PROVABLE BENEFITS OF BLOCK PRECONDITIONING OVER GRADIENT DESCENT UNDER BLOCK-HETEROGENEOUS SMOOTHNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Motivated by the empirical advantage of adaptive optimizers on Transformers, we study the comparison between block-preconditioned gradient descent (Block-PGD) and constant-step gradient descent (GD). We introduce *block-heterogeneous smoothness*, in which parameter blocks have distinct smoothness constants L_k, and measure stationarity by $\mathcal{G}_{\mathbf{L}}^2 = \sum_k \|\nabla_k f\|^2/L_k$, the ordinary gradient norm in curvature-normalized coordinates. Block-PGD attains $\min_{t<T} \mathcal{G}_{\mathbf{L}}^2 \le 2\Delta/T$, whereas GD attains $2\mathcal{R}_H\Delta/T$, where $\mathcal{R}_H = L_{\max}/L_{\min}$. A quadratic instance proves that linear dependence on $\mathcal{R}_H$ is necessary for GD, including under loss and ordinary-gradient criteria, although it does not match the general nonconvex dependence on accuracy. We extend the upper bounds to stochastic gradients, establish a local Zipfian softmax majorizer with ratio V^α, and give a conditional RMSProp-style result with $\sqrt{\mathcal{R}_H}$ dependence. Direct Hessian–vector-product diagnostics on a small GPT-2 model trained on WikiText-2 find persistent architecture-block curvature heterogeneity and strong second-moment/curvature alignment for output-class blocks. These results explain the benefit of block preconditioning; they do not constitute a convergence theory for full Adam with momentum.

1 INTRODUCTION

The choice of optimizer in modern deep learning depends strongly on architecture: Adam (Kingma & Ba, 2015) often outperforms SGD on Transformers (Vaswani et al., 2017), whereas SGD with momentum remains competitive on many CNN workloads (He et al., 2016). This observation motivates our study, but our formal comparison is narrower: we analyze blockwise preconditioning versus a single global step size. We do not prove convergence guarantees for full Adam, whose momentum direction and history-dependent second moment require additional control.

Recent work has proposed several competing hypotheses. Zhang et al. (2024) demonstrated that the Hessian of transformer losses exhibits *block heterogeneity*—different parameter blocks have substantially different spectral properties—and proved separation on convex quadratics. Kunstner et al. (2024) showed that Zipfian class imbalance in language data causes Adam (via sign-like updates) to make uniform progress across classes, while GD is bottlenecked by rare tokens. Jiang et al. (2025) proved the first non-convex separation between AdaGrad and SGD, but using the non-standard ℓ_1-stationarity measure. Li et al. (2025) proved linear training-loss convergence and poor generalization for SignGD on a specified noisy two-layer transformer task. This is not a matched GD separation, and SignGD is not interchangeable with Adam (Orvieto & Gower, 2025). Zhang et al. (2025) showed that Adam's advantage is coordinate-system dependent, requiring rotation-dependent assumptions. These results are informative but fragmented, each isolating one facet (Hessian structure, data imbalance, sign-based updates, or rotation dependence) without clarifying whether they describe the same underlying mechanism.

We isolate a common optimization primitive behind several of these explanations: each block k has a smoothness scale L_k, and a single GD step is constrained by $L_{\max}$ even when progress is needed in a block with scale $L_{\min}$. The resulting ratio $\mathcal{R}_H = L_{\max}/L_{\min}$ quantifies the value of ideal block preconditioning. This abstraction connects Hessian heterogeneity and class imbalance, while

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.

deliberately separating the proved preconditioning result from the harder question of how Adam estimates and uses this geometry.

Our Contributions.

1. We introduce the *block-heterogeneous smoothness* framework (Definition 1), a natural abstraction of the curvature structure observed in transformers, and an associated *weighted stationarity measure* $\mathcal{G}_{\mathbf{L}}^2$ (Definition 3).
2. We prove an $\mathcal{R}_H$ **separation between upper bounds**: Block-PGD finds $\mathcal{G}_{\mathbf{L}}^2 \leq \varepsilon^2$ in $\mathcal{O}(\Delta/\varepsilon^2)$ iterations (Theorem 2), while constant-step GD has an $\mathcal{O}(\mathcal{R}_H \Delta/\varepsilon^2)$ guarantee (Theorem 3). An exact quadratic proves necessary linear dependence on $\mathcal{R}_H$ for GD, also in loss and the ordinary gradient norm, but leaves the joint nonconvex $(\mathcal{R}_H, \varepsilon)$ lower bound open (Theorem 4).
3. We extend to the **stochastic setting** (Theorem 6), showing that the $\mathcal{R}_H$ factor appears in the bias term of the GD upper bound but not in the Block-PGD bound.
4. Under explicit local well-specification and class-conditional moment assumptions, we give a softmax quadratic majorizer with ratio $\mathcal{R}_H = V^\alpha$ (Theorem 7); this is not a global or minimal curvature-ratio claim.
5. We prove that any momentum-free diagonal method whose steps lie in $[c/L_k, 1/L_k]$ attains the corresponding preconditioned rate, and derive a **RMSProp-style $\sqrt{\mathcal{R}_H}$ guarantee** when second moments are comparable to L_k (Theorem 9 and Corollary 10).
6. Direct curvature diagnostics on WikiText-2 show large architecture-block heterogeneity, but Adam's steps do not satisfy the $1/L_k$ bracket on that decomposition. In the output-class decomposition, second moments align strongly with curvature. This decomposition-dependent evidence clarifies both the relevance and the limitation of our theory.

2 Related Work

Adaptive vs. Non-Adaptive Methods: Theory. The study of adaptive methods begins with Duchi et al. (2011), who established data-dependent regret bounds that exploit gradient geometry in online convex optimization. Jiang et al. (2025) extended this to non-convex stochastic optimization, proving a factor-d improvement for AdaGrad under anisotropic smoothness, but using ℓ_1-stationarity. Zhang et al. (2024) proved separation on convex quadratics with heterogeneous block Hessians. Yadav et al. (2025) proved sign descent outperforms normalized GD on a softmax unigram model with heavy-tailed class imbalance. Our contribution is a transparent Block-PGD/GD comparison under a block quadratic majorizer, together with a quadratic showing necessary linear dependence on its heterogeneity ratio. Unlike analyses of Adam or AdaGrad, Block-PGD assumes access to the block scales.

Empirical Understanding of Adam vs. SGD. Kunstner et al. (2023) showed the Adam-SGD gap persists in full-batch settings, ruling out noise as the primary factor. Kunstner et al. (2024) linked the gap to Zipfian class imbalance. Pan & Li (2023) connected Adam's advantage to reduced directional sharpness. Tomihari & Sato (2026) showed gradient heterogeneity across parameters is linked to LayerNorm placement. Orvieto & Gower (2025) demonstrated that Signum consistently underperforms Adam, showing Adam's advantage goes beyond sign normalization.

Block Structure in Optimization. Block-coordinate methods have a long history (Wright, 2015). Our block-heterogeneous smoothness is related to coordinate-wise smoothness (Nesterov, 2012) and (L_0, L_1)-smoothness (Zhang et al., 2020), but specifically targets the inter-block heterogeneity that arises from neural network architecture. We analyze *full-gradient* methods (GD and preconditioned GD) rather than coordinate descent, which selects one block per step. Both methods considered here update all blocks at each iteration; coordinate descent instead counts selective block updates.

What the preconditioner represents. Natural gradient uses information geometry (Amari, 1998); K-FAC approximates curvature with Kronecker factors (Martens and Grosse, 2015), and Shampoo accumulates tensor-structured gradient statistics (Gupta et al., 2018). None is identical to an oracle

block smoothness scale. Nonconvex AdaGrad analyses (Ward et al., 2019; Li and Orabona, 2019) and the Adam convergence counterexample (Reddi et al., 2018) further show why adaptivity alone does not establish our guarantee. Decoupled weight decay (Loshchilov and Hutter, 2019) changes the update beyond its preconditioner. We therefore compare declared update rules, using the stationarity framework of stochastic nonconvex optimization (Ghadimi and Lan, 2013), rather than transferring the result to every adaptive optimizer. Randomized block-coordinate complexity (Richtárik and Takáč, 2014) provides a related baseline, but counts selective block updates rather than our simultaneous full-gradient steps.

