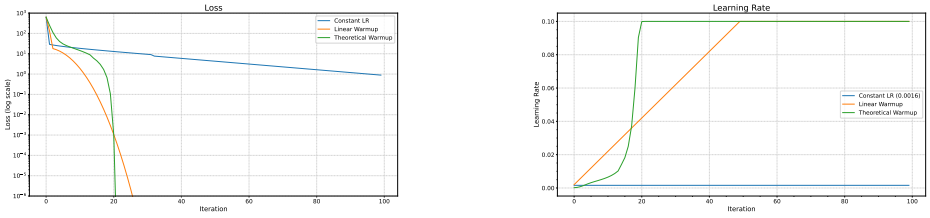
$$h(x) = \begin{cases} \mathrm{e}^{-\sqrt{K_1}x - 1} - \frac{1}{2}, & x \in (-\infty, -\frac{1}{\sqrt{K_1}}) \\ \frac{1}{2}K_1 x^2, & x \in [-\frac{1}{\sqrt{K_1}}, \frac{1}{\sqrt{K_1}}] \\ \mathrm{e}^{\sqrt{K_1}x - 1} - \frac{1}{2}, & x \in (\frac{1}{\sqrt{K_1}}, +\infty) \end{cases}, \quad g(y) = \frac{1}{2}K_1 y^2.$$

Note that f is $(1, K_1, K_1)$-smooth and $f^* = 0$. This function is a simple construction that approximates the river-valley loss landscape presented in Wen et al. (2024), which is believed to capture the properties of neural networks. The landscape is very sharp along the x-axis and mild along the y-axis, where the y-axis can be interpreted as the river. Consider the initialization $x_0 = \frac{\log(\Delta_0)}{\sqrt{K_1}}$ with $\Delta_0 > \mathrm{e}$ and $y_0 > 1$. Then GD with warmup can converge $\tilde{\Theta}(\Delta_0)$ times faster than using constant learning rates. A detailed explanation for this can be found in Appendix F.4. This gap also highlights the importance of warmup when the training doesn't have a good initialization (Δ_0 is large).

Therefore, we can provide a theoretical explanation for the empirical advantages of learning rate warmup Kalra & Barkeshli (2024); Gilmer et al. (2022) based on the theorems. At the beginning of training, the initialization may be poor, leading to a large function gap Δ_0 and high local smoothness $\left\|\nabla^2 f(\mathbf{w}_0)\right\| \leq K_0 + K_\rho \Delta_0^\rho$. Thus, the initial learning rate of a constant (or non-increasing) learning rate schedule must be sufficiently small ($\mathcal{O}(K_\rho^{-1}\Delta_0^{-\rho})$), to prevent oscillation or divergence. In practice, depending on the model's initialization and parameterization, training may start with either a sharpness reduction phase or a progressive sharpening phase Cohen et al. (2021); Kalra & Barkeshli (2024). The former scenario is congruent with our smoothness assumption. In the latter case, once sharpness exceeds the stability threshold, the training enters the edge of stability (EoS) regime Cohen et al. (2021), where warm-up facilitates training by actively reducing sharpness. Therefore, as training progresses, the function gap Δ_t decreases, and a reduction in local smoothness happens, which in turn allows a larger learning rate. Learning rate warmup accelerates this process, getting the model into regions of the loss landscape with lower local smoothness more quickly. Moreover, it enables the use of larger learning rates after entering the smooth regions, which is often denoted as the target learning rate in the warm-up strategy. Our theory shows that this acceleration can be significant, at least in the early phase of training. In Appendix B, we further discuss the tightness of our smoothness assumption from the perspective of the EoS literature.

3.2 An Empirical Validation of the Theoretical Schedule

We do an empirical validation of the simple synthetic river-valley minimization problem described in Example 4 to empirically support our claim that the theoretically derived schedule $\{\eta_t\}$ in Theorem 1



(a) loss convergence (b) lr schedule in the training

Figure 2: An empirical experiment based on the synthetic problem setting in Example 4. The loss convergence curves are on the left side, and the learning rate dynamics are on the right side.

is a valid warmup schedule rather than to show that it performs exceptionally well in practice. In the specific experimental setup, we set $\Delta_0 = 1000$ and $K_1 = 10$, $x_0 = \frac{\log(\Delta_0)}{\sqrt{K_1}}$ and $y_0 = 2$. We consider 3 schedules for comparison: constant, linear warmup, and the theoretical warmup in Theorem 1. For a fair comparison, we carefully tune the learning rate scale for all schedules, i.e., we find an optimal constant to multiply the learning rate to achieve the fastest convergence (without leading to divergence). The results are shown in Figure 2, where we can observe that the theoretical warmup schedule and linear warmup schedule achieve a similar significant acceleration in loss convergence compared to the constant schedule. The learning rate dynamics figure also shows that both warmup schedules enable a much larger stable learning rate compared to the constant schedule without warmup. For an empirical validation on ResNet, please refer to Appendix C.2.

3.3 Lower Bound of GD

We consider the lower bound of GD with non-increasing learning rate schedules under $\rho = 1$.

Theorem 3. *Given $K_1 > 0$, $\Delta > 0$ as the initial loss gap parameter, and a desired accuracy $0 < \epsilon \leq \frac{2\sqrt{K_1}\Delta}{5}$, for GD with any non-increasing learning rate sequence $\{\eta_t\}$, there exists a function $f : \mathbb{R} \to \mathbb{R}$ that is $(1, 3\epsilon\sqrt{K_1}, 4\pi K_1)$-smooth, lower bounded by f^* and $f(\mathbf{w}_0) - f^\star \leq 8\Delta$, such that GD needs at least*

$$\Omega\left(\frac{K_1\Delta^2}{\epsilon^2}\right)$$

iterations to achieve $\|\nabla f(\mathbf{w})\| \leq \epsilon$.

The lower bound matches the upper bound result for constant learning rates in Theorem 1, implying the tightness of the bounds. The proof is presented in Appendix F.3. The specific lower bound construction is based on the use of trigonometric functions, which satisfy the $(1, K_0, K_1)$-smoothness but cannot be adopted by the (ρ, L_0, L_ρ)-smoothness assumptions for any $\rho > 0$ as noted in Example 1.

Note that compared to the lower bound for $(1, L_0, L_1)$-smoothness presented in Zhang et al. (2020b), Theorem 3 is more general since it allows general non-increasing learning rate schedules rather than only constant learning rates. This requires novel construction and proof techniques and may be of interest for future study on lower bounds. Also, Theorem 3 does not have additional logarithmic terms in the lower bound, which serves as evidence for the fact that $(1, K_0, K_1)$-smoothness is strictly weaker than $(1, L_0, L_1)$-smoothness and more difficult to optimize.

4 Theory of SGD

In this section, we analyze SGD for (ρ, K_0, K_ρ)-smooth functions:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t,$$

where $\mathbf{g}_t = \nabla f(\mathbf{w}_t, \xi_t)$ and ξ_t are i.i.d. samples from the underlying distribution. We first make the following assumptions on the noise.

Assumption 3. *$\mathbb{E}_\xi \nabla f(\mathbf{w}, \xi) = \nabla f(\mathbf{w})$ and $\|\nabla f(\mathbf{w}, \xi) - \nabla f(\mathbf{w})\| \leq \sigma$ for some $\sigma > 0$ and all $\mathbf{w} \in \mathbb{R}^d$, with probability 1.*

Assumption 4. $\mathbb{E}_\xi \nabla f(\mathbf{w}, \xi) = \nabla f(\mathbf{w})$ *and* $\|\nabla f(\mathbf{w}, \xi) - \nabla f(\mathbf{w})\|^2 \leq A\left(f(\mathbf{w}) - f^\star\right) + B\|\nabla f(\mathbf{w})\|^2 + \sigma^2$ *for some* $A \geq 0, B \geq 0, \sigma > 0$ *and all* $\mathbf{w} \in \mathbb{R}^d$*, with probability* 1.

Assumption 3 is commonly used in stochastic optimization, especially under generalized smoothness settings Zhang et al. (2020a;b); Li et al. (2023b); Crawshaw et al. (2022), for example, the (L_0, L_1)-smoothness.

The expectation version of Assumption 4 was originally introduced by Khaled & Richtárik (2023) in the expectation form, and is a more general noise assumption compared with the bounded variance assumption, covering a wider range of randomness sources. We adopt this assumption to consider more general settings, and also explore how learning rate warmup can help with convergence when the noise is significant. Also note that the assumptions we considered here can be relaxed to the sub-gaussian noise assumptions, with all the theorems still valid up to some logarithmic terms, and we include the corresponding details in Appendix I.

4.1 Bounded Noise

Let $\Delta_{\text{avg},p} := T^{-1}\sum_{t=0}^{T-1}\Delta_t^p$ and $\ell_\delta := \log(\mathrm{e}/\delta)$. For compactness, write

$$H_{\text{ad}} := K_0 + K_\rho \Delta_{\text{avg},\rho}, \quad R_{\text{ad}} := C_1^{-1}\Delta_{\text{avg},(\rho-1)/2} + C_2^{-1},$$
$$H_{\text{c}} := K_0 + K_\rho \Delta_0^\rho, \quad R_{\text{c}} := C_1^{-1}\Delta_0^{(\rho-1)/2} + C_2^{-1}$$

We will use $\eta_t = h(\Delta_t)$ for warm-up schedule and $\eta_t = h(D), D > 0$ for constant lr, where $h(x)$ is a decreasing function defined below:

$$\eta_t = h(\Delta_t) = \min\left\{\frac{r(\Delta_t)}{G(\Delta_t)+\sigma}, \frac{1}{2L(\Delta_t)}, \sqrt{\frac{\Delta_0}{TL(\Delta_t)\sigma^2}}, \frac{\Delta_0}{\sigma G(\Delta_t)\sqrt{2T\ell_\delta}}\right\},$$

where $r(u) = \min\{C_1 u^{-(\rho-1)/2}, C_2\}, L(u) = 2K_0 + K_\rho(2u)^\rho, G(u) = 2\sqrt{u\left(K_0 + 3^\rho K_\rho u^\rho\right)}$,

We use $\mathcal{O}_\rho$ to hide only numerical factors depending on ρ.

Theorem 4. *Suppose Assumptions 1, 2 and 3 hold with* $\rho \geq 1$*, and let SGD use the learning rate schedule* $h(\Delta_t)$ *defined above. Then with probability at least* $1-\delta$*,* $\Delta_t \leq 3\Delta_0$ *for all* $0 \leq t \leq T$*, and*

$$\min_{t<T}\|\nabla f(\mathbf{w}_t)\|^2 \leq \mathcal{O}_\rho\left(\frac{\Delta_0}{T}(H_{\text{ad}} + \sigma R_{\text{ad}}) + \sigma\sqrt{\frac{\Delta_0 \ell_\delta H_{\text{ad}}}{T}}\right). \tag{2}$$

Theorem 5. *Under the same assumptions as Theorem 4, let SGD use the constant learning rate* $\eta = h(D), D = 3\Delta_0$ *defined above. Then with probability at least* $1-\delta$*,* $\Delta_t \leq 3\Delta_0$ *for all* $0 \leq t \leq T$*, and*

$$\min_{t<T}\|\nabla f(\mathbf{w}_t)\|^2 \leq \mathcal{O}_\rho\left(\frac{\Delta_0}{T}(H_{\text{c}} + \sigma R_{\text{c}}) + \sigma\sqrt{\frac{\Delta_0 \ell_\delta H_{\text{c}}}{T}}\right). \tag{3}$$

We can see that in the stochastic settings, both constant learning rates and the learning rate schedule adapted to Δ_t achieve the $\mathcal{O}(1/\sqrt{T})$ convergence rate with high probability, which matches the optimal rate Arjevani et al. (2023). However, it is not hard to see that to ensure convergence, we take the adaptive learning rate $\eta_t = h(\Delta_t)$ and the constant learning rate $\eta = h(3\Delta_0)$ in the same pattern, where h is a monotonically decreasing function. Therefore, we have $\eta_t \geq \eta$ and $\Delta_{avg,\rho} \leq (3\Delta_0)^\rho$, and the terms introduced by $H_{\text{ad}}, R_{\text{ad}}$ and $H_{\text{c}}, R_{\text{c}}$ in the convergence rates are also correspondingly related to $\Delta_{avg,\rho}$ and $(3\Delta_0)^\rho$, indicating that the convergence rate in Theorem 4 is better than that in Theorem 5. We also note that $\Delta_{avg,\rho}$ can be significantly smaller than $(3\Delta_0)^\rho$, which is reflected in the following convex example.

Example 5. *Suppose further that* f *is convex level-set bounded. Let* $D := 3\Delta_0$*. Similar to the analysis of Theorems 4 and 5 and that in Liu & Zhou (2023), for* $\rho \geq 1$*, we have that* $\sum_{t=0}^{T-1}\Delta_t/T = \mathcal{O}(T^{-1/2})$ *and* $\Delta_t \leq D$ *holds with high probability. Therefore,* $\Delta_{\text{avg},\rho} = \frac{1}{T}\sum_{t=0}^{T-1}\Delta_t^\rho \leq \mathcal{O}(T^{-1/2})$*. Consequently, the smoothness in the warm-up bound satisfies* $H_{\text{ad}} = K_0 + K_\rho\Delta_{\text{avg},\rho} \leq K_0 + \mathcal{O}(T^{-1/2})$*, whereas the corresponding quantity for the*

constant learning rate remains $H_c = K_0 + K_\rho \Delta_0^\rho$. Thus, the K_ρ-dependent contribution in the warm-up bound decreases along the optimization trajectory, while that in the constant-learning-rate bound remains determined by the initial function gap. A similar improvement holds for the radius-dependent quantity R_{ad}. Therefore, the warm-up schedule admits a strictly sharper trajectory-dependent bound: both H_{ad} and R_{ad} decrease as the function gap decreases, while H_c and R_c remain controlled by the initial gap. This illustrates the advantage of adapting the learning rate to the local geometry along the optimization trajectory.

In the stochastic setting, we cannot guarantee that Δ_t decreases at every step. Nevertheless, in practice, as long as SGD does not diverge, Δ_t typically shows a decreasing trend. Hence, η_t presented in Theorem 4 is approximately increasing and can be interpreted as a specific adaptive learning rate warmup strategy, which is also verified in Section 3.2, though we do not aim to introduce a warm-up strategy that is superior to existing methods. Moreover, as training progresses, local smoothness also decreases with the decrease of Δ_t, allowing the model to reach flatter regions of the loss landscape and use larger learning rates afterwards, just like the deterministic case.

4.2 ABC-Inequality

We use the same notations as in Section 4.1. We will use $\eta_t = h_{A,B}(\Delta_t)$ for warm-up schedule and $\eta_t = h_{A,B}(D)$ for constant schedule, where $h_{A,B}(x)$ is a decreasing function similar to $h(t)$ defined in Section 4.1. The explicit warm-up and constant learning rates are stated in (40) and (41) in Appendix H.

Theorem 6. *Suppose Assumptions 1, 2 and 4 hold with $\rho \geq 1$, and let SGD use the learning rate schedule $\eta_t = h_{A,B}(\Delta_t)$ defined in (40). Then with probability at least $1-\delta$, $\Delta_t \leq 3\Delta_0$ for all $t \in [T]$, and*

$$\min_{t<T} \|\nabla f(\mathbf{w}_t)\|^2 \leq \mathcal{O}_\rho \left[\frac{\Delta_0}{T}(H_{ad} + \sigma R_{ad}) + \sqrt{\frac{\Delta_0 \ell_\delta}{T}\left((\sigma^2 + A\Delta_0)H_{ad} + B\Delta_0 H_{ad}^2\right)} \right]. \quad (4)$$

Theorem 7. *Under the same assumptions as Theorem 6, let SGD use the constant learning rate $\eta_t = h_{A,B}(D), D = 3\Delta_0$ defined in (41). Then with probability at least $1-\delta$, $\Delta_t \leq 3\Delta_0$ for all $t \in [T]$, and*

$$\min_{t<T} \|\nabla f(\mathbf{w}_t)\|^2 \leq \mathcal{O}_\rho \left[\frac{\Delta_0}{T}(H_c + \sigma R_c) + \sqrt{\frac{\Delta_0 \ell_\delta}{T}\left((\sigma^2 + A\Delta_0)H_c + B\Delta_0 H_c^2\right)} \right]. \quad (5)$$

Note that by setting $A = B = 0$, Theorem 6 and Theorem 7 cover the results of Theorem 4 and Theorem 5, respectively. Similar to the discussion in Section 4.1, since Δ_t should be generally decreasing in the training process, η_t is approximately increasing and can be regarded as a specific learning rate warmup strategy.

Moreover, the extra gradient noise in Assumption 4 introduces even more benefits of warmup. As we noted in the previous examples, $\Delta_{avg,\rho}$ can be significantly smaller than Δ_0^ρ. Thus, by comparing the results in Theorem 6 and 7, we notice that the specific warmup schedule can reduce the dependence of convergence rates on both A and B, which further demonstrates that learning rate warmup may also be beneficial when the gradient noise is significant and related to the landscape.

5 Conclusion

The paper investigates a theoretical explanation for the benefits of the learning rate warmup strategy. We propose a novel family of generalized smoothness assumptions to better describe the local smoothness variation in the training process. Then, under the novel smoothness assumptions, we prove that GD and SGD can both benefit from the warmup strategy, showing potentially a $\Theta(T)$ times acceleration for the deterministic setting and $\Theta(\sqrt{T})$ times for the stochastic settings in convergence speed over a constant or non-increasing learning rate schedule. A limitation of this work is that the analysis only applies to SGD, but not to SGD with momentum or Adam. However, we believe that our analysis can be extended to these optimizers, since the described benefits of warm-up in this paper arise mainly from our generalized smoothness assumptions, which are independent of the choice of an optimizer. Also, the lower bound results currently only apply to the $(1, K_0, K_1)$-smoothness setting, which may not be general enough. These are potentially interesting future topics.

AI USE STATEMENT

In this work, we used generative AI tools to assist in checking and revising parts of the mathematical arguments and proofs, including identifying gaps in existing arguments and suggesting corrective proof steps. We also used it to improve the clarity and presentation of the manuscript. We did not use generative AI tools to develop the overall theoretical framework, propose research hypotheses, design research methodology or experiments, implement methods, generate or preprocess datasets, translate the manuscript, or interpret experimental results. We manually checked the AI-assisted mathematical revisions against the stated assumptions and verified that the revised arguments support the corresponding claims. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

This paper mainly focuses on the theoretical part, and we include all the theorem statements and assumptions in the paper, and proofs in the appendix.

REFERENCES

Foivos Alimisis, Rustem Islamov, and Aurélien Lucchi. Why Do We Need Warm-up? A Theoretical Perspective. *arXiv.org*, abs/2510.03164, 2025.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2): 165–214, 2023.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in neural information processing systems*, 35:9955–9968, 2022.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in neural information processing systems*, 31, 2018.