3 Problem Setup

3.1 Block-Heterogeneous Smoothness

We consider the minimization of a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ where the parameter vector admits a block decomposition $w = (w_1, \ldots, w_K) \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_K}$ with $\sum_{k=1}^K d_k = d$.

Definition 1 (Block-Heterogeneous Smoothness)**.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is $\mathbf{L}$-*block smooth* with constants $\mathbf{L} = (L_1, \ldots, L_K)$, where $L_k > 0$, if for all $w, \delta \in \mathbb{R}^d$:

$$f(w + \delta) \leq f(w) + \langle \nabla f(w), \delta \rangle + \frac{1}{2} \sum_{k=1}^{K} L_k \left\| \delta_k \right\|^2, \tag{1}$$

where $\delta = (\delta_1, \ldots, \delta_K)$ is the block decomposition of δ.

Remark 1. *The definition implies a scalar quadratic upper model with $L_{\max} = \max_k L_k$, since $\sum_k L_k \|\delta_k\|^2 \leq L_{\max} \|\delta\|^2$. For nonconvex functions this one-sided condition does not itself imply a globally $L_{\max}$-Lipschitz gradient. The analysis below uses only the upper model. A heterogeneous majorizer can be tighter in low-scale blocks; its admissible constants need not be minimal.*

Definition 2 (Heterogeneity Ratio)**.** The *heterogeneity ratio* of an $\mathbf{L}$-block smooth function is $\mathcal{R}_H := L_{\max}/L_{\min}$, where $L_{\max} = \max_k L_k$ and $L_{\min} = \min_k L_k$.

$\mathcal{R}_H$ describes the chosen valid majorizer, not an intrinsic property of f: inflating an admissible L_k can change it without changing f. We compare both algorithms using the same fixed $\mathbf{L}$ and assume access to these constants; local curvature probes do not certify their global validity.

3.2 Weighted Stationarity Measure

Definition 3 (Weighted Stationarity)**.** For an $\mathbf{L}$-block smooth function f, the $\mathbf{L}$-*weighted stationarity measure* is

$$\mathcal{G}_{\mathbf{L}}^2(w) := \sum_{k=1}^{K} \frac{1}{L_k} \left\| \nabla_k f(w) \right\|^2, \tag{2}$$

where $\nabla_k f(w) := \nabla_{w_k} f(w) \in \mathbb{R}^{d_k}$ denotes the block-k partial gradient.

Let $z_k = \sqrt{L_k} w_k$ and write $\widetilde{f}(z) = f(w)$. Then $\mathcal{G}_{\mathbf{L}}^2 = \|\nabla_z \widetilde{f}\|^2$: it is ordinary Euclidean stationarity after normalizing the units of each block. Equivalently, it is the squared norm $\|\mathbf{H}^{-1/2} \nabla f\|^2$ for $\mathbf{H} = \operatorname{diag}(L_1 I_{d_1}, \ldots, L_K I_{d_K})$, and twice the decrease certified by minimizing the block-smooth quadratic upper model. This geometry is natural for block preconditioning, but an improvement in $\mathcal{G}_{\mathbf{L}}$ does not automatically imply an equal improvement in loss or the ordinary gradient norm on every objective.

Proposition 1 (Equivalence with Standard Stationarity)**.** *For any $\mathbf{L}$-block smooth function:*

$$\frac{1}{L_{\max}} \left\| \nabla f(w) \right\|^2 \leq \mathcal{G}_{\mathbf{L}}^2(w) \leq \frac{1}{L_{\min}} \left\| \nabla f(w) \right\|^2. \tag{3}$$

Hence $\mathcal{G}_{\mathbf{L}}^2(w) \leq \varepsilon^2$ implies $\|\nabla f(w)\|^2 \leq L_{\max} \varepsilon^2$, and $\|\nabla f(w)\|^2 \leq \varepsilon^2$ implies $\mathcal{G}_{\mathbf{L}}^2(w) \leq \varepsilon^2 / L_{\min}$.

Proof. Since $L_{\min} \leq L_k \leq L_{\max}$ for all k, we have $\frac{1}{L_{\max}} \leq \frac{1}{L_k} \leq \frac{1}{L_{\min}}$. Multiplying by $\|\nabla_k f\|^2$ and summing over k yields the result. □

Algorithm 1 Block-Preconditioned Gradient Descent (Block-PGD)

Require: Initial point w_0, block smoothness constants $\mathbf{L} = (L_1, \ldots, L_K)$
1: **for** $t = 0, 1, \ldots, T-1$ **do**
2: Compute $g_{t,k} = \nabla_k f(w_t)$ for each block $k \in [K]$
3: Update: $w_{t+1,k} = w_{t,k} - \frac{1}{L_k} g_{t,k}$ for each $k \in [K]$
4: **end for**
5: **return** $w_{\hat{t}}$ where $\hat{t} = \arg\min_{t \in \{0,\ldots,T-1\}} \mathcal{G}^2_{\mathbf{L}}(w_t)$

Remark 2 (Practical Interpretation). *$\mathcal{G}^2_{\mathbf{L}}(w) \le \varepsilon^2$ implies $\|\nabla_k f(w)\|^2 \le L_k \varepsilon^2$ for every block, while Proposition 1 converts it to ordinary stationarity with a factor $L_{\max}$. This conversion can erase the displayed $\mathcal{R}_H$ upper-bound separation. We therefore report both metrics empirically and distinguish the general normalized-stationarity result from the standard-metric separation on our explicit quadratic instance.*

4 Main Results

4.1 Upper Bound for Block-Adaptive Methods

We first analyze block-preconditioned gradient descent (Block-PGD), which uses a per-block step size $1/L_k$.

Theorem 2 (Block-Adaptive Upper Bound). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mathbf{L}$-block smooth with $\Delta := f(w_0) - \inf_w f(w) < \infty$. Then Block-PGD (Algorithm 1) satisfies:*

$$\min_{0 \le t \le T-1} \mathcal{G}^2_{\mathbf{L}}(w_t) \le \frac{2\Delta}{T}. \tag{4}$$

Proof. By the $\mathbf{L}$-block smoothness condition (1) with $\delta_k = -\frac{1}{L_k}\nabla_k f(w_t)$:

$$\begin{aligned} f(w_{t+1}) &\le f(w_t) + \sum_{k=1}^K \left\langle \nabla_k f(w_t), -\frac{1}{L_k}\nabla_k f(w_t) \right\rangle + \frac{1}{2}\sum_{k=1}^K L_k \cdot \frac{1}{L_k^2} \|\nabla_k f(w_t)\|^2 \\ &= f(w_t) - \sum_{k=1}^K \frac{1}{L_k}\|\nabla_k f(w_t)\|^2 + \frac{1}{2}\sum_{k=1}^K \frac{1}{L_k}\|\nabla_k f(w_t)\|^2 \\ &= f(w_t) - \frac{1}{2}\mathcal{G}^2_{\mathbf{L}}(w_t). \end{aligned} \tag{5}$$

Summing from $t = 0$ to $T-1$ and using $f(w_T) \ge \inf f$:

$$\frac{1}{2}\sum_{t=0}^{T-1} \mathcal{G}^2_{\mathbf{L}}(w_t) \le f(w_0) - f(w_T) \le \Delta. \tag{6}$$

The minimum over t is bounded by the average: $\min_t \mathcal{G}^2_{\mathbf{L}}(w_t) \le \frac{1}{T}\sum_t \mathcal{G}^2_{\mathbf{L}}(w_t) \le \frac{2\Delta}{T}$. □

4.2 Upper and Lower Bounds for Gradient Descent

Theorem 3 (GD Upper Bound under Block Smoothness). *Let f be $\mathbf{L}$-block smooth with gap Δ. Gradient descent with step size $\eta = 1/L_{\max}$ satisfies:*

$$\min_{0 \le t \le T-1} \mathcal{G}^2_{\mathbf{L}}(w_t) \le \frac{2\mathcal{R}_H \cdot \Delta}{T}. \tag{7}$$