Justin Gilmer, Behrooz Ghorbani, Ankush Garg, Sneha Kudugunta, Behnam Neyshabur, David Cardoze, George Edward Dahl, Zachary Nado, and Orhan Firat. A loss curvature perspective on training instabilities of deep learning models. In *International Conference on Learning Representations*, 2022.

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*, 2018.

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r14EOsCqKX.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

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.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257 – 317, 2020.

Stanisław Jastrzębski, Zac Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Amos Storkey, and Yoshua Bengio. Three factors influencing minima in SGD, 2018. URL https://openreview.net/forum?id=rJma2bZCW.

Dayal Singh Kalra and Maissam Barkeshli. Why warmup the learning rate? underlying mechanisms and improvements. *Advances in Neural Information Processing Systems*, 37:111760–111801, 2024.

Andrej Karpathy. nanoGPT. https://github.com/karpathy/nanoGPT, 2022.

Ahmed Khaled and Peter Richtárik. Better Theory for SGD in the Nonconvex World. *Transactions on Machine Learning Research (TMLR)*, 2023, 2023.

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. Analyzing & reducing the need for learning rate warmup in gpt training. *Advances in Neural Information Processing Systems*, 37:2914–2942, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=8aunGrXdkl.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of Adam Under Relaxed Assumptions. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023b.

Chaoyue Liu, Dmitriy Drusvyatskiy, Misha Belkin, Damek Davis, and Yian Ma. Aiming towards the minimizers: fast convergence of sgd for overparametrized problems. *Advances in neural information processing systems*, 36:60748–60767, 2023.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the Variance of the Adaptive Learning Rate and Beyond. In *International Conference on Learning Representations (ICLR)*, 2020.

Yuxing Liu, Rui Pan, and Tong Zhang. Adagrad under anisotropic smoothness. *arXiv preprint arXiv:2406.15244*, 2024.

Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. *arXiv preprint arXiv:2312.08531*, 2023.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Yura Malitsky and Konstantin Mishchenko. Adaptive Gradient Descent without Descent. In *International Conference on Machine Learning (ICML)*, pp. 6702–6712, 2020.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Vivak Patel, Shushu Zhang, and Bowen Tian. Global Convergence and Stability of Stochastic Gradient Descent. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Jiang Qian, Yuren Wu, Bojin Zhuang, Shaojun Wang, and Jing Xiao. Understanding Gradient Clipping In Incremental Gradient Methods. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1504–1512, 2021.

Artem Riabinin, A. Veprikov, Arman Bolatov, Martin Takác, and A. Beznosikov. Where Does Warm-Up Come From? Adaptive Scheduling for Norm-Constrained Optimizers. *arXiv.org*, abs/2602.05813, 2026.

Samuel L. Smith, Pieter-Jan Kindermans, and Quoc V. Le. Don't decay the learning rate, increase the batch size. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1Yy1BxCZ.

Marc Teboulle and Yakov Vaisbourd. An elementary approach to tight worst case complexity analysis of gradient based methods. *Mathematical Programming*, 201(1):63–96, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Alexander Tyurin. Toward a Unified Theory of Gradient Descent under Generalized Smoothness. In *Forty-second International Conference on Machine Learning*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020a.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity. In *International Conference on Learning Representations (ICLR)*, 2020b.

Shen-Yi Zhao, Yin-Peng Xie, and Wu-Jun Li. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64:1–13, 2021.

A Related Work

Learning rate warmup. Learning rate warmup is a widely employed heuristic for training deep neural networks. The use of learning rate warmup dates back at least to He et al. (2016), which used a small constant learning rate during the first stage of training. Later, the linear warmup strategy was introduced by Goyal et al. (2017), and soon became popular for training a large range of models, including ResNets He et al. (2016) and transformers Vaswani et al. (2017). Empirical evidence showed that learning rate warmup can enhance training stability to allow large learning rates and improve model performance Gotmare et al. (2019); Gilmer et al. (2022); Kalra & Barkeshli (2024).

Intuitions for the benefits of warmup. In Goyal et al. (2017), the authors proposed that to use a larger batch size, the learning rate should be scaled up proportionally. Smith et al. (2018); Jastrzębski et al. (2018) theoretically studied how the ratio between batch size and learning rate affects the training dynamics of SGD. However, in many cases, the learning rate cannot directly increase proportionally to the batch size in order to maintain training stability. Thus, warmup was introduced by Goyal et al. (2017) as a trick for gradually increasing learning rates. After that, studies on the warmup mechanism appeared. Gotmare et al. (2018) found that warmup prevents training instability by limiting the updates to deep-layer weights through empirical analysis. Liu et al. (2020) specifically studied Adam Kingma & Ba (2014) and attributed training instability to the large variance caused by the adaptive step size of Adam and viewed warmup as a method of variance reduction. Other work suggested that learning rate warmup enables the model to enter smoother regions of the loss landscape, leading to a gradual decrease in local smoothness (sharpness) Gilmer et al. (2022); Kalra & Barkeshli (2024). Based on the relation between learning rates and the local smoothness Nesterov

et al. (2018); Cohen et al. (2021), this enables larger learning rates in the following training process, thereby accelerating the convergence. Wen et al. (2024) also provided a similar understanding by proposing an intuitive river-valley interpretation of the neural networks' landscape.

Generalized Smoothness. The smoothness condition plays a significant role in optimization theory. For a twice-differentiable function, the standard L-smooth assumption assumes an upper bound L on the largest singular value of the Hessian Nesterov et al. (2018), where L is a constant. Zhang et al. (2020b) was probably the first to generalize the upper bound L to be a linear function of the current gradient norm, i.e., $\left\|\nabla^2 f(\mathbf{w})\right\| \leq L(\mathbf{w}) = L_0 + L_1 \left\|\nabla f(\mathbf{w})\right\|$, which is strictly weaker than the standard smoothness condition and is verified to be valid in some small neural networks. The idea was followed by Zhang et al. (2020a), which derived finer properties of the generalized smoothness. Further extensions of this generalized smoothness have also been developed since then. Li et al. (2023a) extended the linear function of $\|\nabla f(\mathbf{w})\|$ to $\|\nabla f(\mathbf{w})\|^\rho$ with $\rho \geq 1$ and proved that GD with a constant learning rate converges if and only if $\rho < 2$. In another direction, Crawshaw et al. (2022); Liu et al. (2024) developed anisotropic versions of the generalized smoothness assumption.

Convergence under Generalized Smoothness. The convergence of SGD under the L-smoothness assumption has been extensively studied. For the (L_0, L_1)-smoothness, most analyses focused on varying learning rates, such as SGD with clipping Zhang et al. (2020b;a); Qian et al. (2021), SignSGD Crawshaw et al. (2022), and normalized SGD Zhao et al. (2021). Moreover, their analyses often rely on the bounded noise assumption or the subgaussian noise assumption. Li et al. (2023a) proved the convergence of SGD with constant learning rate under (ρ, L_0, L_ρ)-smoothness with $0 \leq \rho < 2$, by bounding the gradients along the optimization trajectory. Their constant learning rate depends on the initial suboptimality gap, which in turn depends on both the loss function and initialization. Tyurin (2025) proposes a specific adaptive learning rate, under which GD converges for (ρ, L_0, L_ρ)-smooth functions for any $\rho > 0$. However, the proposed learning rate involves computing an integral, which typically does not have a closed-form expression, and is also not necessarily monotonic, making it less practical and different from our settings. Note that the lower bound of SGD under the L-smoothness and bounded variance conditions is $\Omega(1/T^{1/4})$ Arjevani et al. (2023). The above analyses, under generalized smoothness conditions, also achieve the $O(1/T^{1/4})$ bound, though some of them rely on stronger noise assumptions.

Concurrent Works. Subsequent to the completion of our manuscript, two recent studies that employ smoothness assumptions similar to our (ρ, K_0, K_ρ)-smoothness assumption to explain the benefits of learning rate warm-up also became public Alimisis et al. (2025); Riabinin et al. (2026). We provide a detailed discussion of these works here.

Alimisis et al. (2025) proposed the (H_0, H_1)-smoothness assumption: $f(\mathbf{w}) - f^\star \leq H_0 + H_1(f(\mathbf{w}) - f^\star)$, to investigate the convergence of GD. In comparison, our work offers several key distinctions. First, the (H_0, H_1)-smoothness is a specific instance of our (ρ, K_0, K_ρ)-smoothness framework where $\rho = 1$; thus, our smoothness assumption is more general. Second, while both studies provide empirical validation that the smoothness assumptions approximately hold in neural network training, our theoretical orientations differ. Our work focuses on demonstrating that the (ρ, K_0, K_ρ)-smoothness is better suited than the (ρ, L_0, L_ρ)-smoothness for the characterization of the loss landscape. In contrast, Alimisis et al. (2025) emphasizes proving that the (H_0, H_1)-smoothness holds across a wide range of deep learning architectures, including deep linear networks under balanced condition Du et al. (2018), two-layer neural networks with l_2 regularization, and simplified one-layer Transformers. Third, we theoretically establish the benefits of GD with warm-up schedule in both general convex and nonconvex settings, with a primary focus on the latter. Alimisis et al. (2025) proves the benefits of warm-up (specifically for $\rho = 1$) in structured nonconvex setting: under the PL Polyak (1963) and aiming conditions Liu et al. (2023). Finally, both works establish lower bounds for GD within their respective settings. While Alimisis et al. (2025) establish a lower bound complexity for GD with a constant learning rate in general non-convex cases, we provide a more general lower bound applicable to any non-increasing learning rate. Notably, our lower bound is tighter than that of Alimisis et al. (2025) as it eliminates the logarithmic factor in Alimisis et al. (2025).

Riabinin et al. (2026) introduced the (ρ, K_0, K_1, K_ρ)-smoothness assumption, defined by $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \mathcal{K}(\mathbf{x})\|\mathbf{x} - \mathbf{y}\|$ with $\mathcal{K}(\mathbf{x}) := K_0 + K_1(f(\mathbf{x}) - f^\star) + K_\rho(f(\mathbf{x}) - f^\star)^\rho$. Distinct from

our work, Riabinin et al. (2026) employed the learning rate schedule $\eta_t = \mathcal{O}\left(\frac{\Delta_t}{K_0+K_1\Delta_t+K_\rho\Delta_t^\rho}\right)$ to investigate the convergence of linear minimization oracle (LMO) based algorithms under the assumption of star-convexity. They further proposed an adaptive warm-up scheduler based on function values for practical neural network training. Consequently, although Riabinin et al. (2026) adopts a smoothness assumption similar to the (ρ, K_0, K_ρ)-smoothness, it utilizes a different learning rate schedule and diverges from our work in its theoretical and empirical emphasis.

B Discussion on the Smoothness Assumption and Edge of Stability.

In the literature concerning the Edge of Stability (EoS), sharpness does not necessarily exhibit a monotonic decline. Specifically, during the training of neural networks, sharpness may initially undergo a gradual increase—a phenomenon termed progressive sharpening—before oscillating around the $2/\eta$ threshold (the EoS phenomenon), where η denotes the learning rate of gradient descent Cohen et al. (2021). This phenomenon is not entirely consistent with our (K_0, K_ρ)-smoothness assumption that the upper bound of sharpness decreases as the function value declines. In this section, we discuss the interplay between this phenomenon and our proposed assumption.

First, our research concentrates on the initial phase of training to explore the benefits of learning rate warm-up. If the model is initialized in a steep region of the loss landscape, a sufficiently small initial learning rate must be selected. In this case, sharpness gradually decreases as the warm-up progresses. Consequently, training starts with a sharpness reduction phase, which aligns with our hypothesis. On the other hand, if the model is initialized in a flat region, training may begin with a progressive sharpening phase, during which sharpness increases alongside the warm-up. In this case, our proposed upper bound is loose. However, as the learning rate continues to rise, the sharpness eventually exceeds the stability threshold, and training enters the EoS regime, where sharpness subsequently decreases in response to further increases in the learning rate. In this context, warm-up facilitates training by actively reducing sharpness, thereby guiding the model toward flatter regions and enabling the use of larger learning rates Kalra & Barkeshli (2024); Gilmer et al. (2022). Therefore, in both scenarios, our (ρ, K_0, K_ρ)-smoothness assumption and our theoretical analysis provide an effective explanation for the benefits of learning rate warm-up. Moreover, if the progressive sharpening occurs during the warm-up phase, the effect of learning rate warm-up may be compromised. Kalra & Barkeshli (2024) propose employing a larger initial learning rate to precipitate an early entry into the EoS regime, thereby annealing sharpness during the warm-up period.

Regarding the entire training process, progressive sharpening and EoS could occur during the mid and late training stages. Consequently, the (ρ, K_0, K_ρ)-smoothness assumption may become loose within certain level sets $\left\{\mathbf{w} : f(\mathbf{w}) \leq \bar{f}\right\}$ for some $\bar{f}$. Nevertheless, based on the discussion above, we believe that our proposed assumption remains tight when $f > \bar{f}$ or during the initial phase of training, and therefore, our analysis provides an effective theoretical justification for the benefits of learning rate warm-up.

C Experiment Settings and Additional Experiment Results

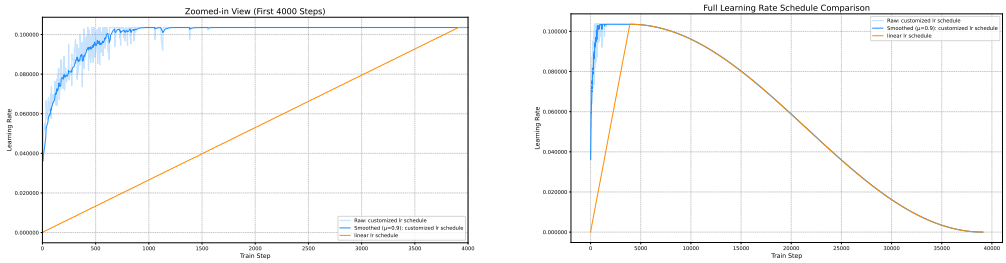
C.1 Experiment Settings of Section 2.3

Given consecutive iterates $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$, we define the update direction $\mathbf{d}_t \triangleq \mathbf{w}_{t+1} - \mathbf{w}_t$. The smoothness $\hat{L}(\mathbf{w}_t)$ is then estimated by:

$$\hat{L}(\mathbf{w}_t) = \max_{\gamma \in \{\delta_1, \ldots, \delta_n\}} \frac{\|\nabla f(\mathbf{w}_t + \gamma \mathbf{d}_t) - \nabla f(\mathbf{w}_t)\|_2}{\|\gamma \mathbf{d}_t\|_2}$$

where the sample points are $\delta_i = i/n$; we use $n = 6$, yielding $\gamma \in \{1/6, 2/6, 3/6, 4/6, 5/6, 1\}$.

The CNN configuration involves training a ResNet18 He et al. (2016) on CIFAR-10 Krizhevsky (2009) for 20 epochs. The Transformer configuration consists of training a NanoGPT model (6 blocks, 384 embedding dimension, 6 attention heads) Karpathy (2022) on the TinyShakespeare character dataset for 600 steps. Both models are trained using SGD with momentum ($\mathrm{lr} = 1e-4$). Both experiments were conducted using a single NVIDIA A100(40GB) PCIE GPU.



(a) lr schedule in warmup period (b) lr schedule in the whole training

Figure 3: A comparison between warmup learning rate schedules in ResNet training. The blue line is the theoretical warmup schedule derived in Theorem 1, and the yellow line is the standard linear warmup. We do smoothing for the blue line in the plot to make it clearer.

C.2 ResNet Results Referred in Section 3.2

We train a ResNet18 model on CIFAR-10 for 100 epochs with SGD (without momentum), using the linear warmup schedule, no warmup (constant) schedule, and our theoretical schedule $\eta_t = \frac{1}{4\sqrt{2}+4} \min\left\{\frac{1}{K_0}, \frac{1}{3^\rho K_\rho}\Delta_t^{-\rho}\right\}$. Following common practice, we set the first 10 epochs as the warm-up phase, and we do cosine decay after this initial phase. We set $\rho = 1$ for the theoretical schedule, and tune $K_0 \in \{1, 4, 8, 16\}$ and $K_\rho \in \{1/4, 1/2, 1, 4\}$. For the constant and linear warmup schedules, we set the target learning rate to be the same as the target learning rate of the theoretical schedule, i.e., $\frac{1}{4\sqrt{2}+4}\frac{1}{K_0}$, to ensure a fair comparison.

As displayed in Figure 3, we can observe that the theoretical schedule increases similarly to the linear warmup schedule, but is steeper in the first place, making a more concave curve. Also, since we use mini-batch gradients instead of full gradients, the loss is not monotonically decreasing, so there is some small oscillation before the schedule reaches a plateau. This learning rate schedule shows a valid warmup phase and achieves the plateau faster than the linear warmup schedule. Moreover, we list the performance in Table 1, showing that the theoretical schedule achieves even better performance than linear warmup, outperforming the constant schedule. Therefore, the theoretical schedule employed in Theorem 1 can be considered a valid representative of the warmup schedules, and thus, our theory built on this schedule does present the benefits of doing warmup.

Table 1: The test epoch accuracy after 100 epochs of training with different warm-up schedules. The results are reported as mean ± standard deviation over 6 independent runs.

	Theoretical Warmup	Linear Warmup	No Warmup (Constant)
Test Epoch Accuracy	0.8589 ± 0.0023	0.8577 ± 0.0023	0.8562 ± 0.0019

The data for all the 6 runs of the ResNet experiment is listed in Table 2.

Table 2: Detailed results for different warm-up schedules, including individual run data, mean, and standard deviation over 6 runs.

Warm-up Schedule	Metric	Individual Runs	Mean ± Std. Dev.
Theoretical Warmup	Val Acc.	[0.8567, 0.8600, 0.8623, 0.8595, 0.8553, 0.8599]	0.8589 ± 0.0023
	Train Loss	[0.0425, 0.0276, 0.0031, 0.0504, 0.0138, 0.0633]	0.0335 ± 0.0208
Linear Warmup	Val Acc.	[0.8549, 0.8593, 0.8570, 0.8612, 0.8550, 0.8590]	0.8577 ± 0.0023
	Train Loss	[0.1043, 0.0413, 0.0122, 0.0242, 0.0244, 0.0242]	0.0384 ± 0.0306
No Warmup	Val Acc.	[0.8532, 0.8546, 0.8559, 0.8573, 0.8578, 0.8585]	0.8562 ± 0.0019
	Train Loss	[0.0196, 0.0300, 0.0344, 0.0355, 0.0675, 0.0819]	0.0448 ± 0.0221

D EXAMPLES IN SECTION 2.2

The two examples are from Patel et al. (2022). Readers can also refer to their paper for a detailed description.

D.1 FEED FORWARD NEURAL NETWORK

We consider that σ is linear and φ is sigmoid. Suppose we have two sample points $(y, z) = (0, 0)$ and $(y, z) = (1, 1)$ with equal probability. The output $\hat{y}$ satisfies: $\hat{y} = \frac{1}{2}$ if $z = 0$ and $\hat{y} = (1 + \exp\{-w_1 w_2 w_3 w_4\})^{-1}$ if $z = 1$. The binary cross entropy with ridge penalty can be written as

$$f_{z,y}(\mathbf{w}) = -y \log \hat{y} - (1 - y) \log (1 - \hat{y}) + \frac{1}{2} \sum_{i=1}^{4} w_i^2.$$

Taking expectation over (z, y), we obtain that

$$f(\mathbf{w}) = \frac{1}{2} \log 2 + \frac{1}{2} \log (1 + \exp\{-w_1 w_2 w_3 w_4\}) + \frac{1}{2} \|\mathbf{w}\|^2 .$$

A simple calculation shows that

$$\nabla f(\mathbf{w}) = \frac{-0.5}{1 + \exp\{w_1 w_2 w_3 w_4\}} \begin{bmatrix} w_2 w_3 w_4 \\ w_1 w_3 w_4 \\ w_1 w_2 w_4 \\ w_1 w_2 w_3 \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix},$$

and

$$\nabla^2 f(\mathbf{w}) = \underbrace{\frac{-0.5}{1 + \exp\{w_1 w_2 w_3 w_4\}} \begin{bmatrix} 0 & w_3 w_4 & w_w w_4 & w_2 w_3 \\ w_3 w_4 & 0 & w_1 w_4 & w_1 w_3 \\ w_2 w_4 & w_1 w_4 & 0 & w_1 w_2 \\ w_2 w_3 & w_1 w_3 & w_1 w_2 & 0 \end{bmatrix}}_{A}$$

$$+ \underbrace{\frac{0.5 \exp(w_1 w_2 w_3 w_4)}{(1 + \exp\{w_1 w_2 w_3 w_4\})^2} \begin{bmatrix} w_2 w_3 w_4 \\ w_1 w_3 w_4 \\ w_1 w_2 w_4 \\ w_1 w_2 w_3 \end{bmatrix} \begin{bmatrix} w_2 w_3 w_4 \\ w_1 w_3 w_4 \\ w_1 w_2 w_4 \\ w_1 w_2 w_3 \end{bmatrix}^\top}_{B} + I_4$$

We first show that f is not (ρ, L_0, L_ρ)-smooth for any $0 \le \rho < 2$. Let $\mathbf{w} = (0, w_4, w_4, w_4)^\top$. Then

$$\|\nabla f(\mathbf{w})\|_1 \le \frac{1}{4}|w_4|^3 + 3|w_4|.$$

Since $\left\|\nabla^2 f(\mathbf{w})\right\|_F$ is lower bounded by $\nabla^2 f(\mathbf{w})_{(1,1)}$, we have

$$\left\|\nabla^2 f(\mathbf{w})\right\|_F \ge \frac{1}{8} w_4^6.$$

Therefore, if f is (ρ, L_0, L_ρ)-smooth, then it must hold that $\rho \ge 2$.

Next, we show that f is (ρ, K_0, K_ρ)-smooth for some $\rho > 0$. Note that

$$\|A\|_F^2 = \frac{0.5}{(1 + \exp\{w_1 w_2 w_3 w_4\})^2} \left(\sum_{1 \le i < j \le 4} (w_i w_j)^2 \right) \le \left(\sum_{i=1}^{4} w_i^2 \right)^2,$$

$$\begin{aligned} \|B\|_F^2 &= \frac{0.25 \exp\{2 w_1 w_2 w_3 w_4\}}{(1 + \exp\{w_1 w_2 w_3 w_4\})^4} \left(\sum_{1 \le i < j < k \le 4} (w_i w_j w_k)^4 + 2 (w_1 w_2 w_3 w_4)^2 \sum_{1 \le i < j \le 4} (w_i w_j)^2 \right) \\ &\le \sum_{1 \le i < j < k \le 4} (w_i w_j w_k)^4 + \sum_{1 \le i < j \le 4} (w_i w_j)^2 \\ &\le \left(\sum_{i=1}^{4} w_i^2 \right)^6 + \left(\sum_{i=1}^{4} w_i^2 \right)^2, \end{aligned}$$

where in the first inequality we use $\frac{e^{2x}}{(1+e^x)^4}x^2 \leq 1$ and $\frac{e^{2x}}{(1+e^x)^4} \leq 1$. Combining the above results, we obtain that

$$\|\nabla^2 f(\mathbf{w})\|_F^2 \leq 3\left(\sum_{i=1}^4 w_i^2\right)^6 + 6\left(\sum_{i=1}^4 w_i^2\right)^2 + 12.$$

Moreover, it is not hard to see that $f^\star \leq \log 2$ and

$$f(\mathbf{w}) - f^\star \geq \frac{1}{2}\sum_{i=1}^4 w_i^2 - \frac{1}{2}\log 2.$$

Therefore, we conclude that f is (ρ, K_0, K_1)-smooth with some $K_0, K_1 > 0$ and $\rho \geq 3$.

D.2 Recurrent Neural Network

We consider that σ is linear and φ is sigmoid. Suppose we have two sample points $(\mathbf{z}, y) = (1,0,0,0,1)$ and $(0,0,0,0,0)$ with equal probability. Fix $h_0 = 0$ and $w_3 = 1$. We have $\hat{y} = \frac{\exp\{w_1^3 w_2 z_0\}}{1+\exp\{w_1^3 w_2 z_0\}}$. The binary cross entropy with ridge penalty can be written as

$$f_{\mathbf{z},y}(\mathbf{w}) = -y\log\hat{y} - (1-y)\log(1-\hat{y}) + \frac{1}{2}\sum_{i=1}^2 w_i^2.$$

Taking expectation over $(\mathbf{z}, y)$ we obtain that

$$f(\mathbf{w}) = \frac{1}{2}\left(\log 2 + \log\left(1+\exp\left(w_1^3 w_2\right)\right) - w_1^3 w_2 + w_1^2 + w_2^2\right)$$

$$\nabla f(\mathbf{w}) = \begin{bmatrix} \frac{-3w_1^2 w_2}{2}\frac{1}{1+\exp\left(w_1^3 w_2\right)} + w_1 \\ \frac{-w_1^3}{2}\frac{1}{1+\exp\left(w_1^3 w_2\right)} + w_2 \end{bmatrix}$$

$$\nabla^2 f(\mathbf{w}) = \begin{bmatrix} \frac{9w_1^4 w_2^2 \exp\left(w_1^3 w_2\right)}{2\left(1+\exp\left(w_1^3 w_2\right)\right)^2} - \frac{3w_1 w_2}{1+\exp\left(w_1^3 w_2\right)} + 1 & \frac{3w_1^5 w_2 \exp\left(w_1^3 w_2\right)}{2\left(1+\exp\left(w_1^3 w_2\right)\right)^2} - \frac{3w_1^2}{2}\frac{1}{1+\exp\left(w_1^3 w_2\right)} \\ \frac{3w_1^5 w_2 \exp\left(w_1^3 w_2\right)}{2\left(1+\exp\left(w_1^3 w_2\right)\right)^2} - \frac{3w_1^2}{2}\frac{1}{1+\exp\left(w_1^3 w_2\right)} & \frac{w_1^6 \exp\left(w_1^3 w_2\right)}{2\left(1+\exp\left(w_1^3 w_2\right)\right)^2} + 1 \end{bmatrix}.$$

We first show that f is not (ρ, L_0, L_ρ)-smooth for any $0 \leq \rho < 2$. Let $w_2 = 0$. We have

$$\|f(\mathbf{w})\|_1 = |w_1| + \frac{|w_1|^3}{4}.$$

Consider the bottom right entry of $\nabla^2 f(\mathbf{w})$, we have

$$\left\|\nabla^2 f(\mathbf{w})\right\|_F > \frac{w_1^6}{8}.$$

Therefore, if f is (ρ, L_0, L_ρ)-smooth, then it must hold that $\rho \geq 2$.

Next, we show that f is (ρ, K_0, K_ρ)-smooth for some $\rho \geq 0$. We directly compute

$$\begin{aligned}
\|\nabla^2 f(\mathbf{w})\|_F^2 &\leq \frac{\exp\left\{2w_1^3 w_2\right\}}{\left(1+\exp\left\{w_1^3 w_2\right\}\right)^4}\left(\frac{243}{4}w_1^8 w_2^4 + 9w_1^{10}w_2^2 + \frac{1}{2}w_1^{12}\right) \\
&\quad + \frac{9}{1+\exp\left\{w_1^3 w_2\right\}}\left(3w_1^2 w_2^2 + w_1^4\right) + 5 \\
&= \underbrace{\frac{\exp\left\{2w_1^3 w_2\right\}}{\left(1+\exp\left\{w_1^3 w_2\right\}\right)^4}w_1^8\left(\frac{243}{4}w_2^4 + 9w_1^2 w_2^2\right)}_{A} + \underbrace{\frac{\exp\left\{2w_1^3 w_2\right\}}{\left(1+\exp\left\{w_1^3 w_2\right\}\right)^4}\frac{1}{2}w_1^{12}}_{B} \\
&\quad + \underbrace{\frac{9}{1+\exp\left\{w_1^3 w_2\right\}}\left(3w_1^2 w_2^2 + w_1^4\right)}_{C} + 5.
\end{aligned}$$

Since $\frac{e^{2x}}{(1+e^x)^4}x^2 \leq 1$ and $\frac{e^{2x}}{(1+e^x)^4} \leq 1$, we have

$$\begin{aligned}
A &= \frac{\exp\left\{2w_1^3w_2\right\}}{\left(1+\exp\left\{w_1^3w_2\right\}\right)^4}\left(w_1^3w_2\right)^2 w_1^2\left(\frac{243}{4}w_2^2+9w_1^2\right)\\
&\leq \frac{243}{4}w_1^2(w_1^2+w_2^2) \leq \frac{243}{4}(w_1^2+w_2^2)^2,\\
B &\leq \frac{1}{2}w_1^{12} \leq \frac{1}{2}(w_1^2+w_2^2)^6,\\
C &\leq 9\times(3w_1^2w_2^2+w_1^4) \leq 9\times\frac{3}{2}(w_1^2+w_2^2)^2.
\end{aligned}$$

Combining the above results, we obtain that

$$\begin{aligned}
\|\nabla^2 f(\theta)\|_F^2 \leq& \frac{243}{4}\left(w_1^2+w_2^2\right)^2+\frac{1}{2}(w_1^2+w_2^2)^6+\frac{27}{2}(w_1^2+w_2^2)^2+5\\
\leq& 256\left(w_1^2+w_2^2\right)^2+\left(w_1^2+w_2^2\right)^6+5
\end{aligned}$$

Moreover, note that $f^\star \leq \log 2$ and

$$f(\mathbf{w})-f^\star \geq \frac{1}{2}\left(w_1^2+w_2^2-\log 2\right).$$

We conclude that f is (ρ, K_0, K_1)-smooth with some $K_0, K_1 > 0$ and $\rho \geq 3$.

E Proofs for Section 2

E.1 Proof of Lemma 1

Proof. Since f is (ρ, L_0, L_ρ)-smooth with $0 \leq \rho < 2$, by (Li et al., 2023a, Lemma 3.5) we have that for all $\mathbf{w} \in \mathbb{R}^d$,

$$f(\mathbf{w})-f^\star \geq \frac{\|\nabla f(\mathbf{w})\|^2}{2L_0+2^{\rho+1}L_\rho\|\nabla f(\mathbf{w})\|^\rho}.$$

If $2L_0 \leq 2^{\rho+1}L_\rho\|\nabla f(\mathbf{w})\|^\rho$, we obtain that

$$f(\mathbf{w})-f^\star \geq \frac{\|\nabla f(\mathbf{w})\|^2}{2^{\rho+2}L_\rho\|\nabla f(\mathbf{w})\|^\rho} = \frac{\|\nabla f(\mathbf{w})\|^{2-\rho}}{2^{\rho+2}L_\rho}.$$

By the definition of (ρ, L_0, L_ρ)-smoothness we have

$$\|\nabla^2 f(\mathbf{w})\| \leq L_0+L_\rho\|\nabla f(\mathbf{w})\|^\rho \leq L_0+L_\rho^{\frac{2}{2-\rho}}2^{\frac{\rho(\rho+2)}{2-\rho}}\left(f(\mathbf{w})-f^\star\right)^{\frac{\rho}{2-\rho}}.$$

If $2L_0 > 2^{\rho+1}L_\rho\|\nabla f(\mathbf{w})\|^\rho$, $\|\nabla f(\mathbf{w})\|$ is bounded:

$$\|\nabla f(\mathbf{w})\|^\rho < \frac{L_0}{2^\rho L_\rho}.$$

Again, by the definition of (ρ, L_0, L_ρ)-smoothness we have

$$\|\nabla^2 f(\mathbf{w})\| \leq L_0+L_\rho\|\nabla f(\mathbf{w})\|^\rho \leq L_0+\frac{L_0}{2^\rho} \leq 2L_0$$

Combining the two cases, we obtain that

$$\|\nabla^2 f(\mathbf{w})\| \leq 2L_0+L_\rho^{\frac{2}{2-\rho}}2^{\frac{\rho(\rho+2)}{2-\rho}}\left(f(\mathbf{w})-f^\star\right)^{\frac{\rho}{2-\rho}}.$$

This completes the proof. □

E.2 Generalized Smoothness for Differentiable Functions

Definition 1 admits the following extension, which does not require a Hessian.

Definition 2. *A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ with $f^\star = \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > -\infty$ is (ρ, K_0, K_ρ)-smooth if, for $K_0, K_\rho \geq 0$ and $\rho > 0$,*

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \leq \|\mathbf{y} - \mathbf{x}\| \int_0^1 \left[K_0 + K_\rho \Delta\big(\mathbf{x} + t(\mathbf{y} - \mathbf{x})\big)^\rho\right] \mathrm{d}t, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \quad (6)$$

where $\Delta(\mathbf{x}) = f(\mathbf{x}) - f^\star$.

Lemma 3 (Equivalence with the Hessian definition.)**.** *For $f \in C^2$, Definition 2 is equivalent to Definition 1, with the same parameters.*

Proof. Let $\mathbf{d} := \mathbf{y} - \mathbf{x}$. If Definition 1 holds, then

$$\begin{aligned}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| &= \left\| \int_0^1 \nabla^2 f(\mathbf{x} + t\mathbf{d})\mathbf{d}\, \mathrm{d}t \right\| \\ &\leq \|\mathbf{d}\| \int_0^1 \left\|\nabla^2 f(\mathbf{x} + t\mathbf{d})\right\| \mathrm{d}t \leq \|\mathbf{d}\| \int_0^1 \left[K_0 + K_\rho \Delta(\mathbf{x} + t\mathbf{d})^\rho\right] \mathrm{d}t.\end{aligned}$$

Conversely, fix $\mathbf{x}$ and a unit vector $\mathbf{v}$. Applying (6) with $\mathbf{y} = \mathbf{x} + s\mathbf{v}$ gives

$$\frac{\|\nabla f(\mathbf{x} + s\mathbf{v}) - \nabla f(\mathbf{x})\|}{|s|} \leq \int_0^1 \left[K_0 + K_\rho \Delta(\mathbf{x} + ts\mathbf{v})^\rho\right] \mathrm{d}t.$$

Letting $s \to 0$ yields $\left\|\nabla^2 f(\mathbf{x})\mathbf{v}\right\| \leq K_0 + K_\rho \Delta(\mathbf{x})^\rho$. Taking the supremum over all unit vectors $\mathbf{v}$ proves (1). □

E.3 Proof of Lemma 2

For any two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, let $\mathbf{d} = \mathbf{y} - \mathbf{x}$ and define

$$h(t) := \int_0^t \left[K_0 + K_\rho \Delta(\mathbf{x} + v\mathbf{d})^\rho\right] \mathrm{d}v, \qquad t \in [0, 1]. \quad (7)$$

Definition 2 implies that ∇f is locally Lipschitz, since its integrand is locally bounded. Applying (6) to $\mathbf{x}$ and $\mathbf{x} + t\mathbf{d}$ gives

$$\|\nabla f(\mathbf{x} + t\mathbf{d}) - \nabla f(\mathbf{x})\| \leq t\,\|\mathbf{d}\| \int_0^1 \left[K_0 + K_\rho \Delta(\mathbf{x} + st\mathbf{d})^\rho\right] \mathrm{d}s = \|\mathbf{d}\|\, h(t). \quad (8)$$

In particular, $\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \leq h(1)\,\|\mathbf{y} - \mathbf{x}\|$. Integrating the directional derivative also yields

$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{x}) &= \langle \nabla f(\mathbf{x}), \mathbf{d}\rangle + \int_0^1 \langle \nabla f(\mathbf{x} + t\mathbf{d}) - \nabla f(\mathbf{x}), \mathbf{d}\rangle \, \mathrm{d}t \\ &\leq \langle \nabla f(\mathbf{x}), \mathbf{d}\rangle + \|\mathbf{d}\|^2 \int_0^1 h(t)\mathrm{d}t \\ &= \langle \nabla f(\mathbf{x}), \mathbf{d}\rangle + \|\mathbf{d}\|^2 \int_0^1 (1 - t)\left[K_0 + K_\rho \Delta(\mathbf{x} + t\mathbf{d})^\rho\right] \mathrm{d}t. \end{aligned} \quad (9)$$

Thus a uniform bound on the integrand gives the required descent estimate. We establish this bound below directly from the differentiable definition.