Proof. GD updates $w_{t+1} = w_t - \eta \nabla f(w_t)$ with $\eta = 1/L_{\max}$. By $\mathbf{L}$-block smoothness:

$$\begin{aligned} f(w_{t+1}) &\le f(w_t) - \eta\|\nabla f(w_t)\|^2 + \frac{\eta^2}{2}\sum_{k=1}^K L_k \|\nabla_k f(w_t)\|^2 \\ &\le f(w_t) - \eta\|\nabla f(w_t)\|^2 + \frac{\eta^2 L_{\max}}{2}\|\nabla f(w_t)\|^2 = f(w_t) - \frac{1}{2L_{\max}}\|\nabla f(w_t)\|^2. \end{aligned} \tag{8}$$

Summing over t: $\min_t \|\nabla f(w_t)\|^2 \le 2L_{\max}\Delta/T$. By Proposition 1:

$$\min_t \mathcal{G}^2_{\mathbf{L}}(w_t) \le \frac{1}{L_{\min}} \cdot \frac{2L_{\max}\Delta}{T} = \frac{2\mathcal{R}_H \cdot \Delta}{T}. \tag{9}$$

□

The following quadratic shows that no guarantee can be uniform in $\mathcal{R}_H$: GD has necessary linear dependence on heterogeneity on this instance. It does not, however, match the $1/\varepsilon^2$ dependence of the preceding general nonconvex upper bound.

Theorem 4 (Necessary Linear Heterogeneity Dependence for GD)**.** *For any $K \ge 2$, any $L_1 \le L_K$ with $\mathcal{R}_H = L_K/L_1 \ge 2$, and any constant step size $0 < \eta \le 1/L_K$, there exists an (L_1, L_K)-block smooth function $f : \mathbb{R}^{d_1+d_K} \to \mathbb{R}$ with $f(w_0) - \inf f = \Delta$ such that constant-step GD requires at least*

$$T \ge \frac{(1-\eta L_1)}{2\eta L_1} \log \frac{\Delta}{\varepsilon^2} \tag{10}$$

iterations to achieve $\mathcal{G}^2_{\mathbf{L}}(w_T) \le \varepsilon^2$. With $\eta = 1/L_K$ and $\mathcal{R}_H \ge 2$, this gives $T \ge \frac{\mathcal{R}_H}{4} \log(\Delta/\varepsilon^2)$.

For $\eta = 1/L_K$, the same instance also satisfies

$$f(w_t) - f^* \ge \frac{\Delta}{2}(1 - 1/\mathcal{R}_H)^{2t}, \qquad \|\nabla f(w_t)\|^2 \ge \Delta L_{\min}(1 - 1/\mathcal{R}_H)^{2t}. \tag{11}$$

Thus GD also needs $\Omega(\mathcal{R}_H \log(1/\varepsilon))$ iterations under loss and ordinary-gradient criteria, whereas Block-PGD reaches the optimum in one step on this instance.

Proof sketch (full proof in Appendix B.1). We construct $f(w_1, w_K) = \frac{L_1}{2}\|w_1 - w_1^*\|^2 + \frac{L_K}{2}\|w_K\|^2$ with initial gap Δ. Since f is separable, GD decomposes: $\mathcal{G}^2_{\mathbf{L}}(w_t) = \Delta[(1-\eta L_1)^{2t} + (1-\eta L_K)^{2t}]$. The slow block dominates; using $\log(1/(1-x)) \le x/(1-x)$ yields the stated lower bound. The loss and ordinary-gradient inequalities follow from retaining the same slow-block term. Full details appear in Appendix B.1. □

Corollary 5 (Separation)**.** *Under $\mathbf{L}$-block smoothness with heterogeneity ratio $\mathcal{R}_H$, the iteration counts for constant-step GD and Block-PGD to achieve $\mathcal{G}^2_{\mathbf{L}} \le \varepsilon^2$ satisfy:*

$$T_{\mathrm{GD}}(\varepsilon) = \mathcal{O}\left(\frac{\mathcal{R}_H \cdot \Delta}{\varepsilon^2}\right), \qquad T_{\text{Block-PGD}}(\varepsilon) = \mathcal{O}\left(\frac{\Delta}{\varepsilon^2}\right). \tag{12}$$

The quadratic lower bound establishes necessary linear dependence on $\mathcal{R}_H$ in one regime, but does not establish tightness of the product $\mathcal{R}_H\Delta/\varepsilon^2$ for general nonconvex functions. Determining the joint worst-case dependence on $(\mathcal{R}_H, \varepsilon)$ remains open.

Appendix B.2 gives an anisotropic, non-Newton construction showing that an $\Omega(\mathcal{R}_H/q)$ advantage survives multiplicative smoothness-estimation error q. Appendix B.3 also compares the two methods' descent-lemma certificates at a common point. This is a comparison of quadratic-model guarantees, not of realized decreases along different trajectories, since either certificate can be loose.

4.3 Stochastic Extension

Assumption 1 (Block-Structured Noise)**.** The stochastic gradient $\hat{g}_t$ satisfies $\mathbb{E}[\hat{g}_t|w_t] = \nabla f(w_t)$ and $\mathbb{E}[\|\hat{g}_{t,k} - \nabla_k f(w_t)\|^2 |w_t] \le \sigma_k^2$ for each block k.

Theorem 6 (Stochastic Separation)**.** *Under Assumption 1 and $\mathbf{L}$-block smoothness with gap Δ:*

***(a) Stochastic Block-PGD** with step sizes $\eta_k = 1/L_k$ satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\mathcal{G}^2_{\mathbf{L}}(w_t)\big] \le \frac{2\Delta}{T} + \sum_{k=1}^{K} \frac{\sigma_k^2}{L_k}. \tag{13}$$

***(b) Stochastic GD** with step size $\eta = 1/L_{\max}$ satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\mathcal{G}^2_{\mathbf{L}}(w_t)\big] \le \frac{2\mathcal{R}_H \cdot \Delta}{T} + \frac{1}{L_{\min}L_{\max}} \sum_{k=1}^{K} L_k\sigma_k^2. \tag{14}$$

The ratio of the bias terms is $\mathcal{R}_H$, matching the deterministic upper-bound comparison. The variance terms are not *uniformly ordered: the contribution of block k differs by the factor $L_k^2/(L_{\min}L_{\max})$, which can be below or above one. Our stochastic separation therefore concerns the transient bias term, not a universal noise-floor advantage.*

Proof sketch (full proof in Appendix B.4). Both parts follow the deterministic analysis with an additional variance term from $\mathbb{E}[\|\xi_{t,k}\|^2] \le \sigma_k^2$. For Block-PGD, the variance term is $\sum_k \sigma_k^2/L_k$; for GD, conversion via $\mathcal{G}_{\mathbf{L}}^2 \le \|\nabla f\|^2/L_{\min}$ gives the displayed, generally incomparable expression. □

5 Application to Transformers

5.1 Softmax Regression with Zipfian Data

Consider the multiclass classification problem with V classes and feature dimension d:

$$f(W) = -\mathbb{E}_{(x,y)}\left[\log \frac{\exp(w_y^\top x)}{\sum_{j=1}^V \exp(w_j^\top x)}\right], \tag{15}$$

where $W = [w_1, \dots, w_V]^\top \in \mathbb{R}^{V\times d}$.

Assumption 2 (Zipfian Class Distribution)**.** The class probabilities satisfy $p_j = \Pr(y = j) = j^{-\alpha}/H_{V,\alpha}$, where $\alpha > 0$ is the Zipf exponent and $H_{V,\alpha} = \sum_{j=1}^V j^{-\alpha}$. The features x are sub-Gaussian with second moment matrix $M_x := \mathbb{E}[xx^\top]$.