Lemma 4 (Lemma A.3, Li et al. (2023a))**.** *Let $u : [a, b] \to [0, \infty)$ and $l : [0, \infty) \to (0, \infty)$ be two continuous functions. Suppose $u'(t) \leq l(u(t))$ for all $t \in [a, b]$, then it holds for all $t \in [a, b]$ that*

$$\int_{u(a)}^{u(t)} \frac{1}{l(w)} \mathrm{d}w \leq t - a.$$

Lemma 5. *Suppose f satisfies Definition 2. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, define $h(t)$ as in (7), let $m > 0$, and set $a = K_0 + K_\rho\big(\Delta(\mathbf{x}) + m\big)^\rho$. If $a\|\mathbf{y}-\mathbf{x}\|^2 + \|\nabla f(\mathbf{x})\|\|\mathbf{y}-\mathbf{x}\| \le m$, then*

$$\Delta\big(\mathbf{x} + t(\mathbf{y}-\mathbf{x})\big) \le \Delta(\mathbf{x}) + m, \quad h(t) \le at, \quad \forall t \in [0,1].$$

Consequently,

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \le a\|\mathbf{y}-\mathbf{x}\|,$$
$$f(\mathbf{y}) \le f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{a}{2}\|\mathbf{y}-\mathbf{x}\|^2.$$

Proof. Let $\mathbf{d} = \mathbf{y}-\mathbf{x}$, $D = \|\mathbf{d}\|$, and $q(t) = f(\mathbf{x}+t\mathbf{d}) - f(\mathbf{x})$. Suppose $q(t) > m$ for some $t \in [0,1]$. By continuity, there is a first hitting time $t_\star \in (0,1)$ such that $q(t_\star) = m$ and $q(t) \le m$ for $0 \le t \le t_\star$. On this part of the segment, $\Delta(\mathbf{x}+t\mathbf{d}) \le \Delta(\mathbf{x}) + m$, so (8) gives $\|\nabla f(\mathbf{x}+t\mathbf{d}) - \nabla f(\mathbf{x})\| \le atD$. Therefore,

$$\begin{aligned} q(t_\star) &= \int_0^{t_\star} \langle \nabla f(\mathbf{x}+t\mathbf{d}), \mathbf{d}\rangle \,\mathrm{d}t \\ &\le t_\star\|\nabla f(\mathbf{x})\| D + \frac{a}{2}t_\star^2 D^2 \\ &\le t_\star\big(\|\nabla f(\mathbf{x})\| D + aD^2\big) \le t_\star m < m, \end{aligned}$$

contradicting $q(t_\star) = m$. Hence the function-gap bound holds on the entire segment, and $h(t) \le at$ follows from (7). Substituting this bound into (8) and (9) gives the gradient bound and the descent estimate, since $\int_0^1 at\,\mathrm{d}t = a/2$. □

Lemma 6 (Bounded Gradient)**.** *Suppose Assumption 2 holds. Let $\Delta = f(\mathbf{x}) - f^\star$. It holds that*

$$\|\nabla f(\mathbf{x})\| \le 2\sqrt{K_0\Delta + K_\rho 3^\rho \Delta^{\rho+1}}.$$

Proof. By Lemma 5, let $m = 2\Delta$, we obtain that $h(1) \le K_0 + K_\rho(3\Delta)^\rho =: a$, if

$$\|\mathbf{y}-\mathbf{x}\|^2\left(K_0 + K_\rho(3\Delta)^\rho\right) + \|\mathbf{y}-\mathbf{x}\|\|\nabla f(\mathbf{x})\| \le 2\Delta.$$

Equivalently,

$$\|\mathbf{y}-\mathbf{x}\| \le \frac{-\|\nabla f(\mathbf{x})\| + \sqrt{\|\nabla f(\mathbf{x})\|^2 + 8a\Delta}}{2a} =: r.$$

For $\mathbf{y}$ satisfying $\|\mathbf{y}-\mathbf{x}\| \le r$, the descent estimate in Lemma 5 gives

$$f(\mathbf{y}) - f(\mathbf{x}) \le \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{a}{2}\|\mathbf{y}-\mathbf{x}\|^2.$$

Letting $\mathbf{y} = \mathbf{x} - \frac{\eta}{\|\nabla f(\mathbf{x})\|}\nabla f(\mathbf{x})$, we obtain that

$$-\Delta \le f(\mathbf{y}) - f(\mathbf{x}) \le -\eta\|\nabla f(\mathbf{x})\| + \frac{a}{2}\eta^2.$$

Therefore, we obtain that

$$g(\eta) := \frac{a}{2}\eta^2 - \eta\|\nabla f(\mathbf{x})\| + \Delta \ge 0, \forall \eta \in [0, r]. \tag{10}$$

It is not hard to see that $\arg\min_{\eta\in\mathbb{R}} g(\eta) = \frac{\|\nabla f(\mathbf{x})\|}{a}$. We then consider two cases: $\frac{\|\nabla f(\mathbf{x})\|}{a} \le r$ and $\frac{\|\nabla f(\mathbf{x})\|}{a} > r$. Suppose $\frac{\|\nabla f(\mathbf{x})\|}{a} \le r$. Equivalently, $\|\nabla f(\mathbf{x})\| \le \sqrt{a\Delta}$. By (10) we need

$$\|\nabla f(\mathbf{x})\| \le \sqrt{2a\Delta}.$$

Now suppose $\frac{\|\nabla f(\mathbf{x})\|}{a} > r$. Equivalently, $\|\nabla f(\mathbf{x})\| > \sqrt{a\Delta}$. By (10) we need

$$\frac{a}{2}r^2 - r\|\nabla f(\mathbf{x})\| + \Delta \ge 0.$$

Equivalently,

$$\|\nabla f(\mathbf{x})\| \le \sqrt{\frac{8}{3}a\Delta}.$$

Combining the above two cases, we conclude that $\|\nabla f(\mathbf{x})\| \le \sqrt{\frac{8}{3}a\Delta} \le 2\sqrt{a\Delta} = 2\sqrt{K_0\Delta + K_\rho 3^\rho\Delta^{\rho+1}}$. □

Proof of Lemma 2

Proof. By Lemma 5, for any $m > 0$, we have $h(1) \leq K_0 + K_\rho (m+\Delta)^\rho =: a$, if

$$\|\mathbf{y} - \mathbf{x}\| \leq \frac{2m}{\|\nabla f(\mathbf{x})\| + \sqrt{\|\nabla f(\mathbf{x})\|^2 + 4am}} =: r.$$

Let $A = K_0\Delta + K_\rho 3^\rho \Delta^{\rho+1}$ and $B = A + m\left(K_0 + K_\rho (m+\Delta)^\rho\right)$. By Lemma 6, we have $\|\nabla f(\mathbf{x})\| \leq 2\sqrt{A}$, and thus

$$r \geq \frac{m}{\sqrt{A} + \sqrt{B}}.$$

Let

$$m = \max\left\{\Delta, \frac{1}{3}\left(\frac{K_0}{K_1}\right)^{1/\rho}\right\}.$$

If $K_0 \leq K_1 3^\rho \Delta^\rho$, we have $m = \Delta$, $A \leq 2K_1 3^\rho \Delta^{\rho+1}$ and

$$B = 2K_0\Delta + K_1 3^\rho \Delta^{\rho+1} + K_1 2^\rho \Delta^{\rho+1} \leq 4K_1 3^\rho \Delta^{\rho+1}.$$

Thus

$$\frac{m}{\sqrt{A} + \sqrt{B}} \geq \frac{1}{\left(2+\sqrt{2}\right)\sqrt{3^\rho K_1}\Delta^{\frac{\rho-1}{2}}} =: C_1 \Delta^{-\frac{\rho-1}{2}}.$$

If $K_0 > K_1 3^\rho \Delta^\rho$, we have $m = \frac{1}{3}\left(\frac{K_0}{K_1}\right)^{1/\rho}$, $A \leq 2K_0\Delta \leq 2K_0 m$ and

$$B \leq 2K_0 m + m\left(K_0 + K_1 (2m)^\rho\right) \leq 4mK_0.$$

Thus

$$\frac{m}{\sqrt{A} + \sqrt{B}} \geq \frac{m}{\left(2+\sqrt{2}\right)\sqrt{K_0 m}} = \frac{1}{2\sqrt{3}+\sqrt{6}} \frac{K_0^{\frac{1}{2\rho}-\frac{1}{2}}}{K_1^{\frac{1}{2\rho}}} =: C_2.$$

Next we bound $a = K_0 + K_1 (m+\Delta)^\rho$. If $m = \Delta$, we have

$$K_0 + K_1 (m+\Delta)^\rho = K_0 + K_1 2^\rho \Delta^\rho.$$

If $m = \frac{1}{3}\left(\frac{K_0}{K_1}\right)^{1/\rho}$, we have

$$K_0 + K_1 (m+\Delta)^\rho \leq K_0 + K_1 2^\rho m^\rho = K_0 + \left(\frac{2}{3}\right)^\rho K_0 \leq 2K_0.$$

Combining the above results, we get the desired result. □

E.4 Closeness of (ρ, K_0, K_ρ)-smoothness

Lemma 7. *Let $f_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \ldots, m$ be $(\rho_i, K_{0,i}, K_{\rho_i,i})$-smooth with $\rho_i > 0$ and lower bounded by $f_i^\star$. Define $F(\mathbf{x}) := \sum_{i=1}^m f_i(\mathbf{x})$, $\underline{F}^\star := \sum_{i=1}^m f_i^\star$ and $F^\star := \inf_{\mathbf{x}} F(\mathbf{x}) \geq \underline{F}^\star$. Then we have that*

$$\left\|\nabla^2 F(\mathbf{x})\right\|_2 \leq \widetilde{K}_0 + \widetilde{K}_\rho \left(F(\mathbf{x}) - F^\star\right)^{\rho_{\max}},$$

where

$$\widetilde{K}_0 = \sum_{i=1}^m K_{0,i} + \left(C_{\max}(F^\star - \underline{F}^\star)^{\rho_{\max}} + 1\right)\sum_{i=1}^m K_{\rho_i,i}, \qquad \widetilde{K}_\rho = C_{\max}\sum_{i=1}^m K_{\rho_i,i},$$

and $C_{\max} = \max\{1, 2^{\rho_{\max}}\}$.

Proof. Let $F(\mathbf{x}) = \sum_{i=1}^m f_i(\mathbf{x})$, then

$$\nabla^2 F(\mathbf{x}) = \sum_{i=1}^m \nabla^2 f_i(\mathbf{x}), \qquad F(\mathbf{x}) - \underline{F}^\star = \sum_{i=1}^m (f_i(\mathbf{x}) - f_i^\star).$$

By the triangle inequality,

$$\|\nabla^2 F(\mathbf{x})\|_2 \le \sum_{i=1}^m \|\nabla^2 f_i(\mathbf{x})\|_2 \le \sum_{i=1}^m K_{0,i} + \sum_{i=1}^m K_{\rho_i,i}(f_i(\mathbf{x}) - f_i^\star)^{\rho_i}.$$

Let $\rho_{\max} := \max_i \rho_i$. Since $f_i(\mathbf{x}) - f_i^\star \le F(\mathbf{x}) - \underline{F}^\star$,

$$(f_i(\mathbf{x}) - f_i^\star)^{\rho_i} \le 1 + (F(\mathbf{x}) - \underline{F}^\star)^{\rho_{\max}}.$$

Using this bound, we get

$$\|\nabla^2 F(\mathbf{x})\|_2 \le \sum_{i=1}^m K_{0,i} + \sum_{i=1}^m K_{\rho_i,i} + \left(\sum_{i=1}^m K_{\rho_i,i}\right)(F(\mathbf{x}) - \underline{F}^\star)^{\rho_{\max}}.$$

Using

$$(a+b)^r \le C_r\,(a^r + b^r)\,,\ C_r = \begin{cases} 1, & 0 < r \le 1 \\ 2^{r-1}, & r \ge 1 \end{cases}$$

with $a, b, r > 0$ and $F(\mathbf{x}) - \underline{F}^\star = F(\mathbf{x}) - F^\star + F^\star - \underline{F}^\star$, we obtain the desired result. □

Lemma 8. *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is (ρ, K_0, K_ρ)-smooth and lower bounded by $f^\star$. For $A \in \mathbb{R}^{d\times n}$ and $\mathbf{b} \in \mathbb{R}^d$, define*

$$g(\mathbf{x}) := f(A\mathbf{x} + \mathbf{b}).$$

Then g is $(\widetilde{K}_0, \widetilde{K}_\rho)$-smooth with

$$\widetilde{K}_0 = \|A\|_2^2\,(K_0 + c_\rho K_\rho\,(g^\star - f^\star)^\rho)\,, \qquad \widetilde{K}_\rho = \|A\|_2^2 c_\rho K_\rho,$$

where $g^\star := \inf_{\mathbf{x}} g(\mathbf{x}) \ge f^\star$ and $c_\rho = \max\left\{1, 2^{\rho-1}\right\}$.

Proof. Note that

$$\nabla^2 g(\mathbf{x}) = A^\top \nabla^2 f(A\mathbf{x} + \mathbf{b}) A.$$

Applying the (ρ, K_0, K_ρ)-smoothness of f and using the rescaling argument similar to the previous proof, we obtain the desired result. □

Lemma 9 (Propositions B.4, B.5, Alimisis et al. (2025))**.** *There exists (L_0, L_1)-smooth functions $f, f_1, f_2 : \mathbb{R}^d \to \mathbb{R}$ such that $g(\mathbf{x}) := f_1(\mathbf{x}) + f_2(\mathbf{x})$ and $h(\mathbf{x}) := f(A\mathbf{x} + \mathbf{b})$ with $A \in \mathbb{R}^{d\times n}, \mathbf{b} \in \mathbb{R}^d$ are not (L_0, L_1)-smooth functions.*

F Proofs for Section 3

For simplicity, we let $\Delta_t = f(\mathbf{w}_t) - f^\star$, $r_t = \min\left\{C_1 \Delta_t^{-\frac{\rho-1}{2}}, C_2\right\}$ and $L_t = 2K_0 + K_\rho\,(2\Delta_t)^\rho$.

F.1 Proof of Theorem 1

Proof. Let $c_\eta = \frac{1}{4\sqrt{2}+4}$, $\mathbf{g}_t = \nabla f(\mathbf{w}_t)$ and $a_t = K_0 + K_\rho\,(3\Delta_t)^\rho$. By Lemma 6, $\|\mathbf{g}_t\|^2 \le 4a_t\Delta_t$. Moreover, the learning rate $\eta_t = c_\eta \min\left\{\frac{1}{K_0}, \frac{1}{3^\rho K_\rho \Delta_t^\rho}\right\}$ satisfies $\eta_t a_t \le 2c_\eta$. Therefore,

$$\begin{aligned} &a_t \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 + \|\mathbf{g}_t\|\,\|\mathbf{w}_{t+1} - \mathbf{w}_t\| \\ &= \eta_t \|\mathbf{g}_t\|^2 (1 + a_t\eta_t) \le 8c_\eta\,(1 + 2c_\eta)\,\Delta_t = \Delta_t \le 2\Delta_t. \end{aligned}$$

Applying Lemma 5 with $m = 2\Delta_t$ gives

$$f(\mathbf{w}_{t+1}) \leq f(\mathbf{w}_t) - \eta_t \|\mathbf{g}_t\|^2 + \frac{a_t}{2}\eta_t^2 \|\mathbf{g}_t\|^2$$
$$\leq f(\mathbf{w}_t) - \frac{\eta_t}{2} \|\mathbf{g}_t\|^2 \leq f(\mathbf{w}_t).$$

Note also that

$$\eta_t \leq \frac{1}{\left(2+2\sqrt{2}\right)\left(K_0 + K_\rho \left(3\Delta_t\right)^\rho\right)} \leq \frac{1}{2K_0 + K_\rho \left(2\Delta_t\right)^\rho} = \frac{1}{L_t}. \tag{11}$$

Telescoping the above inequation from $t = 0$ to $t = T-1$ we obtain that

$$\sum_{t=0}^{T-1} \eta_t \|\nabla f(\mathbf{w}_t)\|^2 \leq 2\left(f(\mathbf{w}_0) - f(\mathbf{w}_T)\right) \leq 2\Delta_0.$$

Note that $1/\eta_t \leq (4\sqrt{2}+4)\left(K_0 + K_\rho \left(3\Delta_t\right)^\rho\right)$. Using the AM–HM inequality, we have

$$\sum_{t=0}^{T-1} \eta_t \geq \frac{T^2}{\sum_{t=0}^{T-1} 1/\eta_t} \geq \frac{1}{4\sqrt{2}+4} \frac{T^2}{\sum_{t=0}^{T-1} K_0 + K_\rho \left(3\Delta_t\right)^\rho}.$$

This completes the proof for the increasing learning rate.

Now suppose we use the constant learning rate $\eta \leq \frac{1}{4\sqrt{2}+4} \min\left\{\frac{1}{K_0}, \frac{1}{K_\rho(3\Delta_0)^\rho}\right\}$. Applying the same one-step argument as above with η in place of η_0 gives

$$f(\mathbf{w}_1) \leq f(\mathbf{w}_0) - \frac{\eta}{2} \|\nabla f(\mathbf{w}_0)\|^2 \leq f(\mathbf{w}_0).$$

This means $\Delta_1 \leq \Delta_0$ and thus $\eta \leq \frac{1}{4\sqrt{2}+4} \min\left\{\frac{1}{K_0}, \frac{1}{K_\rho(3\Delta_1)^\rho}\right\}$. By induction, it is not hard to see that $\eta \leq \frac{1}{4\sqrt{2}+4} \min\left\{\frac{1}{K_0}, \frac{1}{K_\rho(3\Delta_t)^\rho}\right\}, \forall t \in [T]$. Therefore, following the same one-step analysis as above, we have

$$\eta \sum_{t=0}^{T-1} \|\nabla f(\mathbf{w}_t)\|^2 \leq 2\Delta_0.$$

This completes the proof. □

F.2 Proof of Theorem 2

We first present the following local smoothness property and then prove the theorem using convexity and the gradient bound.

Lemma 10. *Under the same conditions as in Lemma 2, for any $\mathbf{x}_1, \mathbf{x}_2$ such that $\|\mathbf{x}_1 - \mathbf{x}\| \leq r(\mathbf{x})$ and $\|\mathbf{x}_2 - \mathbf{x}\| \leq r(\mathbf{x})$, where $r(\mathbf{x})$ is defined in Lemma 2. Then we have*

$$\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\| \leq L(\mathbf{x}) \|\mathbf{x}_1 - \mathbf{x}_2\|,$$

and

$$f(\mathbf{x}_2) \leq f(\mathbf{x}_1) + \langle \nabla f(\mathbf{x}_1), \mathbf{x}_2 - \mathbf{x}_1 \rangle + \frac{L(\mathbf{x})}{2} \|\mathbf{x}_2 - \mathbf{x}_1\|^2,$$

where $L(\mathbf{x})$ is defined in Lemma 2.