Assumption 3 (Local Class-Frequency Majorizer)**.** Assume $\|M_x\|_{\mathrm{op}} > 0$. The class-conditional moments satisfy $\mathbb{E}[xx^\top \mid y{=}j] \preceq C_0 M_x$ for all j. Moreover, there is a convex neighborhood $\mathcal{W}$ of a well-specified Bayes classifier W^* such that, for every $W \in \mathcal{W}$ and class j,

$$\mathbb{E}_x[\widehat{p}_j(W,x)xx^\top] \preceq 2C_0 p_j M_x, \tag{16}$$

where $\widehat{p}_j(W,x) = \mathrm{softmax}(Wx)_j$ and $M_x = \mathbb{E}[xx^\top]$.

The first condition controls feature anisotropy across classes. At a well-specified W^*, $\widehat{p}_j(W^*,x) = \Pr(y = j \mid x)$ and hence $\mathbb{E}[\widehat{p}_j(W^*,x)xx^\top] = p_j\mathbb{E}[xx^\top \mid y = j] \preceq C_0 p_j M_x$; continuity gives such a neighborhood, although the assumption does not quantify its radius.

Theorem 7 (Local Block Smoothness of Softmax with Zipfian Data)**.** *Under Assumptions 2 and 3, the softmax loss* (15) *is* $\mathbf{L}$*-block smooth on $\mathcal{W}$ with respect to the per-class decomposition $W = (w_1, \dots, w_V)$, with $L_j = 2C_0\|M_x\|_{\mathrm{op}} \cdot p_j$ (where C_0 is from Assumption 3). Consequently:*

$$\mathcal{R}_H = \frac{L_1}{L_V} = \frac{p_1}{p_V} = V^\alpha. \tag{17}$$

(The C_0 and $\|M_x\|_{\mathrm{op}}$ factors cancel in the ratio.)

Proof sketch (full proof in Appendix B.5). The categorical covariance obeys $\mathrm{diag}(\widehat{p}) - \widehat{p}\widehat{p}^\top \preceq \mathrm{diag}(\widehat{p})$. Consequently the full Hessian is bounded by the block-diagonal matrix whose jth block is $\mathbb{E}[\widehat{p}_j xx^\top]$. Assumption 3 then gives $L_j = 2C_0 p_j\|M_x\|_{\mathrm{op}}$, and Zipf's law gives $\mathcal{R}_H = p_1/p_V = V^\alpha$. □

Corollary 8 (Local Separation for Softmax Regression)**.** *For softmax regression with Zipfian vocabulary, the admissible local block majorizer in Theorem 7 has ratio V^α. Conditional on the initialization and every iterate remaining in $\mathcal{W}$, the GD upper bound has an $\mathcal{O}(V^\alpha)$ factor relative to Block-PGD. This conditional comparison requires each update segment to remain in $\mathcal{W}$, so that its local majorizer applies throughout the step.*

This is a local result and does not by itself prove that an optimization trajectory enters or remains in $\mathcal{W}$. Globally, $\mathrm{diag}(\widehat{p}) - \widehat{p}\widehat{p}^\top \preceq I$, so the common admissible choice $L_j = \|M_x\|_{\mathrm{op}}$ has ratio one and gives no separation. The frequency-sensitive majorizer arises only in the well-specified neighborhood. Section 7 tracks the curvature/frequency correlation during training as a diagnostic, not as a proof of trajectory containment.

5.2 Unification: Architecture and Data Heterogeneity

Our framework places two explanations in a common language but under different decompositions: architecture components can have different local Hessian scales (Zhang et al., 2024), while output-class blocks inherit frequency-dependent scales under Assumption 3. These ratios cannot simply be multiplied without a joint block majorizer; formalizing their interaction remains open.

6 Diagonal Preconditioning and a RMSProp-Style Consequence

Block-PGD assumes exact knowledge of L_k. We next give a sufficient condition for a general momentum-free diagonal method, followed by a weaker condition motivated by RMSProp. Neither result covers full Adam with momentum.

Theorem 9 (Diagonal-Preconditioned GD)**.** *Let f be $\mathbf{L}$-block smooth with gap Δ. Consider diagonal-preconditioned GD: $w_{t+1} = w_t - D_t \nabla f(w_t)$, where $D_t = \mathrm{diag}(d_{t,1}, \ldots, d_{t,d})$ with $d_{t,i} > 0$. If, for all t and all i in block k:*

$$\frac{c}{L_k} \le d_{t,i} \le \frac{1}{L_k}, \quad \textit{for a constant } 0 < c \le 1, \tag{18}$$

then $\min_{t<T} \mathcal{G}^2_{\mathbf{L}}(w_t) \le \frac{2\Delta}{cT}$.

Proof. By block smoothness with $\delta_i = -d_{t,i}\nabla_i f$:

$$f(w_{t+1}) \le f(w_t) - \sum_k \sum_{i \in k} d_{t,i} (\nabla_i f)^2 + \frac{1}{2} \sum_k L_k \sum_{i \in k} d_{t,i}^2 (\nabla_i f)^2. \tag{19}$$

Since $d_{t,i} \le 1/L_k$, we have $L_k d_{t,i}^2 \le d_{t,i}$, so $\frac{1}{2} L_k d_{t,i}^2 \le \frac{1}{2} d_{t,i}$. Thus:

$$f(w_{t+1}) \le f(w_t) - \frac{1}{2} \sum_k \sum_{i \in k} d_{t,i} (\nabla_i f)^2 \le f(w_t) - \frac{c}{2} \sum_k \frac{1}{L_k} \|\nabla_k f\|^2 = f(w_t) - \frac{c}{2} \mathcal{G}^2_{\mathbf{L}}(w_t). \tag{20}$$

Telescoping: $\min_t \mathcal{G}^2_{\mathbf{L}}(w_t) \le 2\Delta/(cT)$. □

Corollary 10 (RMSProp-Style Square-Root Preconditioning)**.** *Suppose positive second-moment estimates satisfy $aL_k \le v_{t,i} \le bL_k$ for every coordinate i in block k, for constants $0 < a \le b$. Consider the momentum-free update*

$$w_{t+1,i} = w_{t,i} - \frac{\eta}{\sqrt{v_{t,i}}} \nabla_i f(w_t), \qquad \eta = \sqrt{a/L_{\max}}. \tag{21}$$

Then

$$\min_{t<T} \mathcal{G}^2_{\mathbf{L}}(w_t) \le \frac{2\sqrt{b/a}\sqrt{\mathcal{R}_H}\Delta}{T}. \tag{22}$$

Proof. The effective step satisfies

$$\frac{\sqrt{a/b}}{\sqrt{\mathcal{R}_H} L_k} \le \frac{\eta}{\sqrt{v_{t,i}}} \le \frac{1}{L_k}. \tag{23}$$

Apply Theorem 9 with $c = \sqrt{a/b}/\sqrt{\mathcal{R}_H}$. □

Remark 3 (Scope and Softmax Interpretation)**.** *At a well-specified softmax optimum, the per-example gradient $g_j = (\widehat{p}_j - \mathbf{1}\{y = j\})x$ satisfies the Fisher identity*

$$\mathbb{E}[g_j g_j^\top] = \mathbb{E}[\widehat{p}_j(1 - \widehat{p}_j) x x^\top] = \nabla^2_{w_j} f. \tag{24}$$

Thus $v_{t,i} \asymp L_k$ is meaningful when coordinatewise second moments and the block spectral scale are comparable. It is still an explicit assumption, and our architecture-block diagnostics show that the stronger $1/L_k$ step bracket need not hold. Full Adam uses $D_t m_t$ rather than $D_t \nabla f(w_t)$; controlling its momentum buffer and evolving second moment is outside our results.

7 EXPERIMENTS

We first reproduce the controlled quadratic and Zipfian-softmax phenomena, then directly measure local curvature and adaptive steps on real text. The final diagnostics are intentionally reported even where they do not support the strongest possible connection to Adam.

7.1 SYNTHETIC: BLOCK-HETEROGENEOUS QUADRATIC

Setup. We construct $f(w_1, w_2) = f_1(w_1) + f_2(w_2)$ where each block is a quadratic $f_k(w) = \frac{L_k}{2}\|w - w_k^*\|^2$, with $w \in \mathbb{R}^{100}$ and w_k^* drawn from $\mathcal{N}(0, 4I)$, matching the construction in Theorem 4. We use $L_1 = 1$ and $L_2 \in \{1, 10, 100, 1000\}$ and compare constant-step GD ($\eta = 1/L_2$) and Block-PGD ($\eta_k = 1/L_k$) for $T = 3{,}000$ iterations with fixed random seed.

Results. Figure 2(a) shows $\mathcal{G}_{\mathbf{L}}^2$ vs. iteration count. Block-PGD converges at the same rate regardless of $\mathcal{R}_H$, while GD slows down proportionally to $\mathcal{R}_H = L_2/L_1$, confirming Theorem 4. Figure 2(b) shows the empirical iteration count ratio $T_{\text{GD}}/T_{\text{Block-PGD}}$ to reach a fixed target grows with $\mathcal{R}_H$.

7.2 SOFTMAX REGRESSION WITH ZIPFIAN CLASSES

Setup. Softmax regression on synthetic data: $d = 50$ features, $V \in \{100, 500, 2000\}$ classes, $n = 50{,}000$ samples. For each class j, $x \mid y{=}j \sim \mathcal{N}(\mu_j, 0.25 I_d)$ with fixed random centroids $\mu_j \sim \mathcal{N}(0, I_d)$. Class frequencies follow Zipf's law with $\alpha = 1$. Adam uses lr $= 10^{-3}$ with default (β_1, β_2); SGD lr is grid-searched over $\{1, 10^{-1}, 10^{-2}, 10^{-3}\}$. Batch size 256, 50 epochs, averaged over 3 seeds.

Results. Figure 3(a) shows that Adam reaches lower loss within the fixed budget and that the gap grows with V. Figure 3(b) reports the final-loss ratio; this is a qualitative finite-budget diagnostic, not a test of the iteration-complexity exponent. Figure 3(c) shows frequency-dependent per-class gradients, which is consistent with the Zipfian mechanism but is not a direct smoothness measurement.

7.3 DIRECT CURVATURE AND EFFECTIVE-STEP DIAGNOSTICS

Setup. We train a randomly initialized GPT-2 model (3 layers, 3 heads, hidden size 192, vocabulary 50,257) on WikiText-2 sequences of length 96. Adam uses lr 6×10^{-4}; SGD is tuned over $\{0.003, 0.01, 0.03, 0.1, 0.3\}$. We use three matched seeds. For each architecture block we estimate the spectral norm of the block-restricted local Hessian using eight Hessian–vector-product power iterations. For 64 output tokens sampled over log-spaced frequency ranks, we also compute the exact top eigenvalue of

$$\frac{1}{N}\sum_n \widehat{p}_j(h_n)(1 - \widehat{p}_j(h_n))h_n h_n^\top \tag{25}$$

with hidden features fixed, and compare it with Adam's bias-corrected second moment and effective step. GPT-2 ties its input and output embeddings, so the recorded second moment also includes the input-embedding role.

Results. At step 300, Adam and the best SGD configuration attain mean validation losses 6.756 and 7.486. They reach fixed loss 7.6 in 50 and 228 steps on average (4.6x). In a seed-0 detailed diagnostic, their ordinary squared gradient norms are 1.335 and 1.337, while their local weighted metrics are 0.352 and 0.266; hence the weighted metric does not monotonically track the loss advantage. The architecture-block HVP ratio is 1,428–3,517 across seeds at step 50 and 810–1,120 at step 2,000. However, architecture-block effective steps are negatively correlated with $1/L_k$ (seed-0 step-300 Spearman -0.13), so Theorem 9's bracket is not empirically verified on this decomposition.

For output-class blocks, the step-2,000 Spearman ranges are 0.964–0.967 for curvature versus token frequency, 0.971–0.976 for second moment versus curvature, and 0.972–0.976 for effective step versus $1/\sqrt{L}$. These rank associations are consistent with inverse-curvature scaling on the softmax decomposition, but cannot identify the square-root exponent in Corollary 10: $1/L$ and $1/\sqrt{L}$ induce identical ranks. At random initialization the curvature/frequency correlation is -0.10 in seed 0, rising to 0.964 by step 50; this trajectory is evidence for emergence of the local structure, not proof that the iterates remain in $\mathcal{W}$.

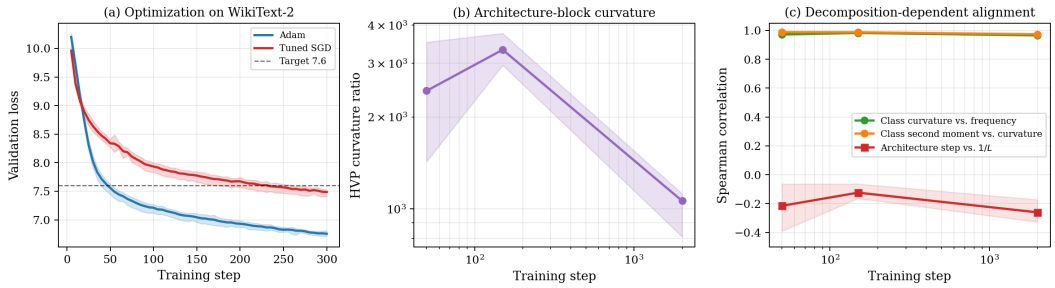

Figure 1: Direct WikiText-2 diagnostics (mean and seed range). **(a)** Adam versus tuned SGD validation loss. **(b)** Architecture-block local curvature ratios estimated by HVPs. **(c)** Output-class frequency/curvature and second-moment/curvature alignment are strong, while architecture-block effective steps do not align with $1/L_k$.

8 Discussion and Conclusion

Under a block quadratic majorizer, block preconditioning improves the convergence bound over a single global GD step. The upper-bound comparison differs by $\mathcal{R}_H$ in normalized stationarity, and explicit quadratics require linear heterogeneity dependence for GD under normalized stationarity, loss, and the ordinary gradient norm. The nonconvex accuracy dependence is not matched. A local Zipfian softmax majorizer and the output-class diagnostics connect the framework to language modeling, while the architecture-block diagnostics show that this connection depends critically on the chosen decomposition.

On the Stationarity Measure. $\mathcal{G}_{\mathbf{L}}^2$ is ordinary Euclidean stationarity in curvature-normalized coordinates and the decrease certificate of the block upper model. The general upper-bound separation is in this metric; converting it to $\|\nabla f\|$ can lose the advantage. Our quadratic lower bound supplies a standard-metric separation on one family, but no theorem says that $\mathcal{G}_{\mathbf{L}}$ must track training loss across practical optimizers. Indeed, Figure 1 shows an Adam loss advantage without a smaller measured weighted metric. Reporting loss, both stationarity measures, and direct curvature prevents conflating these distinct claims.

9 Limitations

First, Block-PGD assumes known block smoothness constants; Corollary 10 replaces this with an explicit second-moment/curvature comparability assumption rather than proving that an optimizer learns it. Second, full Adam is not analyzed: momentum changes the update direction, and our architecture-block measurements fail the stronger $1/L_k$ bracket. Third, the V^α result is local, assumes well-specification and controlled class-conditional moments, and applies to output-class blocks with hidden features treated as fixed. We neither quantify the radius of $\mathcal{W}$ nor prove trajectory containment. Fourth, HVP power iteration estimates local block-restricted spectral norms on a small model and finite batch; it is not a global smoothness certificate. Fifth, the quadratic lower bound establishes linear $\mathcal{R}_H$ dependence but does not match the nonconvex $1/\varepsilon^2$ dependence. Finally, our theory does not explain why SGD remains competitive for CNNs. Extending the analysis to momentum, learned time-varying preconditioners, joint attention/MLP/output geometry, and matching nonconvex lower bounds are open problems.

Reproducibility Statement

The supplement provides the synthetic optimization and language-model training scripts, saved numerical diagnostics, and analysis code. Additional experiments compare amplitude fits and conservative scale-estimation errors; their settings and complete outcomes are included. The local curvature measurements do not certify global smoothness constants.

AI Use Statement

OpenAI Codex was used to aid and polish manuscript writing and to assist with literature retrieval and discovery, including finding related work. The authors are responsible for the final manuscript and citations.

References

Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 1998. https://doi.org/10.1162/089976698300017746.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12:2121–2159, 2011.