Proof. Let $m > 0$, $\Delta = f(\mathbf{x}) - f^\star$, $a = K_0 + K_\rho (m+\Delta)^\rho$ and $r := \frac{2m}{\|\nabla f(\mathbf{x})\| + \sqrt{\|\nabla f(\mathbf{x})\|^2 + 4am}}$. Let $\|\mathbf{y} - \mathbf{x}\| \le r$, by Lemma 5, we have

$$\begin{aligned} f(\mathbf{y}) &\le f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{a}{2} \|\mathbf{y} - \mathbf{x}\|^2 \\ &\le f(\mathbf{x}) + \|\nabla f(\mathbf{x})\| \|\mathbf{y} - \mathbf{x}\| + \frac{a}{2} \|\mathbf{y} - \mathbf{x}\|^2 \\ &\le f(\mathbf{x}) + \frac{1}{8a} \left(\sqrt{\|\nabla f(\mathbf{x})\|^2 + 4am} - \|\nabla f(\mathbf{x})\| \right) \left(3\|\nabla f(\mathbf{x})\| + \sqrt{\|\nabla f(\mathbf{x})\|^2 + 4am} \right) \\ &= f(\mathbf{x}) + \frac{m}{2} + \frac{\|\nabla f(\mathbf{x})\|}{4a} \left(\sqrt{\|\nabla f(\mathbf{x})\|^2 + 4am} - \|\nabla f(\mathbf{x})\| \right) \\ &= f(\mathbf{x}) + \frac{m}{2} + \frac{m}{1 + \sqrt{1 + \frac{4am}{\|\nabla f(\mathbf{x})\|^2}}} \le f(\mathbf{x}) + m. \end{aligned} \tag{12}$$

Then, for $\|\mathbf{x}_1 - \mathbf{x}\| \le r$ and $\|\mathbf{x}_2 - \mathbf{x}\| \le r$, we have

$$\begin{aligned} \|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\| &\le \|\mathbf{x}_1 - \mathbf{x}_2\| \int_0^1 \left\| \nabla^2 f\left(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)\right) \right\| \mathrm{d}t \\ &\le \|\mathbf{x}_1 - \mathbf{x}_2\| \int_0^1 \left(K_0 + K_\rho \left(f(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)) - f^\star \right)^\rho \right) \mathrm{d}t \\ &\le \|\mathbf{x}_2 - \mathbf{x}_1\| \left(K_0 + K_\rho \left(f(\mathbf{x}) - f^\star + m \right)^\rho \right), \end{aligned}$$

where the second inequality is due to Assumption 2 and in the last inequality we use $\|\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1) - \mathbf{x}\| \le r$ for $t \in [0,1]$ and (12). Setting $m = \max\left\{ f(\mathbf{x}) - f^\star, \frac{1}{3}\left(\frac{K_0}{K_\rho}\right)^{1/\rho} \right\}$ and following the proof of Lemma 2 we obtain the desired result. □

Proof of Theorem 2

Proof. Let $\Delta_t = f(\mathbf{w}_t) - f(\mathbf{w}^\star)$, $D_t = \|\mathbf{w}_t - \mathbf{w}^\star\|$, $\mathbf{g}_t = \nabla f(\mathbf{w}_t)$, and $c_\eta = 1/(8\sqrt{2} + 8)$. If $\Delta_t = 0$, GD has reached a minimizer and the conclusions are immediate, so we consider iterations with $\Delta_t > 0$. By Lemma 6,

$$\|\mathbf{g}_t\|^2 \le 4\Delta_t \left(K_0 + 3^\rho K_\rho \Delta_t^\rho \right).$$

The adaptive learning rate therefore satisfies

$$\eta_t \|\mathbf{g}_t\|^2 \le 8c_\eta \Delta_t = (\sqrt{2} - 1)\Delta_t. \tag{13}$$

Convexity gives $\langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}^\star \rangle \ge \Delta_t$, so

$$\begin{aligned} D_{t+1}^2 &= D_t^2 - 2\eta_t \langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}^\star \rangle + \eta_t^2 \|\mathbf{g}_t\|^2 \\ &\le D_t^2 - (3 - \sqrt{2})\eta_t \Delta_t \le D_t^2 - \eta_t \Delta_t. \end{aligned}$$

Telescoping from $t = 0$ to $T - 1$ gives

$$\sum_{t=0}^{T-1} \eta_t \Delta_t \le D_0^2. \tag{14}$$

To verify descent directly, let $a_t = K_0 + 3^\rho K_\rho \Delta_t^\rho$. Since $\eta_t a_t \le 2c_\eta$, we have

$$a_t \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 + \|\mathbf{g}_t\| \|\mathbf{w}_{t+1} - \mathbf{w}_t\| = \eta_t \|\mathbf{g}_t\|^2 (1 + a_t \eta_t) \le 8c_\eta (1 + 2c_\eta)\Delta_t \le 2\Delta_t.$$

Lemma 5 with $m = 2\Delta_t$ thus yields

$$\Delta_{t+1} \le \Delta_t - \eta_t \left(1 - \frac{a_t \eta_t}{2} \right) \|\mathbf{g}_t\|^2 \le \Delta_t - \frac{\eta_t}{2} \|\mathbf{g}_t\|^2.$$

In particular, Δ_t is non-increasing. By (14) and the definition of η_t,

$$\min_{0 \le t < T} \min \left\{ \frac{\Delta_t}{K_0}, \frac{\Delta_t^{1-\rho}}{3^\rho K_\rho} \right\} \le (8\sqrt{2} + 8) \frac{D_0^2}{T}.$$

If $\rho \geq 1$, monotonicity gives

$$\min\left\{\frac{\Delta_{T-1}}{K_0}, \frac{\Delta_0^{1-\rho}}{3^\rho K_\rho}\right\} \leq (8\sqrt{2}+8)\frac{D_0^2}{T}.$$

Thus, for $T > (8\sqrt{2}+8)3^\rho D_0^2 K_\rho \Delta_0^{\rho-1}$, we have $\Delta_{T-1} \leq (8\sqrt{2}+8)D_0^2 K_0/T$. If $0 < \rho < 1$, both terms are non-decreasing functions of Δ_t, and hence

$$\min\left\{\frac{\Delta_{T-1}}{K_0}, \frac{\Delta_{T-1}^{1-\rho}}{3^\rho K_\rho}\right\} \leq (8\sqrt{2}+8)\frac{D_0^2}{T}.$$

This gives the claimed bound for $\rho < 1$.

For the constant learning rate $\eta = c_\eta \min\{1/K_0, 1/(3^\rho K_\rho \Delta_0^\rho)\}$, the same descent argument applies inductively: $\Delta_t \leq \Delta_0$ implies $\eta \leq c_\eta \min\{1/K_0, 1/(3^\rho K_\rho \Delta_t^\rho)\}$, which ensures $\Delta_{t+1} \leq \Delta_t$. The bounds (13) and (14) therefore also hold with $\eta_t = \eta$, giving

$$\Delta_{T-1} \leq \frac{D_0^2}{\eta T} \leq (8\sqrt{2}+8)\left(\frac{D_0^2 K_0}{T} + \frac{D_0^2 K_\rho (3\Delta_0)^\rho}{T}\right).$$

This proves all the claims. □

F.3 Proof of Theorem 3

Proof. We consider three cases of $\{\eta_t\}$:

1. $\eta_t \leq \frac{2}{K_1\Delta}, \forall t$.
2. $\eta_t > \frac{2}{K_1\Delta}, \forall t$.
3. $\exists \tau, \eta_t > \frac{2}{K_1\Delta}, t \leq \tau$ and $\eta_t \leq \frac{2}{K_1\Delta}, t > \tau$.

Case 1

We construct the following function:

$$h(x) = \begin{cases} -2\epsilon\left(x + \frac{1}{\sqrt{K_1}}\right) + \frac{5\epsilon}{4\sqrt{K_1}}, & x \in (-\infty, -\frac{1}{\sqrt{K_1}}) \\ \frac{\epsilon}{4}\left(6\sqrt{K_1}x^2 - K_1^{\frac{3}{2}}x^4\right), & x \in \left[-\frac{1}{\sqrt{K_1}}, \frac{1}{\sqrt{K_1}}\right] \\ 2\epsilon\left(x - \frac{1}{\sqrt{K_1}}\right) + \frac{5\epsilon}{4\sqrt{K_1}}, & x \in \left(\frac{1}{\sqrt{K_1}}, +\infty\right) \end{cases} \tag{15}$$

with initial point $y_0 = \frac{1}{\sqrt{K_1}} + \frac{\Delta}{\epsilon}$. We have $h^\star = h(0) = 0$ and $|h''(x)| \leq 3\epsilon\sqrt{K_1}$, so h is $(1, 3\epsilon\sqrt{K_1}, 0)$-smooth. The condition $\epsilon \leq 2\sqrt{K_1}\Delta/5$ gives $h(y_0) = 2\Delta + 5\epsilon/(4\sqrt{K_1}) \leq 5\Delta/2 < 8\Delta$. On the linear tail $y_t \geq 1/\sqrt{K_1}$, the gradient is 2ϵ and each step has length $2\epsilon\eta_t \leq 4\epsilon/(K_1\Delta)$. Therefore, it takes at least

$$\frac{y_0 - \frac{1}{\sqrt{K_1}}}{\frac{4\epsilon}{K_1\Delta}} \geq \frac{K_1\Delta^2}{4\epsilon^2}$$

iterations.

Case 2

Given x_0, we define $x_t = x_0 - \sum_{s=0}^{t-1}\eta_s\sqrt{K_1}\Delta, \forall t \in \mathbb{N}$. We define

$$f(x) = f_t(x), x \in (x_{t+1}, x_t],$$

and

$$f_t(x) = a_t \sin(b_t x + c_t) + d_t, \quad x \in (x_{t+1}, x_t]$$

with

$$b_t = \frac{2\pi}{x_t - x_{t+1}} = \frac{2\pi}{\eta_t\sqrt{K_1}\Delta} \leq \pi\sqrt{K_1}, \quad c_t = \arctan\left(-\frac{\eta_t}{2\pi}K_1\Delta\right) - b_t x_t$$

$$a_t = \frac{\eta_t}{2\pi} K_1 \Delta^2 \sqrt{1 + \frac{\eta_t^2}{4\pi^2} K_1^2 \Delta^2}, \quad d_t = a_t + \Delta - \frac{\alpha_t}{\alpha_t + \sqrt{1+\alpha_t^2}} \Delta,$$

where $\alpha_t = \frac{\sqrt{K_1}}{b_t}$. It is not hard to verify that

$$\begin{aligned} f_t(x_{t+1}) =& f_{t+1}(x_{t+1}) = \Delta, \\ f_t'(x_{t+1}) =& f_{t+1}'(x_{t+1}) = \sqrt{K_1}\Delta, \\ f_t''(x_{t+1}) =& f_{t+1}''(x_{t+1}) = K_1 \Delta. \end{aligned}$$

Then we can link all these f_t together. Moreover, note that

$$|f_t''(x)| = \left| a_t b_t^2 \sin\left(b_t x + c_t\right) \right| \le a_t b_t^2 = K_1 \Delta \sqrt{1 + \frac{1}{\alpha_t^2}} \le 2\pi K_1 \Delta,$$

where in the last inequality we use $\alpha_t = \frac{1}{2\pi} \eta_t K_1 \Delta > \frac{1}{\pi}$. Also note that

$$f_t(x) = a_t \sin\left(b_t x + c_t\right) + d_t \ge d_t - a_t = \Delta - \frac{\alpha_t}{\alpha_t + \sqrt{1+\alpha_t^2}} \Delta > \Delta - \frac{1}{2}\Delta = \frac{1}{2}\Delta,$$

where in the second inequality we use $\frac{\alpha_t}{\alpha_t + \sqrt{1+\alpha_t^2}} < \frac{1}{2}$. If $f^\star = 0$, we immediately obtain that f is $(1, 0, 4\pi K_1)$-smooth. Next, we extend $f(x)$ from x_0 to $+\infty$ to achieve this. Define

$$G(t) = \Delta \left[1 + 2\sqrt{K_1}\left(t - x_0\right) + 2K_1 \left(t - x_0\right)^2 \right] e^{-\sqrt{K_1}(t - x_0)}, t > x_0.$$

It is not hard to verify that $G(x_0) = \Delta$, $G'(x_0) = \sqrt{K_1}\Delta$ and $G''(x_0) = K_1 \Delta$. Moreover, we have $G(t) > 0, t > x_0$ and $G(t) \to 0$ as $t \to \infty$. Therefore, $G^\star = 0$. We compute that

$$G''(t) = \Delta K_1 e^{-\sqrt{K_1}(t - x_0)} \left[1 - 6\sqrt{K_1}\left(t - x_0\right) + 2K_1 \left(t - x_0\right)^2 \right].$$

Therefore, we have

$$|G''(t)| \le \Delta K_1 e^{-\sqrt{K_1}(t - x_0)} \left[1 + 6\sqrt{K_1}\left(t - x_0\right) + 2K_1 \left(t - x_0\right)^2 \right]$$

We immediately obtain that $G(t)$ is $(1, 0, 3K_1)$-smooth. Finally, we define

$$f(x) = \begin{cases} G(x), & x > x_0, \\ f_t(x), & x \in (x_{t+1}, x_t], t \in \mathbb{N} \end{cases}$$

We have that f is $(1, 0, 4\pi K_1)$-smooth, $f^\star = 0$, $f(x_0) - f^\star = \Delta$ and $f'(x_t) = \sqrt{K_1}\Delta, \forall t \in \mathbb{N}$.

Case 3 We use only the accuracy condition $0 < \epsilon \le 2\sqrt{K_1}\Delta/5$ stated in the theorem. Set $q = 1/\sqrt{K_1}$ and $H = 7\Delta$.

We construct the function:

$$f(x) = \begin{cases} h(x), & x \in (-\infty, x_{\tau+1}] \\ g(x), & x \in (x_{\tau+1}, x_\tau] \\ f_t(x), & x \in (x_{t+1}, x_t] \text{ for } 0 \le t \le \tau - 1 \\ F(x), & x > x_0 \end{cases}$$

where h is defined in (15), and g and f_t are defined below. For $x > x_0$, set

$$F(x) = H + \sqrt{K_1}\Delta(x - x_0) + \frac{K_1 \Delta}{2}(x - x_0)^2.$$

Let $x_{t+1} = x_0 - \sum_{s=0}^t \eta_s \sqrt{K_1}\Delta$ for $0 \le t \le \tau$, and choose

$$y_0 = q + \frac{7\Delta - \frac{5\epsilon}{4\sqrt{K_1}}}{2\epsilon} > q, \qquad x_0 = y_0 + \sum_{s=0}^{\tau} \eta_s \sqrt{K_1}\Delta.$$

Then $x_{\tau+1} = y_0$, $h(y_0) = H$, $h'(y_0) = 2\epsilon$, and $h''(y_0) = 0$.

The basic idea of our construction in this case is to let the $(\tau+1)$-th iterate be $x_{\tau+1} = y_0$. Then, the worst-case convergence in Case 1 can be applied. We then want the iterates with large learning rates ($t \le \tau$), making no progress in convergence. We construct trigonometric functions $f_t, t \le \tau - 1$ to achieve this. Finally, we use a piecewise polynomial function g to connect the trigonometric part to h. When $\tau = 0$, the trigonometric part is absent and g connects directly to F.

For $t \le \tau - 1$, we define

$$f_t(x) = a_t \sin(b_t x + c_t) + d_t, \quad x \in (x_{t+1}, x_t] \tag{16}$$

with

$$b_t = \frac{2\pi}{x_t - x_{t+1}} = \frac{2\pi}{\eta_t \sqrt{K_1}\Delta} \le \pi\sqrt{K_1}, \quad c_t = \arctan\left(-\frac{\eta_t}{2\pi} K_1 \Delta\right) - b_t x_t$$
$$a_t = \frac{\eta_t}{2\pi} K_1 \Delta^2 \sqrt{1 + \frac{\eta_t^2}{4\pi^2} K_1^2 \Delta^2}, \quad d_t = a_t + g(x_\tau) - \frac{\alpha_t}{\alpha_t + \sqrt{1+\alpha_t^2}}\Delta,$$

where $\alpha_t = \frac{\sqrt{K_1}}{b_t}$ and $g(x_\tau) = H = 7\Delta$. Note that $f^* = 0$ (achieved when $x = 0$). By Lemma 11, we have that f_t is $(1, 0, K_1)$-smooth. Also, with the above parameter choices, we have

$$\begin{aligned} f_t(x_{t+1}) &= f_t(x_t) = H, \\ f_t'(x_{t+1}) &= f_t'(x_t) = \sqrt{K_1}\Delta, \\ f_t''(x_{t+1}) &= f_t''(x_t) = K_1\Delta. \end{aligned}$$

Then we can link all these $f_t, t \le \tau - 1$ together to achieve $(1, 0, K_1)$-smooth function, as in the boundary x_t and x_{t+1}, the function value and derivatives are the same.

Let $x_L = x_{\tau+1}$ and $x_R = x_\tau$. Define the connector

$$g(x) = \begin{cases} H + 2\epsilon q\, p\left(\dfrac{x - x_L}{q}\right), & x \in [x_L, x_L + q], \\ H, & x \in [x_L + q, x_R - q], \\ H + \Delta r\left(\dfrac{x - (x_R - q)}{q}\right), & x \in [x_R - q, x_R], \end{cases} \tag{17}$$

where

$$p(s) = s(1-s)^3(3s+1), \qquad r(s) = \frac{1}{2}s^3(1-s)(5s-7), \qquad s \in [0,1].$$

The middle interval is nonempty because

$$x_R - x_L = \eta_\tau \sqrt{K_1}\Delta > \frac{2}{\sqrt{K_1}} = 2q.$$

By Lemma 12, g is C^2 and satisfies

$$\begin{aligned} g(x_L) &= H, \quad g'(x_L) = 2\epsilon, \quad g''(x_L) = 0, \\ g(x_R) &= H, \quad g'(x_R) = \sqrt{K_1}\Delta, \quad g''(x_R) = K_1\Delta, \end{aligned}$$

as well as $g(x) > 6\Delta$ and $|g''(x)| \le \epsilon\sqrt{K_1} + 4\pi K_1 g(x)$. Thus the values and first two derivatives agree at every junction. The extension F also matches these derivatives at x_0 and satisfies $F(x) \ge H$ and $|F''(x)| = K_1\Delta \le K_1 F(x)$ for $x > x_0$. Together with $|h''(x)| \le 3\epsilon\sqrt{K_1}$ and Lemma 11, this proves that f is $(1, 3\epsilon\sqrt{K_1}, 4\pi K_1)$-smooth, with $f^\star = f(0) = 0$ and $f(x_0) - f^\star = 7\Delta \le 8\Delta$.

For $t \le \tau$, the gradient at every iterate is $\sqrt{K_1}\Delta > \epsilon$, so GD follows the prescribed points x_t and reaches $x_{\tau+1} = x_L$ in one step from x_τ. After this transition, the iterate lies on the linear tail of h, at distance

$$x_L - q = \frac{7\Delta - \frac{5\epsilon}{4\sqrt{K_1}}}{2\epsilon} \ge \frac{13\Delta}{4\epsilon}$$

from its endpoint. While $x_t \geq q$ and $t > \tau$, the gradient is 2ϵ and $x_t - x_{t+1} = 2\epsilon\eta_t \leq 4\epsilon/(K_1\Delta)$. Consequently, reaching an ϵ-stationary point requires at least

$$\frac{13\Delta/(4\epsilon)}{4\epsilon/(K_1\Delta)} = \frac{13}{16}\frac{K_1\Delta^2}{\epsilon^2}$$

additional iterations. This completes the proof.