Saeed Ghadimi and Guanghui Lan. Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. *SIAM Journal on Optimization*, 2013. https://doi.org/10.1137/120880811.

Vineet Gupta, Tomer Koren, Yoram Singer. Shampoo: Preconditioned Stochastic Tensor Optimization. *International Conference on Machine Learning*, 2018. https://proceedings.mlr.press/v80/gupta18a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

Ruichen Jiang, Devyani Maladkar, and Aryan Mokhtari. Provable complexity improvement of AdaGrad over SGD: Upper and lower bounds in stochastic non-convex optimization. In *COLT*, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between SGD and Adam on transformers, but sign descent might be. In *ICLR*, 2023.

Frederik Kunstner, Alan Milligan, Robin Yadav, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why Adam outperforms gradient descent on language models. In *NeurIPS*, 2024.

Xiaoyu Li, Francesco Orabona. On the Convergence of Stochastic Gradient Descent with Adaptive Stepsizes. *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019. https://proceedings.mlr.press/v89/li19c.html.

Bingrui Li, Wei Huang, Andi Han, Zhanpeng Zhou, Taiji Suzuki, Jun Zhu, and Jianfei Chen. On the optimization and generalization of two-layer transformers with sign gradient descent. In *ICLR*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *International Conference on Learning Representations*, 2019. https://openreview.net/forum?id=Bkg6RiCqY7.

James Martens, Roger Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. *International Conference on Machine Learning*, 2015. https://proceedings.mlr.press/v37/martens15.html.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Antonio Orvieto and Robert M. Gower. In search of Adam's secret sauce. In *NeurIPS*, 2025.

Yan Pan and Yuanzhi Li. Toward understanding why Adam converges faster than SGD for transformers. *arXiv preprint arXiv:2306.00204*, 2023.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the Convergence of Adam and Beyond. *International Conference on Learning Representations*, 2018. https://openreview.net/forum?id=ryQu7f-RZ.

Peter Richtárik and Martin Takáč. Iteration Complexity of Randomized Block-Coordinate Descent Methods for Minimizing a Composite Function. *Mathematical Programming*, 2014. https://doi.org/10.1007/s10107-012-0614-y.

Akiyoshi Tomihari and Issei Sato. Gradient heterogeneity complements Hessian heterogeneity in transformer optimization. *Transactions on Machine Learning Research*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Rachel Ward, Xiaoxia Wu, Leon Bottou. AdaGrad Stepsizes: Sharp Convergence Over Nonconvex Landscapes. *International Conference on Machine Learning*, 2019. https://proceedings.mlr.press/v97/ward19a.html.

Stephen J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

Robin Yadav, Shuo Xie, Tianhao Wang, and Zhiyuan Li. Provable benefit of sign descent: A minimal model under heavy-tailed class imbalance. *arXiv preprint arXiv:2512.00763*, 2025.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *ICLR*, 2020.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need Adam: A Hessian perspective. In *NeurIPS*, 2024.

Tianyue H. Zhang, Lucas Maes, Alan Milligan, Alexia Jolicoeur-Martineau, Ioannis Mitliagkas, Damien Scieur, Simon Lacoste-Julien, and Charles Guille-Escuret. Understanding Adam requires better rotation dependent assumptions. In *NeurIPS*, 2025.

A Additional Experimental Figures

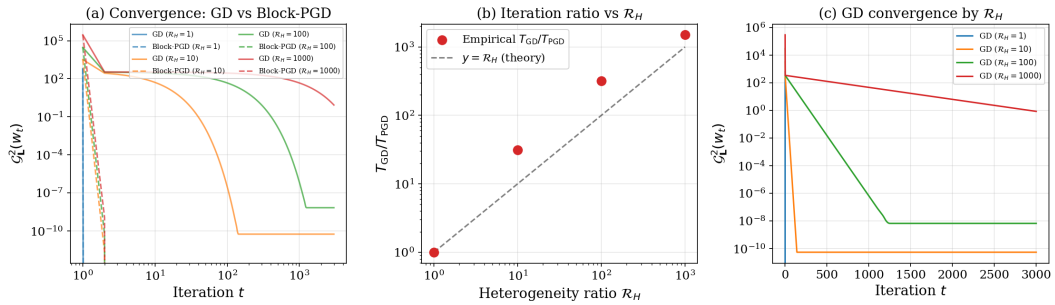

Figure 2: **(a)** $\mathcal{G}_{\mathbf{L}}^2$ vs. iterations for the separable quadratic hard instance. Block-PGD (dashed) converges in one step; constant-step GD (solid) slows proportionally to $\mathcal{R}_H$. **(b)** Empirical iteration ratio $T_{\text{GD}}/T_{\text{Block-PGD}}$ tracks $y = \mathcal{R}_H$. **(c)** GD convergence curves.

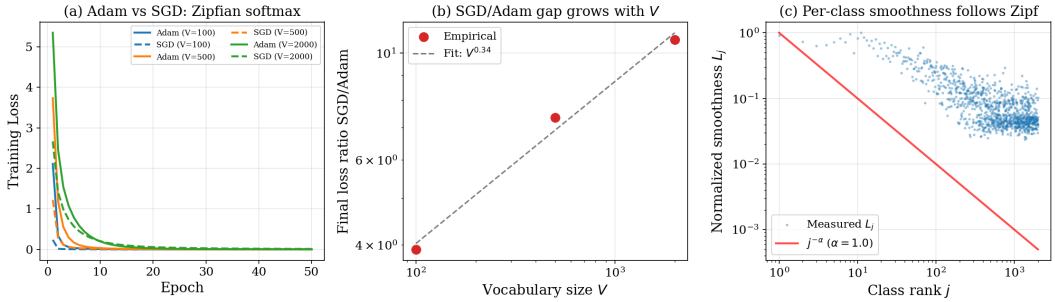

Figure 3: **(a)** Training loss for Zipfian softmax regression. **(b)** Finite-budget final-loss ratio (not an iteration-complexity estimate). **(c)** Per-class gradients exhibit frequency structure; direct curvature is measured in Figure 1.

B Deferred Proofs

B.1 Proof of Theorem 4 (GD Lower Bound)

We construct a separable instance $f(w_1, w_K) = f_1(w_1) + f_K(w_K)$.

Construction. Let $f_K(w_K) = \frac{L_K}{2} \|w_K\|^2$ with $w_{K,0}$ chosen so that $f_K(w_{K,0}) = \Delta/2$. This is exactly L_K-smooth and requires $\eta \le 2/L_K$ for GD convergence.

Let $f_1(w_1) = \frac{L_1}{2} \|w_1 - w_1^*\|^2$ with $w_{1,0} = 0$ and $\|w_1^*\|^2 = \Delta/L_1$, so $f_1(w_{1,0}) - \inf f_1 = \Delta/2$. This is exactly L_1-smooth and convex.

Upper bound direction. Follows from Theorem 3: $\min_t \mathcal{G}_{\mathbf{L}}^2(w_t) \le 2\mathcal{R}_H \Delta/T$.

Lower bound direction. Since f is separable, GD with constant step η decomposes as:

$$w_{1,t} - w_1^* = (1 - \eta L_1)^t (w_{1,0} - w_1^*), \qquad w_{K,t} = (1 - \eta L_K)^t w_{K,0}. \tag{26}$$

The weighted stationarity is:

$$\begin{aligned} \mathcal{G}_{\mathbf{L}}^2(w_t) &= \tfrac{1}{L_1} \|\nabla f_1(w_{1,t})\|^2 + \tfrac{1}{L_K} \|\nabla f_K(w_{K,t})\|^2 \\ &= L_1 (1 - \eta L_1)^{2t} \|w_{1,0} - w_1^*\|^2 + L_K (1 - \eta L_K)^{2t} \|w_{K,0}\|^2 . \end{aligned} \tag{27}$$

Substituting the initial conditions:

$$\mathcal{G}_{\mathbf{L}}^2(w_t) = \Delta\left[(1-\eta L_1)^{2t} + (1-\eta L_K)^{2t}\right]. \tag{28}$$

Since both terms are non-negative and the first term $(1-\eta L_1)^{2t}$ dominates (it decays slower because $\eta L_1 \leq \eta L_K$):

$$\mathcal{G}_{\mathbf{L}}^2(w_t) \geq \Delta(1-\eta L_1)^{2t}. \tag{29}$$

For $\mathcal{G}_{\mathbf{L}}^2(w_t) \leq \varepsilon^2$, we need $(1-\eta L_1)^{2t} \leq \varepsilon^2/\Delta$, i.e.:

$$t \geq \frac{\log(\Delta/\varepsilon^2)}{2\log(1/(1-\eta L_1))}. \tag{30}$$

Using $\log(1/(1-x)) \leq x/(1-x)$ for $x \in (0,1)$, which gives an upper bound on the denominator:

$$t \geq \frac{(1-\eta L_1)\log(\Delta/\varepsilon^2)}{2\eta L_1}. \tag{31}$$

With $\eta = 1/L_K$:

$$t \geq \frac{(1-L_1/L_K)}{2} \cdot \mathcal{R}_H \cdot \log\frac{\Delta}{\varepsilon^2} = \frac{\mathcal{R}_H - 1}{2}\log\frac{\Delta}{\varepsilon^2}. \tag{32}$$

For $\mathcal{R}_H \geq 2$, this gives $t \geq \frac{\mathcal{R}_H}{4}\log(\Delta/\varepsilon^2)$. This is a rigorous lower bound on the number of iterations for constant-step GD to achieve $\mathcal{G}_{\mathbf{L}}^2 \leq \varepsilon^2$ on this specific instance. In contrast, Block-PGD with $\eta_1 = 1/L_1$, $\eta_K = 1/L_K$ converges in one step on each quadratic block:

$$w_{1,1} = w_1^*, \quad w_{K,1} = 0 \quad \Rightarrow \quad \mathcal{G}_{\mathbf{L}}^2(w_1) = 0. \tag{33}$$

The gap between Block-PGD (converges in 1 step on quadratics) and constant-step GD (requires $\Omega(\mathcal{R}_H \log(1/\varepsilon))$ steps, with constant $1/4$) is $\Theta(\mathcal{R}_H)$ in the strongly convex/quadratic setting.

For completeness, the same construction yields standard-metric lower bounds. With $\eta = 1/L_K$, nonnegativity of the fast-block term gives

$$f(w_t) - f^* = \frac{\Delta}{2}\left[(1-1/\mathcal{R}_H)^{2t} + 0^{2t}\right] \geq \frac{\Delta}{2}(1-1/\mathcal{R}_H)^{2t}, \tag{34}$$

$$\|\nabla f(w_t)\|^2 = \Delta L_1(1-1/\mathcal{R}_H)^{2t} + \Delta L_K 0^{2t} \geq \Delta L_{\min}(1-1/\mathcal{R}_H)^{2t}. \tag{35}$$

Thus reaching loss at most ε or ordinary squared gradient at most ε^2 also requires $\Omega(\mathcal{R}_H \log(1/\varepsilon))$ iterations on this instance.

Implications for the non-convex setting. For general non-convex $\mathbf{L}$-block smooth functions, Theorems 3 and 2 give upper bounds differing by $\mathcal{R}_H$. Equation (28) rules out guarantees completely independent of $\mathcal{R}_H$, but does not prove that $\mathcal{R}_H$ must multiply $1/\varepsilon^2$. Whether the displayed nonconvex upper bound is jointly tight remains open.

B.2 Proof of Proposition 11

Proposition 11 (Anisotropic Blocks and Estimated Smoothness)**.** *Let $A_k = \mathrm{diag}(\theta L_k, L_k)$ with $0 < \theta < 1$ and $f(w) = \frac{1}{2}\sum_k w_k^\top A_k w_k$. Suppose $L_k \leq \widehat{L}_k \leq qL_k$. Block-PGD using $1/\widehat{L}_k$ has contraction at most $1-\theta/q$, whereas GD using $1/\widehat{L}_{\max}$ has contraction at least $1-\theta/\mathcal{R}_H$ along the slow direction of the $L_{\min}$ block. Their loss and ordinary-gradient complexities differ by $\Omega(\mathcal{R}_H/q)$; for equal relative estimation error, the ratio is $\Theta(\mathcal{R}_H)$.*

The eigenvalues of A_k are θL_k and L_k. With step $1/\widehat{L}_k$ and $L_k \leq \widehat{L}_k \leq qL_k$, both eigendirections contract by a nonnegative factor and the worst factor satisfies

$$\max_{\lambda \in \{\theta L_k, L_k\}} \left|1 - \frac{\lambda}{\widehat{L}_k}\right| \leq 1 - \frac{\theta}{q}. \tag{36}$$

Thus Block-PGD reaches either a loss or ordinary-gradient threshold in $O((q/\theta)\log(1/\varepsilon))$ iterations. For GD with step $1/\widehat{L}_{\max}$, initialize the slow eigendirection of an $L_{\min}$ block with nonzero gap. Its contraction factor is

$$1 - \frac{\theta L_{\min}}{\widehat{L}_{\max}} \geq 1 - \frac{\theta}{\mathcal{R}_H}, \tag{37}$$

so GD needs $\Omega((\mathcal{R}_H/\theta)\log(1/\varepsilon))$ iterations. Dividing the two scalings gives $\Omega(\mathcal{R}_H/q)$; if $\widehat{L}_k/L_k$ is common across blocks, both factors contain the same estimation multiplier and the ratio is $\Theta(\mathcal{R}_H)$.

B.3 Proof of Theorem 12 (Descent-Certificate Comparison)

Theorem 12 (One-Step Descent-Certificate Comparison)**.** *At every point w with $\nabla f(w) \neq 0$, the Block-PGD certificate is at least the best GD certificate from the same block quadratic upper model (both certificates vanish at stationary points, where the ratio is undefined):*

$$\frac{1}{2}\mathcal{G}^2_{\mathbf{L}}(w) \geq \max_{\eta>0} \eta\left(1 - \frac{\eta L_{\text{eff}}(w)}{2}\right)\|\nabla f(w)\|^2 = \frac{\|\nabla f(w)\|^2}{2L_{\text{eff}}(w)}, \tag{38}$$

where $L_{\text{eff}} = \sum_k L_k \|\nabla_k f\|^2 / \|\nabla f\|^2$. Moreover,

$$1 \leq \frac{\mathcal{G}^2_{\mathbf{L}}}{\|\nabla f\|^2 / L_{\text{eff}}} \leq \frac{(\mathcal{R}_H + 1)^2}{4\mathcal{R}_H}, \tag{39}$$

and both bounds are tight.

By the Cauchy–Schwarz inequality applied to vectors $(\sqrt{L_k}\,\|\nabla_k f\|)_k$ and $(\|\nabla_k f\|/\sqrt{L_k})_k$:

$$\|\nabla f\|^4 = \left(\sum_k \|\nabla_k f\|^2\right)^2 \leq \underbrace{\left(\sum_k L_k \|\nabla_k f\|^2\right)}_{=L_{\text{eff}}\|\nabla f\|^2} \cdot \underbrace{\left(\sum_k \frac{1}{L_k}\|\nabla_k f\|^2\right)}_{=\mathcal{G}^2_{\mathbf{L}}}. \tag{40}$$

Rearranging: $\|\nabla f\|^2 / L_{\text{eff}} \leq \mathcal{G}^2_{\mathbf{L}}$, which proves the lower bound of 1.

For the upper bound, let $\alpha_k = \|\nabla_k f\|^2 / \|\nabla f\|^2 \in [0,1]$ with $\sum_k \alpha_k = 1$. Then:

$$\frac{\mathcal{G}^2_{\mathbf{L}}}{\|\nabla f\|^2 / L_{\text{eff}}} = \left(\sum_k \frac{\alpha_k}{L_k}\right)\left(\sum_k \alpha_k L_k\right). \tag{41}$$

By the Kantorovich inequality, for positive weights α_k summing to 1 and positive values $L_k \in [L_{\min}, L_{\max}]$:

$$\left(\sum_k \frac{\alpha_k}{L_k}\right)\left(\sum_k \alpha_k L_k\right) \leq \frac{(L_{\max} + L_{\min})^2}{4L_{\max}L_{\min}} = \frac{(\mathcal{R}_H + 1)^2}{4\mathcal{R}_H}. \tag{42}$$

This is tight when $K = 2$ with $L_1 = L_{\min}$, $L_2 = L_{\max}$, and $\alpha_1 = \alpha_2 = 1/2$.