□

Lemma 11. *Suppose $g(x_\tau) > 2\pi\Delta + \frac{1}{2}\Delta$. Consider the function defined in (16). We have $f_t(x) > 0$, $|f''(x)| \leq K_1 f(x)$ and $f_t(x_t) = g(x_\tau)$ for all $x \in (x_{t+1}, x_t]$, $t \leq \tau - 1$.*

Proof. By the definition of $\alpha_t := \frac{\sqrt{K_1}}{b_t}$, we have

$$a_t = \Delta\alpha_t\sqrt{1+\alpha_t^2}, \quad b_t = \frac{\sqrt{K_1}}{\alpha_t}.$$

We calculate $f_t''(x)$:

$$|f_t''(x)| = \left|a_t b_t^2 \sin\left(b_t x + c_t\right)\right| \leq a_t b_t^2 = K_1\Delta\sqrt{1+\frac{1}{\alpha_t^2}} \leq 2\pi K_1\Delta,$$

where in the last inequality we use $\alpha_t = \frac{1}{2\pi}\eta_t K_1\Delta > \frac{1}{\pi}$. Then, we calculate $f_t(x)$:

$$f_t(x) = a_t \sin\left(b_t x + c_t\right) + d_t \geq d_t - a_t = g(x_\tau) - \frac{\alpha_t}{\alpha_t + \sqrt{1+\alpha_t^2}}\Delta > g(x_\tau) - \frac{1}{2}\Delta > 2\pi\Delta,$$

where in the second inequality we use $\frac{\alpha_t}{\alpha_t+\sqrt{1+\alpha_t^2}} < \frac{1}{2}$ and the last inequality is due to the condition on $g(x_\tau)$. Then we immediately obtain that $|f''(x)| \leq K_1 f(x)$.

We calculate $f_t(x_t)$:

$$\begin{aligned}
f_t(x_t) &= a_t \sin(c_t) + d_t = -\Delta\alpha_t^2 + d_t \\
&= -\Delta\alpha_t^2 + a_t + g(x_\tau) - \frac{\alpha_t}{\alpha_t + \sqrt{1+\alpha_t^2}}\Delta \\
&= -\Delta\alpha_t^2 + \Delta\alpha_t\sqrt{1+\alpha_t^2} + g(x_\tau) - \frac{\alpha_t}{\alpha_t + \sqrt{1+\alpha_t^2}}\Delta \\
&= g(x_\tau).
\end{aligned}$$

□

Lemma 12. *Suppose $0 < \epsilon \leq 2\sqrt{K_1}\Delta/5$. The function g in (17) is C^2 on $[x_L, x_R]$ and satisfies*

$$\begin{aligned}
g(x_L) = H, \quad g'(x_L) = 2\epsilon, \quad g''(x_L) = 0, \\
g(x_R) = H, \quad g'(x_R) = \sqrt{K_1}\Delta, \quad g''(x_R) = K_1\Delta.
\end{aligned}$$

Moreover, $g(x) > 6\Delta$ and $|g''(x)| \leq \epsilon\sqrt{K_1} + 4\pi K_1 g(x)$ on this interval.

Proof. Differentiation gives

$$\begin{aligned}
p'(s) &= 1 - 18s^2 + 32s^3 - 15s^4, & p''(s) &= -36s + 96s^2 - 60s^3, \\
r'(s) &= -\frac{21}{2}s^2 + 24s^3 - \frac{25}{2}s^4, & r''(s) &= -21s + 72s^2 - 50s^3.
\end{aligned}$$

Their endpoint values are

$$\begin{aligned}
p(0) = 0, \quad p'(0) = 1, \quad p''(0) = 0, \quad p(1) = p'(1) = p''(1) = 0, \\
r(0) = r'(0) = r''(0) = 0, \quad r(1) = 0, \quad r'(1) = r''(1) = 1.
\end{aligned}$$

These identities establish the endpoint conditions and show that g and its first two derivatives match the constant middle piece at both internal junctions.

Since $p'(s) = (1-s)^2(1-3s)(1+5s)$, we have $0 \le p(s) \le p(1/3) = 16/81$ on $[0,1]$. Also, $0 \le -r(s) \le p(1-s)$, so $-16/81 \le r(s) \le 0$. Evaluating p'' at its critical points $(8 \pm \sqrt{19})/15$ and the endpoints, and r'' at $(24 \pm \sqrt{226})/50$ and the endpoints, gives

$$\max_{s\in[0,1]} |p''(s)| < 4, \qquad \max_{s\in[0,1]} |r''(s)| < 4.$$

On the left transition $[x_L, x_L + q]$, it follows that

$$g(x) \ge H = 7\Delta, \qquad |g''(x)| < \frac{8\epsilon}{q} = 8\epsilon\sqrt{K_1} \le \epsilon\sqrt{K_1} + 4\pi K_1 g(x),$$

where the last inequality uses $\epsilon \le 2\sqrt{K_1}\Delta/5$. On the right transition $[x_R - q, x_R]$,

$$g(x) \ge \left(7 - \frac{16}{81}\right)\Delta > 6\Delta, \qquad |g''(x)| < \frac{4\Delta}{q^2} = 4K_1\Delta \le 4\pi K_1 g(x).$$

The middle interval has value H and zero curvature. This proves all the claims. □

F.4 Explanation of Example 4

Proof. First note that f is C^2, $f^\star = f(0,0) = 0$, and

$$\left\|\nabla^2 f(x,y)\right\| = K_1 \max\left\{1, \mathrm{e}^{\sqrt{K_1}|x|-1}\right\} \le K_1 + K_1 f(x,y).$$

Thus f is $(1, K_1, K_1)$-smooth.

Write $R > \mathrm{e}$ for the parameter denoted by Δ_0 in the initialization of Example 4, so that $x_0 = \log R/\sqrt{K_1}$. Fix $K_1 > 0$, $y_0 > 1$, and an accuracy $0 < \epsilon < \min\{1/2, B\}$, where $B = g(y_0) = K_1 y_0^2/2$. The actual initial gap and derivative are

$$\Delta_0 = f(x_0, y_0) - f^\star = \frac{R}{\mathrm{e}} - \frac{1}{2} + B, \qquad h'(x_0) = \frac{\sqrt{K_1}R}{\mathrm{e}}.$$

In particular, $\Delta_0 = \Theta(R)$ as $R \to \infty$. Below, let $z_t = \sqrt{K_1}x_t$ and $\Delta_t = f(x_t, y_t) - f^\star$, and compare the iteration counts needed to reach $\Delta_t \le \epsilon$.

Consider first a constant learning rate $\eta > 0$, and let $a = \eta K_1$. If $a \ge 2z_0\mathrm{e}^{1-z_0}$, then $z_1 = z_0 - a\mathrm{e}^{z_0-1} \le -z_0$. Since e^{r-1}/r is non-decreasing for $r \ge 1$, induction gives

$$|z_{t+1}| = a\mathrm{e}^{|z_t|-1} - |z_t| \ge |z_t| \ge z_0.$$

Thus the target accuracy cannot be reached, and any successful constant learning rate must satisfy

$$a < 2z_0\mathrm{e}^{1-z_0} = \frac{2\mathrm{e}\log R}{R}.$$

For sufficiently large R, this implies $0 < a \le 1/2$. Along the y-axis, we have

$$y_{t+1} = y_t - \eta K_1 y_t = (1 - \eta K_1)y_t = (1 - \eta K_1)^{t+1} y_0. \tag{18}$$

Since $\Delta_t \ge g(y_t) = B(1-a)^{2t}$, reaching $\Delta_t \le \epsilon$ requires

$$t \ge \frac{\log(B/\epsilon)}{-2\log(1-a)} \ge \frac{\log(B/\epsilon)}{4a} = \Omega\left(\frac{R}{\log R}\right).$$

Conversely, taking $a = z_0\mathrm{e}^{1-z_0} = \mathrm{e}\log R/R$ gives $z_1 = 0$, after which $z_t = 0$. Equation (18) then gives a matching $\mathcal{O}(R/\log R)$ upper bound. Hence the smallest iteration count over all positive constant learning rates is

$$T^{\mathrm{opt}}_{\mathrm{const}} = \Theta\left(\frac{R}{\log R}\right).$$

Next, use the adaptive learning rate in Theorem 1, namely

$$\eta_t = \frac{c_\eta}{K_1} \min\left\{1, \frac{1}{3\Delta_t}\right\}, \qquad c_\eta = \frac{1}{4\sqrt{2}+4}.$$

Since $0 < \eta_t K_1 \le c_\eta < 1$, we have $0 < y_t \le y_0$. While $z_t > 1$, we have $\Delta_t = \mathrm{e}^{z_t - 1} - 1/2 + g(y_t) > 1/2$, and therefore

$$\frac{c_\eta}{3(1+B)} \le z_t - z_{t+1} = \frac{c_\eta \mathrm{e}^{z_t - 1}}{3\Delta_t} \le \frac{2c_\eta}{3} < 1.$$

It follows that the first index τ with $z_\tau \le 1$ satisfies $\tau = \Theta(\log R)$ and $0 < z_\tau \le 1$. For $t \ge \tau$, the iterates remain in the quadratic region, where $\Delta_t \le 1/2 + B$ and

$$a_* := c_\eta \min\left\{1, \frac{1}{3(1/2+B)}\right\} \le \eta_t K_1 \le c_\eta < 1.$$

Both coordinates contract by the same factor, so

$$\Delta_{t+1} = (1 - \eta_t K_1)^2 \Delta_t \le (1 - a_*)^2 \Delta_t.$$

Thus reaching accuracy ϵ takes at most $\mathcal{O}(\log(1/\epsilon))$ further iterations. Since $\Delta_t > 1/2 > \epsilon$ for $t < \tau$, the adaptive schedule has $T_{\mathrm{ad}} = \Theta(\log R)$ for the fixed accuracy considered here. Combining the two iteration counts gives

$$\frac{T^{\mathrm{opt}}_{\mathrm{const}}}{T_{\mathrm{ad}}} = \Theta\left(\frac{R}{(\log R)^2}\right) = \tilde{\Theta}(\Delta_0).$$

This proves the claimed acceleration. □

G Unified SGD Analysis and Proofs for Section 4.1

We present the stochastic framework used by Theorems 4–7. The distinction between bounded noise and the almost-sure ABC condition is the deterministic envelope used for $\|\mathbf{n}_t\|$.

Let

$$\Delta_t := f(\mathbf{w}_t) - f^\star, \quad \mathbf{v}_t := \nabla f(\mathbf{w}_t), \quad \mathbf{n}_t := \mathbf{g}_t - \mathbf{v}_t,$$

and let $\mathcal{F}_t = \sigma(\mathbf{w}_0, \xi_0, \ldots, \xi_{t-1})$. Recall the noise assumption in Section 4 gives

$$\mathbb{E}[\mathbf{n}_t \mid \mathcal{F}_t] = 0, \quad \|\mathbf{n}_t\|^2 \le A\Delta_t + B\|\mathbf{v}_t\|^2 + \sigma^2 \quad \text{almost surely}, \tag{19}$$

where $A = B = 0$ recovers Assumption 3. We set

$$\ell_\delta := \log\frac{\mathrm{e}}{\delta}, \quad D := 3\Delta_0.$$

If $\Delta_0 = 0$, the stationarity conclusion already holds at $t = 0$. For $u \ge 0$, define

$$\begin{aligned} r(u) &:= \min\{C_1 u^{-(\rho-1)/2}, C_2\}, & L(u) &:= 2K_0 + K_\rho (2u)^\rho, \\ H(u) &:= K_0 + 3^\rho K_\rho u^\rho, & G(u) &:= 2\sqrt{uH(u)}, \\ N_{A,B}(u) &:= \sqrt{Au + BG(u)^2 + \sigma^2}. \end{aligned} \tag{20}$$

Whenever $\Delta_t \le u$, Lemma 6 and (19) imply $\|\mathbf{v}_t\| \le G(u)$ and $\|\mathbf{n}_t\| \le N_{A,B}(u)$ almost surely.

We will use the following learning rate rule

$$h_{A,B}(u) := \min\left\{\frac{r(u)}{G(u) + N_{A,B}(u)}, \frac{1}{2L(u)}, \sqrt{\frac{\Delta_0}{TL(u)N_{A,B}(u)^2}}, \frac{\Delta_0}{G(u)N_{A,B}(u)\sqrt{2T\ell_\delta}}\right\}. \tag{21}$$

The four entries respectively keep one update inside the local smoothness radius, preserve a descent coefficient, budget the full quadratic noise term, and provide a deterministic Azuma–Hoeffding increment envelope. The bounded-noise specializations used in Theorems 4 and 5 are

$$\eta_t = h_{0,0}(\Delta_t) = \min\left\{\frac{r(\Delta_t)}{G(\Delta_t) + \sigma}, \frac{1}{2L(\Delta_t)}, \sqrt{\frac{\Delta_0}{TL(\Delta_t)\sigma^2}}, \frac{\Delta_0}{\sigma G(\Delta_t)\sqrt{2T\ell_\delta}}\right\}, \tag{22}$$

$$\eta = h_{0,0}(D) = \min\left\{\frac{r(D)}{G(D)+\sigma}, \frac{1}{2L(D)}, \sqrt{\frac{\Delta_0}{TL(D)\sigma^2}}, \frac{\Delta_0}{\sigma G(D)\sqrt{2T\ell_\delta}}\right\}. \tag{23}$$

For the rate calculation, write

$$\begin{aligned} \Delta_{\mathrm{avg},p} &:= \frac{1}{T}\sum_{t=0}^{T-1}\Delta_t^p, \\ H_{\mathrm{ad}} &:= K_0 + K_\rho \Delta_{\mathrm{avg},\rho}, \quad R_{\mathrm{ad}} := C_1^{-1}\Delta_{\mathrm{avg},(\rho-1)/2} + C_2^{-1}, \\ H_{\mathrm{c}} &:= K_0 + K_\rho \Delta_0^\rho, \qquad R_{\mathrm{c}} := C_1^{-1}\Delta_0^{(\rho-1)/2} + C_2^{-1}. \end{aligned} \tag{24}$$

We use $\Delta_{\mathrm{avg},0} = 1$. The notation $\mathcal{O}_\rho$ in Section 4 hides only numerical factors depending on ρ.

G.1 A STOPPED DESCENT ARGUMENT

Lemma 13 (Azuma–Hoeffding inequality)**.** *Let $(M_k, \mathcal{F}_k)_{k=0}^T$ be a real-valued martingale. Suppose that deterministic numbers $b_0, \dots, b_{T-1} \geq 0$ satisfy $|M_{t+1} - M_t| \leq b_t$ almost surely for $0 \leq t < T$. If $\sum_{t=0}^{T-1} b_t^2 > 0$, then, for every $a > 0$,*

$$\mathbb{P}(M_T - M_0 \geq a) \leq \exp\left(-\frac{a^2}{2\sum_{t=0}^{T-1} b_t^2}\right).$$

Equivalently,

$$\mathbb{P}\left(M_T - M_0 \leq \sqrt{2\sum_{t=1}^{n} b_t^2 \log(1/\delta)}\right) \geq 1 - \delta.$$

Lemma 14. *Suppose (19) holds and $\rho \geq 1$. At an index t for which $\Delta_t \leq u_t$, set*

$$\widehat{r}_t = r(u_t), \quad \widehat{L}_t = L(u_t), \quad \widehat{G}_t = G(u_t), \quad \widehat{N}_t = N_{A,B}(u_t).$$

If $0 < \eta_t \leq h_{A,B}(u_t)$, then almost surely

$$\begin{aligned} \eta_t \|\mathbf{g}_t\| &\leq \widehat{r}_t, & \widehat{L}_t\eta_t &\leq \frac{1}{2}, \\ \widehat{L}_t\eta_t^2 \|\mathbf{n}_t\|^2 &\leq \frac{\Delta_0}{T}, & \eta_t \|\mathbf{v}_t\| \|\mathbf{n}_t\| &\leq \frac{\Delta_0}{\sqrt{2T\ell_\delta}}. \end{aligned} \tag{25}$$

Proof. Lemma 6 and (19) imply $\|\mathbf{v}_t\| \leq \widehat{G}_t$ and $\|\mathbf{n}_t\| \leq \widehat{N}_t$, since G and $N_{A,B}$ are non-decreasing. Then we have that $\|\mathbf{g}_t\| \leq \widehat{G}_t + \widehat{N}_t$. Each inequality in (25) now follows from the corresponding entry of (21). □

Lemma 15. *Suppose Assumptions 1,2 and (19) hold with $\rho \geq 1$. Let either $u_t = \Delta_t$ or $u_t = D$, $D = 3\Delta_0$ for every t, and choose learning rates satisfying $0 < \eta_t \leq h_{A,B}(u_t)$. Then, with probability at least $1 - \delta$, it holds that*

$$\Delta_t \leq 3\Delta_0 \quad (0 \leq t \leq T), \tag{26}$$

and

$$\min_{0\leq t<T} \|\nabla f(\mathbf{w}_t)\|^2 \leq \frac{6\Delta_0}{\sum_{t=0}^{T-1}\eta_t}. \tag{27}$$

Proof. Define the exit time and stopped horizon by

$$\tau := \inf\{t \geq 0 : \Delta_t > D\}, \quad s := T \wedge \tau. \tag{28}$$

Since Δ_t is $\mathcal{F}_t$-measurable, τ is a stopping time and $\mathbf{1}_{\{t<\tau\}}$ is $\mathcal{F}_t$-measurable. For $t < \tau$, we have $\Delta_t \leq u_t$ for either choice of u_t. Lemma 14 therefore applies before the exit. In the constant case, $r(D) \leq r(\Delta_t)$ and $L(\Delta_t) \leq L(D)$.

For $t < \tau$, by Lemma 14, we have that $\widehat{L}_t \eta_t \le 1/2$ and $\eta_t \|\mathbf{g}_t\| \le \hat{r}_t$. Then by Lemma 2, we have that for $t < \tau$,

$$\begin{aligned}\Delta_{t+1} &\le \Delta_t - \eta_t \|\mathbf{v}_t\|^2 - \eta_t \langle \mathbf{v}_t, \mathbf{n}_t \rangle + \frac{\widehat{L}_t \eta_t^2}{2} \|\mathbf{v}_t + \mathbf{n}_t\|^2 \\ &\le \Delta_t - \frac{\eta_t}{2} \|\mathbf{v}_t\|^2 - \eta_t \langle \mathbf{v}_t, \mathbf{n}_t \rangle + \widehat{L}_t \eta_t^2 \|\mathbf{n}_t\|^2 ,\end{aligned} \tag{29}$$

where we used $\|\mathbf{v}_t + \mathbf{n}_t\|^2 \le 2\|\mathbf{v}_t\|^2 + 2\|\mathbf{n}_t\|^2$. Summing (29) from $t = 0$ to $s - 1$ gives the exact stopped telescoping inequality

$$\Delta_s + \frac{1}{2} \sum_{t=0}^{s-1} \eta_t \|\mathbf{v}_t\|^2 \le \Delta_0 + M_T + Q_T, \tag{30}$$

where we define

$$Y_t := -\eta_t \langle \mathbf{v}_t, \mathbf{n}_t \rangle \mathbf{1}_{\{t<\tau\}}, \qquad M_k := \sum_{t=0}^{k-1} Y_t \quad (0 \le k \le T),$$

and the full stopped quadratic noise sum

$$Q_T := \sum_{t=0}^{T-1} \widehat{L}_t \eta_t^2 \|\mathbf{n}_t\|^2 \mathbf{1}_{\{t<\tau\}}.$$

By Lemma 14, $\widehat{L}_t \eta_t^2 \|\mathbf{n}_t\|^2 \le \frac{\Delta_0}{T}$ and therefore $Q_T \le \Delta_0$. The multiplier $\eta_t \mathbf{v}_t \mathbf{1}_{\{t<\tau\}}$ is measurable before the current sample is drawn, and

$$\mathbb{E}[Y_t \mid \mathcal{F}_t] = -\eta_t \mathbf{1}_{\{t<\tau\}} \langle \mathbf{v}_t, \mathbb{E}[\mathbf{n}_t \mid \mathcal{F}_t] \rangle = 0.$$

For $t < \tau$, Lemma 14 gives that $|Y_t| \le \frac{\Delta_0}{\sqrt{2T\ell_\delta}}$; for $t \ge \tau$, $Y_t = 0$. Thus

$$|Y_t| \le b := \frac{\Delta_0}{\sqrt{2T\ell_\delta}}.$$

Consequently $(M_k, \mathcal{F}_k)$ is an integrable martingale with deterministic increment envelope b. Lemma 13 implies

$$\mathbb{P}(M_T > \Delta_0) \le \exp(-\ell_\delta) = \delta/\mathrm{e} \le \delta.$$

On the event $\mathcal{E} := \{M_T \le \Delta_0\}$, whose probability is at least $1 - \delta$, we obtain that

$$\Delta_s \le \Delta_s + \frac{1}{2} \sum_{t=0}^{s-1} \eta_t \|\mathbf{v}_t\|^2 \le 3\Delta_0.$$

If $\tau \le T$, this would give $\Delta_\tau \le 3\Delta_0 = D$, contradicting the definition of τ. Hence $\tau > T$ on $\mathcal{E}$, and $\Delta_t \le D$ for every $0 \le t \le T$. Dividing the weighted gradient sum by $\sum_{t=0}^{T-1} \eta_t$ we get the desired result. □

G.2 A COMMON STEP-SIZE COMPLEXITY BOUND

Lemma 16. *Let $0 < u_t \le D$ for $0 \le t < T$ and $\eta_t = h_{A,B}(u_t)$. Suppose $K_0 + K_\rho > 0$, and define*

$$\overline{H} := \frac{1}{T} \sum_{t=0}^{T-1} H(u_t), \qquad \overline{R} := \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{r(u_t)}. \tag{31}$$

Then the following inequality holds:

$$\frac{\Delta_0}{\sum_{t=0}^{T-1} \eta_t} \le \mathcal{O}\left[\frac{\Delta_0}{T} (\overline{H} + \sigma \overline{R}) + \sqrt{\frac{\Delta_0 \ell_\delta}{T} \left((\sigma^2 + A\Delta_0)\overline{H} + B\Delta_0 \overline{H}^2 \right)} \right]. \tag{32}$$

Proof. Cauchy–Schwarz yields

$$\frac{1}{\sum_{t=0}^{T-1}\eta_t} \le \frac{1}{T^2}\sum_{t=0}^{T-1}\frac{1}{\eta_t}. \tag{33}$$

By the definition of η_t in (21), we have that

$$\begin{aligned}\frac{1}{\eta_t} \le& \frac{G(u_t)+N_{A,B}(u_t)}{r(u_t)} + 2L(u_t)\\ &+\sqrt{\frac{T}{\Delta_0}}\sqrt{L(u_t)}N_{A,B}(u_t) + \frac{\sqrt{2T\ell_\delta}}{\Delta_0}G(u_t)N_{A,B}(u_t).\end{aligned} \tag{34}$$

We bound these four contributions before simplifying the result.

Set $c_r := 2+\sqrt{2}$. The definitions of the local radius and envelopes give, for $u>0$,

$$\frac{\sqrt{u}}{r(u)} \le c_r\sqrt{H(u)}, \qquad \frac{G(u)}{r(u)} \le 2c_rH(u), \qquad L(u)\le 2H(u). \tag{35}$$

To verify the first inequality for positive smoothness parameters, let $u_\star = (K_0/K_\rho)^{1/\rho}/3$. For $u \ge u_\star$, the radius equals $C_1u^{-(\rho-1)/2}$, and $\sqrt{u}/r(u) = c_r\sqrt{3^\rho K_\rho u^\rho} \le c_r\sqrt{H(u)}$. For $u \le u_\star$, the radius equals C_2 and $\sqrt{u}/C_2 \le \sqrt{u_\star}/C_2 = c_r\sqrt{K_0}$. The other two inequalities follow from $G(u) = 2\sqrt{uH(u)}$ and the definition of $L(u)$.

Using $N_{A,B}(u) \le \sqrt{A}\sqrt{u} + \sqrt{B}G(u) + \sigma$ and (35), we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{G(u_t)+N_{A,B}(u_t)}{r(u_t)} + 2L(u_t)\right] \le C\left[(1+\sqrt{B})\overline{H} + \sqrt{A}\sqrt{\overline{H}} + \sigma\overline{R}\right] \tag{36}$$

for an absolute constant C, where $T^{-1}\sum_t\sqrt{H(u_t)} \le \sqrt{\overline{H}}$.

Next, since $u \le D$ and $G(u) = 2\sqrt{uH(u)}$,

$$\sqrt{L(u)}N_{A,B}(u) \le \sqrt{2}(\sigma+\sqrt{AD})\sqrt{H(u)} + 2\sqrt{2BD}\,H(u), \tag{37}$$

$$G(u)N_{A,B}(u) \le 2\sqrt{D}(\sigma+\sqrt{AD})\sqrt{H(u)} + 4\sqrt{B}\,DH(u). \tag{38}$$

Evaluating (37) and (38) at $u=u_t$, averaging over $t=0,\dots,T-1$, and using $\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{H(u_t)} \le \left(\frac{1}{T}\sum_{t=0}^{T-1}H(u_t)\right)^{1/2} = \sqrt{\overline{H}}$, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{L(u_t)}N_{A,B}(u_t) \le \sqrt{2}(\sigma+\sqrt{AD})\sqrt{\overline{H}} + 2\sqrt{2BD}\,\overline{H},$$

$$\frac{1}{T}\sum_{t=0}^{T-1}G(u_t)N_{A,B}(u_t) \le 2\sqrt{D}(\sigma+\sqrt{AD})\sqrt{\overline{H}} + 4\sqrt{B}\,D\overline{H}.$$

Next, multiplying (33) by Δ_0 and substituting the bound on $1/\eta_t$ from (34) gives

$$\begin{aligned}\frac{\Delta_0}{\sum_{t=0}^{T-1}\eta_t} &\le \frac{\Delta_0}{T}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\eta_t}\right)\\ &\le \frac{\Delta_0}{T}\left\{\frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{G(u_t)+N_{A,B}(u_t)}{r(u_t)} + 2L(u_t)\right]\right\}\\ &\quad+\sqrt{\frac{\Delta_0}{T}}\left\{\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{L(u_t)}N_{A,B}(u_t)\right\}\\ &\quad+\sqrt{\frac{2\ell_\delta}{T}}\left\{\frac{1}{T}\sum_{t=0}^{T-1}G(u_t)N_{A,B}(u_t)\right\}.\end{aligned}$$

We bound the first average by (36) and the remaining two averages by the inequalities above. Substituting $D = 3\Delta_0$ and collecting the numerical factors then yields

$$\begin{aligned}\frac{\Delta_0}{\sum_{t=0}^{T-1}\eta_t} \le \frac{C\Delta_0}{T}\left[(1+\sqrt{B})\overline{H} + \sqrt{A}\sqrt{\overline{H}} + \sigma\overline{R}\right] \\ + \sqrt{\frac{\Delta_0}{T}}\left[\sqrt{2}(\sigma + \sqrt{3A\Delta_0})\sqrt{\overline{H}} + 2\sqrt{6B\Delta_0}\,\overline{H}\right] \\ + \sqrt{\frac{\Delta_0\ell_\delta}{T}}\left[2\sqrt{6}(\sigma + \sqrt{3A\Delta_0})\sqrt{\overline{H}} + 12\sqrt{2B\Delta_0}\,\overline{H}\right],\end{aligned}$$

where C is the absolute constant in (36).

To simplify the first term, separate the local contribution into

$$\frac{\Delta_0}{T}\left[(1+\sqrt{B})\overline{H} + \sqrt{A}\sqrt{\overline{H}} + \sigma\overline{R}\right] = \frac{\Delta_0}{T}(\overline{H} + \sigma\overline{R}) + \frac{\sqrt{A}\,\Delta_0}{T}\sqrt{\overline{H}} + \frac{\sqrt{B}\,\Delta_0}{T}\overline{H}.$$

The first term already has the form required in (32).

Since $T \ge 1$ and $\ell_\delta \ge 1$, we have $1/\sqrt{T\ell_\delta} \le 1$. Consequently,

$$\frac{\sqrt{A}\,\Delta_0}{T}\sqrt{\overline{H}} = \frac{1}{\sqrt{T\ell_\delta}}\sqrt{\frac{\Delta_0\ell_\delta}{T}}\,\sqrt{A\Delta_0}\sqrt{\overline{H}} \le \sqrt{\frac{\Delta_0\ell_\delta}{T}}\,\sqrt{A\Delta_0}\sqrt{\overline{H}},$$

$$\frac{\sqrt{B}\,\Delta_0}{T}\overline{H} = \frac{1}{\sqrt{T\ell_\delta}}\sqrt{\frac{\Delta_0\ell_\delta}{T}}\,\sqrt{B\Delta_0}\,\overline{H} \le \sqrt{\frac{\Delta_0\ell_\delta}{T}}\,\sqrt{B\Delta_0}\,\overline{H}.$$

Thus, each additional local term is bounded by its corresponding component in the stochastic term. In particular, their sum satisfies

$$\frac{\sqrt{A}\,\Delta_0}{T}\sqrt{\overline{H}} + \frac{\sqrt{B}\,\Delta_0}{T}\overline{H} \le \sqrt{\frac{\Delta_0\ell_\delta}{T}}\left[\sqrt{A\Delta_0}\sqrt{\overline{H}} + \sqrt{B\Delta_0}\,\overline{H}\right].$$

We obtain

$$\frac{\Delta_0}{\sum_{t=0}^{T-1}\eta_t} \le \mathcal{O}\left[\frac{\Delta_0}{T}(\overline{H} + \sigma\overline{R}) + \sqrt{\frac{\Delta_0\ell_\delta}{T}}\left(\sigma\sqrt{\overline{H}} + \sqrt{A\Delta_0}\sqrt{\overline{H}} + \sqrt{B\Delta_0}\,\overline{H}\right)\right].$$

It remains to combine the three nonnegative quantities inside the last parentheses into a single square root. By Cauchy–Schwarz,

$$\left(\sigma\sqrt{\overline{H}} + \sqrt{A\Delta_0}\sqrt{\overline{H}} + \sqrt{B\Delta_0}\,\overline{H}\right)^2 \le 3\left((\sigma^2 + A\Delta_0)\overline{H} + B\Delta_0\overline{H}^2\right).$$

Taking square roots gives $\sigma\sqrt{\overline{H}} + \sqrt{A\Delta_0}\sqrt{\overline{H}} + \sqrt{B\Delta_0}\,\overline{H} \le \sqrt{3}\sqrt{(\sigma^2 + A\Delta_0)\overline{H} + B\Delta_0\overline{H}^2}$. Substituting this estimate into the preceding bound and absorbing the numerical factor $\sqrt{3}$ into the implicit constant yields

$$\frac{\Delta_0}{\sum_{t=0}^{T-1}\eta_t} \le \mathcal{O}\left[\frac{\Delta_0}{T}(\overline{H} + \sigma\overline{R}) + \sqrt{\frac{\Delta_0\ell_\delta}{T}\left((\sigma^2 + A\Delta_0)\overline{H} + B\Delta_0\overline{H}^2\right)}\right].$$

This proves (32).

□

For the applications to the theorems, the averages in (31) satisfy

$$\begin{aligned} u_t = \Delta_t: &\quad \overline{H} = K_0 + 3^\rho K_\rho \Delta_{\mathrm{avg},\rho} \le 3^\rho H_{\mathrm{ad}}, &\quad \overline{R} \le R_{\mathrm{ad}}, \\ u_t = D: &\quad \overline{H} \le K_0 + 12^\rho K_\rho \Delta_0^\rho \le 12^\rho H_{\mathrm{c}}, &\quad \overline{R} \le 4^{(\rho-1)/2} R_{\mathrm{c}}. \end{aligned} \tag{39}$$

The radius bounds use $1/r(u) = \max\{C_1^{-1}u^{(\rho-1)/2}, C_2^{-1}\} \le C_1^{-1}u^{(\rho-1)/2} + C_2^{-1}$. The factors in (39) are precisely the factors absorbed by $\mathcal{O}_\rho$ in the main text.

G.3 Proof of Theorem 4

Proof. Assumption 3 gives (19) with $A = B = 0$ and $N_{0,0}(u) = \sigma$. Apply Lemma 15 with $u_t = \Delta_t$ and $\eta_t = h_{0,0}(\Delta_t)$. On its event of probability at least $1 - \delta$, the entire trajectory through T satisfies $\Delta_t \le D$ and the stationarity bound (27) holds. Lemma 16, with $A = B = 0$, together with the adaptive specialization in (39), gives (2). □

G.4 Proof of Theorem 5

Proof. Take $A = B = 0$, but now use $u_t = D$ and $\eta_t \equiv \eta = h_{0,0}(D)$. Lemma 15 gives $\Delta_t \le D$ for $0 \le t \le T$ and

$$\min_{0 \le t < T} \|\nabla f(\mathbf{w}_t)\|^2 \le \frac{6\Delta_0}{T\eta}$$

on one event of probability at least $1 - \delta$. Lemma 16 applies with the deterministic positive inputs $u_t = D$ and yields (3) through the constant specialization in (39). □

H Proofs for Section 4.2

The almost-sure ABC results are instances of the same analysis in Appendix G. We use the same notations as in Appendix G. In particular, we use $D = 3\Delta_0$.

For Theorem 6, the adaptive specialization of (21) is

$$\begin{aligned} \eta_t &= h_{A,B}(\Delta_t) \\ &= \min\left\{ \frac{r(\Delta_t)}{G(\Delta_t) + N_{A,B}(\Delta_t)}, \frac{1}{2L(\Delta_t)}, \sqrt{\frac{\Delta_0}{TL(\Delta_t)N_{A,B}(\Delta_t)^2}}, \frac{\Delta_0}{G(\Delta_t)N_{A,B}(\Delta_t)\sqrt{2T\ell_\delta}} \right\}. \end{aligned} \tag{40}$$

For Theorem 7, the constant specialization is

$$\begin{aligned} \eta &= h_{A,B}(D) \\ &= \min\left\{ \frac{r(D)}{G(D) + N_{A,B}(D)}, \frac{1}{2L(D)}, \sqrt{\frac{\Delta_0}{TL(D)N_{A,B}(D)^2}}, \frac{\Delta_0}{G(D)N_{A,B}(D)\sqrt{2T\ell_\delta}} \right\}. \end{aligned} \tag{41}$$

These two expressions have exactly the same four components as the bounded-noise rule, and both reduce to that rule when $A = B = 0$.

Budgeting the full ABC noise. For either choice $u_t = \Delta_t$ or $u_t = D$, the third step-size entry gives, before the exit time in (28),

$$\begin{aligned} \widehat{L}_t \eta_t^2 \|\mathbf{n}_t\|^2 &\le \widehat{L}_t \eta_t^2 \left(Au_t + BG(u_t)^2 + \sigma^2\right) \\ &= \widehat{L}_t \eta_t^2 N_{A,B}(u_t)^2 \le \frac{\Delta_0}{T}. \end{aligned} \tag{42}$$

Thus the entire quadratic noise sum is controlled by $Q_T \le \Delta_0$. The remaining analysis for the ABC-noise follows similarly.

H.1 Proof of Theorem 6

Proof. By Assumption 4 (19) holds. The learning rate in (40) is precisely $h_{A,B}(u_t)$ with $u_t = \Delta_t$. Apply Lemma 15. Its single event of probability at least $1 - \delta$ gives $\Delta_t \le 3\Delta_0$ for all $0 \le t \le T$ and

$$\min_{0 \le t < T} \|\nabla f(\mathbf{w}_t)\|^2 \le \frac{6\Delta_0}{\sum_{t=0}^{T-1} \eta_t}.$$

If $\Delta_t = 0$ at some $t < T$, the stationarity claim is already immediate. Otherwise use Lemma 16 on this event, followed by the adaptive specialization in (39). This gives the complete bound (4). □

H.2 Proof of Theorem 7

Proof. Now use $u_t = D$ and the constant step size $\eta = h_{A,B}(D)$ in (41). By Lemma 15, with probability at least $1-\delta$, we have that

$$\Delta_t \le 3\Delta_0 \quad (0 \le t \le T), \qquad \min_{0\le t<T} \|\nabla f(\mathbf{w}_t)\|^2 \le \frac{6\Delta_0}{T\eta}.$$

Lemma 16 with $u_t = D$ and the constant specialization in (39) give (5). □

I Extension to Sub-gaussian Noise

We extend Theorems 4–7 to sub-Gaussian noise for both warmup schedules and constant schedules. An exponential supermartingale controls the linear noise, while a separate norm event ensures local descent.

Assumption 5. *There exist constants $A, B \ge 0$ and $C > 0$ such that, for every $\mathbf{x} \in \mathbb{R}^d$, $\mathbb{E}_\xi[\nabla f(\mathbf{x}, \xi)] = \nabla f(\mathbf{x})$ and*

$$\mathbb{E}_\xi \left[\exp\left(\frac{\|\nabla f(\mathbf{x},\xi) - \nabla f(\mathbf{x})\|^2}{A\big(f(\mathbf{x}) - f^\star\big) + B\,\|\nabla f(\mathbf{x})\|^2 + C} \right) \right] \le \mathrm{e}.$$

Throughout this appendix, suppose Assumptions 1 and 2 hold with $\rho \ge 1$. Write $\mathbf{v}_t = \nabla f(\mathbf{w}_t)$, $\mathbf{n}_t = \mathbf{g}_t - \mathbf{v}_t$. Define

$$D = 3\Delta_0, \qquad \ell = \log\frac{2}{\delta}, \qquad \kappa = \log\frac{2\mathrm{e}T}{\delta}; \qquad 1 \le \kappa, \quad \ell \le \kappa.$$

Retain r, L, H, G from (20) and set

$$S(u) := Au + BG(u)^2 + C, \qquad S_t := S(\Delta_t).$$

Thus $\|\mathbf{v}_t\| \le G(\Delta_t)$, $\mathbb{E}_t[\mathbf{n}_t] = 0$, and $\mathbb{E}_t[\exp(\|\mathbf{n}_t\|^2/S_t)] \le \mathrm{e}$.

The two schedules use the common rule

$$h^{\mathrm{sg}}_{A,B}(u) := \min\left\{ \frac{r(u)}{G(u) + \sqrt{\kappa S(u)}}, \quad \frac{1}{2L(u)}, \sqrt{\frac{\Delta_0}{T\kappa L(u) S(u)}}, \quad \frac{\Delta_0}{G(u)\sqrt{8TS(u)\ell}} \right\}. \tag{43}$$

The dependence on C, T, δ, Δ_0 is suppressed.

I.1 Shared concentration and complexity estimates

Lemma 17. *Under the assumptions above and Assumption 5, let SGD use positive, $\mathcal{F}_t$-measurable learning rates with $\eta_t \le h^{\mathrm{sg}}_{A,B}(z_t)$, where either $z_t = \Delta_t$ for all t or $z_t = D$ for all t. With probability at least $1-\delta$, simultaneously,*

$$\Delta_t \le D \quad (0 \le t \le T), \qquad \sum_{t=0}^{T-1} \eta_t \|\mathbf{v}_t\|^2 \le 6\Delta_0. \tag{44}$$

Proof. For any $\mathcal{F}_t$-measurable vector $\mathbf{u}$, set $X = \langle \mathbf{u}, \mathbf{n}_t \rangle$ and $a = S_t\|\mathbf{u}\|^2$. If $a = 0$, then $X = 0$. By the subgaussian assumption on $\mathbf{n}_t$, we obtain, for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}_t e^{\lambda X} \le e^{2\lambda^2 a}. \tag{45}$$

Let $\tau = \inf\{t \ge 0 : \Delta_t > D\}$, with $\inf \varnothing = +\infty$, and define

$$Y_t = -\eta_t \langle \mathbf{v}_t, \mathbf{n}_t \rangle \mathbf{1}_{\{t<\tau\}}, \quad V_t = 4\eta_t^2 S_t \|\mathbf{v}_t\|^2 \mathbf{1}_{\{t<\tau\}}, \quad M_s = \sum_{t=0}^{s-1} Y_t.$$

Since $\{t < \tau\} \in \mathcal{F}_t$, (45) implies $\mathbb{E}_t e^{\lambda Y_t} \le e^{\lambda^2 V_t/2}$. Consequently, for each fixed $\lambda > 0$,

$$Z_s(\lambda) := \exp\left(\lambda M_s - \frac{\lambda^2}{2}\sum_{t=0}^{s-1} V_t\right), \qquad 0 \le s \le T,$$

is a nonnegative supermartingale Howard et al. (2020): $\mathbb{E}_t Z_{t+1}(\lambda) \le Z_t(\lambda)$ and $Z_0(\lambda) = 1$. Before the exit, $S_t \le S(z_t)$ and $\|\mathbf{v}_t\| \le G(z_t)$. The last entry of (43) therefore gives the deterministic budget

$$V_t \le \frac{\Delta_0^2}{2T\ell}, \quad \sum_{t=0}^{T-1} V_t \le V := \frac{\Delta_0^2}{2\ell} \quad \text{almost surely,}$$

since $V_t = 0$ after the exit. By the exponential-supermartingale concentration Howard et al. (2020), we obtain that

$$\mathbb{P}(M_T > \Delta_0) \le \exp\left(-\frac{\Delta_0^2}{2V}\right) = e^{-\ell} = \frac{\delta}{2}.$$

Conditional Markov's inequality gives $\mathbb{P}(\|\mathbf{n}_t\|^2 > \kappa S_t \mid \mathcal{F}_t) \le e^{1-\kappa} = \delta/(2T)$. Therefore

$$\mathcal{E}_N := \bigcap_{t=0}^{T-1} \{\|\mathbf{n}_t\|^2 \le \kappa S_t\}, \quad \mathbb{P}(\mathcal{E}_N) \ge 1 - \frac{\delta}{2}.$$

On $\mathcal{E}_N$ and for $t < \tau$, the first three entries of the learning rate rule imply

$$\eta_t\|\mathbf{g}_t\| \le r(z_t) \le r(\Delta_t), \qquad L(z_t)\eta_t \le \frac{1}{2}, \qquad L(z_t)\eta_t^2\|\mathbf{n}_t\|^2 \le \frac{\Delta_0}{T}.$$

Here r is nonincreasing and L is nondecreasing. Thus the local smoothness argument underlying (29) yields

$$\Delta_{t+1} \le \Delta_t - \frac{\eta_t}{2}\|\mathbf{v}_t\|^2 - \eta_t \langle \mathbf{v}_t, \mathbf{n}_t\rangle + L(z_t)\eta_t^2\|\mathbf{n}_t\|^2.$$

Set $s = T \wedge \tau$ and $Q_T = \sum_{t=0}^{T-1} L(z_t)\eta_t^2\|\mathbf{n}_t\|^2 \mathbf{1}_{\{t<\tau\}}$. On $\mathcal{E}_N$, $Q_T \le \Delta_0$. Summing to $s-1$ on $\mathcal{E}_N \cap \{M_T \le \Delta_0\}$ gives

$$\Delta_s + \frac{1}{2}\sum_{t=0}^{s-1} \eta_t\|\mathbf{v}_t\|^2 \le \Delta_0 + M_T + Q_T \le 3\Delta_0.$$

This event has probability at least $1 - \delta$. If $\tau \le T$, this contradicts $\Delta_\tau > D$; hence $\tau > T$. Taking $s = T$ and using $\Delta_T \ge 0$ proves (44). □

Lemma 18. *Let $0 < z_t \le D$ and $\eta_t = h^{\mathrm{sg}}_{A,B}(z_t)$, and let $\overline{H} = T^{-1}\sum_{t<T} H(z_t)$ and $\overline{R} = T^{-1}\sum_{t<T} r(z_t)^{-1}$. Then we have*

$$\frac{\Delta_0}{\sum_{t=0}^{T-1}\eta_t} \le \mathcal{O}\left[\frac{\Delta_0}{T}\left(\overline{H} + \sqrt{\kappa C}\,\overline{R}\right) + \sqrt{\frac{\Delta_0\kappa}{T}\left((C + A\Delta_0)\overline{H} + B\Delta_0\overline{H}^2\right)}\right], \tag{46}$$

with an absolute numerical implicit constant.

Proof. Cauchy–Schwarz yields

$$\frac{1}{\sum_{t=0}^{T-1}\eta_t} \le \frac{1}{T^2}\sum_{t=0}^{T-1}\frac{1}{\eta_t}. \tag{47}$$

By the definition of η_t in (43), we have

$$\begin{aligned}\frac{1}{\eta_t} \le{}& \frac{G(z_t) + \sqrt{\kappa S(z_t)}}{r(z_t)} + 2L(z_t) \\ &+ \sqrt{\frac{T\kappa}{\Delta_0}}\sqrt{L(z_t)S(z_t)} + \frac{\sqrt{8T\ell}}{\Delta_0}G(z_t)\sqrt{S(z_t)}.\end{aligned} \tag{48}$$

We bound these four contributions before simplifying the result.

Set $c_r := 2 + \sqrt{2}$. Recall from (35) that, for $u > 0$,

$$\frac{\sqrt{u}}{r(u)} \le c_r \sqrt{H(u)}, \qquad \frac{G(u)}{r(u)} \le 2c_r H(u), \qquad L(u) \le 2H(u).$$

Using

$$\sqrt{S(u)} \le \sqrt{C} + \sqrt{A}\sqrt{u} + \sqrt{B}\, G(u),$$

we obtain

$$\begin{aligned} &\frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{G(z_t) + \sqrt{\kappa S(z_t)}}{r(z_t)} + 2L(z_t)\right] \\ &\qquad \le c_0\left[(1+\sqrt{\kappa B})\overline{H} + \sqrt{\kappa A}\sqrt{\overline{H}} + \sqrt{\kappa C}\,\overline{R}\right] \end{aligned} \tag{49}$$

for an absolute numerical constant c_0, where we used $\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{H(z_t)} \le \left(\frac{1}{T}\sum_{t=0}^{T-1} H(z_t)\right)^{1/2} = \sqrt{\overline{H}}$.

Next, since $u \le D$ and $G(u) = 2\sqrt{uH(u)}$,

$$\begin{aligned} \sqrt{L(u)S(u)} &\le \sqrt{2}(\sqrt{C} + \sqrt{AD})\sqrt{H(u)} + 2\sqrt{2BD}\, H(u), \\ G(u)\sqrt{S(u)} &\le 2\sqrt{D}(\sqrt{C} + \sqrt{AD})\sqrt{H(u)} + 4\sqrt{B}\, DH(u). \end{aligned}$$

Evaluating these inequalities at $u = z_t$, averaging over $t = 0, \dots, T-1$, and using the preceding bound on $T^{-1}\sum_t \sqrt{H(z_t)}$, we obtain

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1}\sqrt{L(z_t)S(z_t)} &\le \sqrt{2}(\sqrt{C} + \sqrt{AD})\sqrt{\overline{H}} + 2\sqrt{2BD}\,\overline{H}, \\ \frac{1}{T}\sum_{t=0}^{T-1} G(z_t)\sqrt{S(z_t)} &\le 2\sqrt{D}(\sqrt{C} + \sqrt{AD})\sqrt{\overline{H}} + 4\sqrt{B}\, D\overline{H}. \end{aligned}$$

Multiplying (47) by Δ_0 and substituting (48) gives

$$\begin{aligned} \frac{\Delta_0}{\sum_{t=0}^{T-1}\eta_t} &\le \frac{\Delta_0}{T}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\eta_t}\right) \\ &\le \frac{\Delta_0}{T}\left\{\frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{G(z_t) + \sqrt{\kappa S(z_t)}}{r(z_t)} + 2L(z_t)\right]\right\} \\ &\quad + \sqrt{\frac{\Delta_0\kappa}{T}}\left\{\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{L(z_t)S(z_t)}\right\} \\ &\quad + \sqrt{\frac{8\ell}{T}}\left\{\frac{1}{T}\sum_{t=0}^{T-1} G(z_t)\sqrt{S(z_t)}\right\}. \end{aligned}$$

We bound the first average by (49) and the remaining two averages by the inequalities above. Substituting $D = 3\Delta_0$ yields

$$\begin{aligned} \frac{\Delta_0}{\sum_{t=0}^{T-1}\eta_t} &\le \frac{c_0\Delta_0}{T}\left[(1+\sqrt{\kappa B})\overline{H} + \sqrt{\kappa A}\sqrt{\overline{H}} + \sqrt{\kappa C}\,\overline{R}\right] \\ &\quad + \sqrt{\frac{\Delta_0\kappa}{T}}\left[\sqrt{2}(\sqrt{C} + \sqrt{3A\Delta_0})\sqrt{\overline{H}} + 2\sqrt{6B\Delta_0}\,\overline{H}\right] \\ &\quad + \sqrt{\frac{\Delta_0\ell}{T}}\left[4\sqrt{6}(\sqrt{C} + \sqrt{3A\Delta_0})\sqrt{\overline{H}} + 24\sqrt{2B\Delta_0}\,\overline{H}\right]. \end{aligned}$$

Since $\ell \le \kappa$, the last two terms can be combined by collecting their numerical factors. Consequently,

$$\begin{aligned} \frac{\Delta_0}{\sum_{t=0}^{T-1}\eta_t} \le \mathcal{O}\Bigg[&\frac{\Delta_0}{T}\left[(1+\sqrt{\kappa B})\overline{H} + \sqrt{\kappa A}\sqrt{\overline{H}} + \sqrt{\kappa C}\,\overline{R}\right] \\ &+ \sqrt{\frac{\Delta_0\kappa}{T}}\left[(\sqrt{C} + \sqrt{A\Delta_0})\sqrt{\overline{H}} + \sqrt{B\Delta_0}\,\overline{H}\right]\Bigg]. \end{aligned} \tag{50}$$

To simplify the first term, separate the local contribution into

$$\frac{\Delta_0}{T}\left[(1+\sqrt{\kappa B})\overline{H}+\sqrt{\kappa A}\sqrt{\overline{H}}+\sqrt{\kappa C}\,\overline{R}\right]$$
$$=\frac{\Delta_0}{T}(\overline{H}+\sqrt{\kappa C}\,\overline{R})+\frac{\sqrt{\kappa A}\,\Delta_0}{T}\sqrt{\overline{H}}+\frac{\sqrt{\kappa B}\,\Delta_0}{T}\overline{H}.$$

The first term already has the form required in (46). Since $T \geq 1$, the remaining two terms satisfy

$$\frac{\sqrt{\kappa A}\,\Delta_0}{T}\sqrt{\overline{H}}=\frac{1}{\sqrt{T}}\sqrt{\frac{\Delta_0\kappa}{T}}\sqrt{A\Delta_0}\sqrt{\overline{H}}\leq\sqrt{\frac{\Delta_0\kappa}{T}}\sqrt{A\Delta_0}\sqrt{\overline{H}},$$

$$\frac{\sqrt{\kappa B}\,\Delta_0}{T}\overline{H}=\frac{1}{\sqrt{T}}\sqrt{\frac{\Delta_0\kappa}{T}}\sqrt{B\Delta_0}\,\overline{H}\leq\sqrt{\frac{\Delta_0\kappa}{T}}\sqrt{B\Delta_0}\,\overline{H}.$$

Thus, each additional local term is bounded by its corresponding component in the stochastic term. Absorbing these contributions into (50), we obtain

$$\frac{\Delta_0}{\sum_{t=0}^{T-1}\eta_t}\leq\mathcal{O}\Bigg[\frac{\Delta_0}{T}(\overline{H}+\sqrt{\kappa C}\,\overline{R})$$
$$+\sqrt{\frac{\Delta_0\kappa}{T}}\left(\sqrt{C}\sqrt{\overline{H}}+\sqrt{A\Delta_0}\sqrt{\overline{H}}+\sqrt{B\Delta_0}\,\overline{H}\right)\Bigg].$$

It remains to combine the three nonnegative quantities inside the last parentheses into a single square root. By Cauchy–Schwarz,

$$\left(\sqrt{C}\sqrt{\overline{H}}+\sqrt{A\Delta_0}\sqrt{\overline{H}}+\sqrt{B\Delta_0}\,\overline{H}\right)^2\leq 3\left[(C+A\Delta_0)\overline{H}+B\Delta_0\overline{H}^2\right].$$

Taking square roots, substituting this estimate into the preceding bound, and absorbing the numerical factor $\sqrt{3}$ into the implicit constant yields

$$\frac{\Delta_0}{\sum_{t=0}^{T-1}\eta_t}\leq\mathcal{O}\left[\frac{\Delta_0}{T}(\overline{H}+\sqrt{\kappa C}\,\overline{R})+\sqrt{\frac{\Delta_0\kappa}{T}\left[(C+A\Delta_0)\overline{H}+B\Delta_0\overline{H}^2\right]}\right].$$

This proves (46). □

I.2 Warmup and constant learning rate results

Use $H_{\mathrm{ad}}, R_{\mathrm{ad}}, H_{\mathrm{c}}, R_{\mathrm{c}}$ from (24). $\mathcal{O}_\rho$ hides only numerical factors depending on ρ.

Theorem 8. *Under the assumptions of Lemma 17, let $\eta_t = h^{\mathrm{sg}}_{A,B}(\Delta_t)$. With probability at least $1-\delta$, $\Delta_t \leq 3\Delta_0$ for all $0 \leq t \leq T$, and*

$$\min_{t<T}\|\nabla f(\mathbf{w}_t)\|^2\leq\mathcal{O}_\rho\Bigg[\frac{\Delta_0}{T}\left(H_{\mathrm{ad}}+\sqrt{\kappa C}\,R_{\mathrm{ad}}\right)$$
$$+\sqrt{\frac{\Delta_0\kappa}{T}\left((C+A\Delta_0)H_{\mathrm{ad}}+B\Delta_0H_{\mathrm{ad}}^2\right)}\Bigg].$$

Proof. Lemma 17 gives, on one event of probability at least $1-\delta$, the claimed trajectory bound and $\min_{t<T}\|\mathbf{v}_t\|^2 \leq 6\Delta_0/\sum_{t<T}\eta_t$. If some $\Delta_t = 0$ with $t < T$, the minimum gradient norm is zero. Otherwise apply Lemma 18 with $z_t = \Delta_t$. The adaptive specialization in (39) gives $\overline{H} \leq 3^\rho H_{\mathrm{ad}}$ and $\overline{R} \leq R_{\mathrm{ad}}$, proving the stated bound on the same event. □

Theorem 9. *Under the same assumptions as Theorem 8, let $\eta_t \equiv \eta = h^{\mathrm{sg}}_{A,B}(3\Delta_0)$. With probability at least $1-\delta$, $\Delta_t \leq 3\Delta_0$ for all $0 \leq t \leq T$, and*

$$\min_{t<T}\|\nabla f(\mathbf{w}_t)\|^2\leq\mathcal{O}_\rho\Bigg[\frac{\Delta_0}{T}\left(H_{\mathrm{c}}+\sqrt{\kappa C}\,R_{\mathrm{c}}\right)$$
$$+\sqrt{\frac{\Delta_0\kappa}{T}\left((C+A\Delta_0)H_{\mathrm{c}}+B\Delta_0H_{\mathrm{c}}^2\right)}\Bigg].$$

Proof. Apply Lemma 17 with $z_t = D$ to obtain the trajectory bound and $\min_{t<T} \|\mathbf{v}_t\|^2 \leq 6\Delta_0/(T\eta)$. Lemma 18 now has $\overline{H} = H(D) \leq 12^\rho H_c$ and $\overline{R} = r(D)^{-1} \leq 4^{(\rho-1)/2} R_c$, by (39). Substitution proves the result. □

J LICENSES

The CIFAR-10 dataset Krizhevsky (2009) (https://huggingface.co/datasets/uoft-cs/cifar10) is released without license. The NanoGPT Karpathy (2022) (https://github.com/karpathy/nanogpt) codebase is under the MIT license.