Tightness. Lower bound: L_k constant $\Rightarrow$ ratio $= 1$. Upper bound: $K = 2$, $L_1 = 1$, $L_2 = \mathcal{R}_H$, $\|\nabla_1 f\| = \|\nabla_2 f\| = g > 0$ gives ratio exactly $(\mathcal{R}_H + 1)^2/(4\mathcal{R}_H)$.

B.4 Proof of Theorem 6 (Stochastic Separation)

Part (a). With stochastic gradients $\hat{g}_{t,k} = \nabla_k f(w_t) + \xi_{t,k}$ and step $\eta_k = 1/L_k$:

$$\begin{aligned}
\mathbb{E}[f(w_{t+1})] &\leq f(w_t) + \sum_k \mathbb{E}\left[\left\langle \nabla_k f, -\frac{1}{L_k}\hat{g}_{t,k}\right\rangle\right] + \frac{1}{2}\sum_k \frac{1}{L_k}\mathbb{E}\left[\|\hat{g}_{t,k}\|^2\right] \\
&= f(w_t) - \sum_k \frac{1}{L_k}\|\nabla_k f\|^2 + \frac{1}{2}\sum_k \frac{1}{L_k}\left(\|\nabla_k f\|^2 + \sigma_k^2\right) \\
&= f(w_t) - \frac{1}{2}\mathcal{G}^2_{\mathbf{L}}(w_t) + \frac{1}{2}\sum_k \frac{\sigma_k^2}{L_k}.
\end{aligned} \tag{43}$$

Summing and rearranging yields the result.

Part (b). With step $\eta = 1/L_{\max}$:

$$\mathbb{E}[f(w_{t+1})] \leq f(w_t) - \frac{1}{2L_{\max}}\|\nabla f\|^2 + \frac{1}{2L^2_{\max}}\sum_k L_k \sigma_k^2. \tag{44}$$

Summing: $\frac{1}{T}\sum_t \|\nabla f\|^2 \leq 2L_{\max}\Delta/T + \frac{1}{L_{\max}}\sum_k L_k\sigma_k^2$. Converting via $\mathcal{G}^2_{\mathbf{L}} \leq \frac{1}{L_{\min}}\|\nabla f\|^2$ gives the stated bound.

B.5 Proof of Theorem 7 (Softmax + Zipf)

The Hessian of the softmax cross-entropy loss with respect to the full parameter matrix W is, for a single input x:

$$\nabla^2 f(W) = \left(\mathrm{diag}(\hat{p}) - \hat{p}\hat{p}^\top\right) \otimes xx^\top, \tag{45}$$

where $\hat{p} = \mathrm{softmax}(Wx) \in \mathbb{R}^V$ and $\otimes$ denotes the Kronecker product.

For any block perturbation $\delta = (\delta_1, \ldots, \delta_V) \in \mathbb{R}^{V \times d}$, the quadratic form evaluated at a single x is:

$$\delta^\top \nabla^2 f \cdot \delta = \sum_{j=1}^{V} \hat{p}_j (\delta_j^\top x)^2 - \left(\sum_{j=1}^{V} \hat{p}_j \delta_j^\top x \right)^2. \tag{46}$$

Since $\left(\sum_j \hat{p}_j \delta_j^\top x\right)^2 \geq 0$, we immediately obtain:

$$\delta^\top \nabla^2 f \cdot \delta \leq \sum_{j=1}^{V} \hat{p}_j (\delta_j^\top x)^2. \tag{47}$$

Taking expectations in (47) and applying Assumption 3 gives

$$\delta^\top \nabla^2 f(W) \delta \leq \sum_j \delta_j^\top \mathbb{E}[\hat{p}_j(W, x) xx^\top] \delta_j \leq \sum_j 2C_0 p_j \left\|M_x\right\|_{\mathrm{op}} \left\|\delta_j\right\|^2. \tag{48}$$

Integrating this Hessian bound along any line segment contained in $\mathcal{W}$ yields the local block-smoothness inequality with $L_j = 2C_0 p_j \left\|M_x\right\|_{\mathrm{op}}$.

The heterogeneity ratio under Zipf's law:

$$\mathcal{R}_H = \frac{L_1}{L_V} = \frac{p_1}{p_V} = \frac{1^{-\alpha}/H_{V,\alpha}}{V^{-\alpha}/H_{V,\alpha}} = V^\alpha. \tag{49}$$

B.6 Strongly Convex Case

In the strongly convex setting, our separable quadratic construction gives the exact gap: Block-PGD converges in one step, while constant-step GD with $\eta = 1/L_K$ requires $\Omega(\mathcal{R}_H \log(1/\varepsilon))$ iterations. This gap appears not only in $\mathcal{G}_{\mathbf{L}}$ but also in function suboptimality and the ordinary gradient norm, as shown above. More generally, standard GD has contraction $1 - \mu/L_{\max}$; on this construction $\mu = L_{\min}$, so the condition number is precisely $\mathcal{R}_H$.

B.7 Discussion of Adam with Momentum

A complete convergence analysis of Adam under block-heterogeneous smoothness is open. Its preconditioner depends on gradient history and its update uses m_t, not $\nabla f(w_t)$. Corollary 10 therefore applies only to a momentum-free update under explicit second-moment bounds. The direct diagnostics further show that alignment is decomposition dependent: it is strong for output-class blocks but absent for the coarse architecture blocks we measured. We make no convergence-rate claim for full Adam.

C SENSITIVITY TO ESTIMATED BLOCK SCALES

Proposition 13 (Multiplicative scale error)**.** *Under Definition 1, take the deterministic update* $w_{t+1,k} = w_{t,k} - \nabla_k f(w_t)/\widehat{L}_{t,k}$. *Suppose* $0 < a \le \widehat{L}_{t,k}/L_k \le b < \infty$ *for every* t, k, *with* $a > 1/2$. *Define* $\kappa = \min\{2/a - 1/a^2,\ 2/b - 1/b^2\} > 0$. *Then* $\min_{t<T} \mathcal{G}^2_{\mathbf{L}}(w_t) \le 2\Delta/(\kappa T)$.

Proof. Set $\rho_{t,k} = L_k/\widehat{L}_{t,k} \in [1/b, 1/a]$. Substitution in the upper model gives a decrease of at least $\frac{1}{2}\sum_k \rho_{t,k}(2-\rho_{t,k})\|\nabla_k f(w_t)\|^2/L_k$. The concave quadratic $\rho(2-\rho)$ attains its minimum on this interval at an endpoint, yielding $\kappa\mathcal{G}^2_{\mathbf{L}}/2$. Sum over iterations and use the lower bound on f. □

This calculation makes the cost of inaccurate scales explicit. Underestimating a valid scale by a factor of two reaches the stability boundary even on a scalar quadratic. Sampled Hessian eigenvalues do not certify the required uniform inequalities; the result neither assumes nor proves that Adam's second moments satisfy them. The block partition and majorizer must be fixed consistently when comparing the diagnostic with this guarantee.

D ADDITIONAL-SEED OPTIMIZER DIAGNOSTICS

We repeat the three-layer, width-192 WikiText-2 diagnostic with two additional seeds (20260920 and 20260921), 300 training steps, and six HVP iterations. Adam and RMSProp use learning rate 6×10^{-4}; SGD uses 0.1. These rates are fixed from the earlier diagnostic, with no new sweep. Each seed fixes an eight-sequence validation batch of length 96; this is a small-batch diagnostic, not a full-validation-set benchmark. The two new seeds are reported separately from the previous runs. Adam's lower loss in these runs does not establish the scale-estimation premise of Appendix C.

Table 1: Validation-batch loss at step 300 in the additional-seed diagnostic.

Optimizer	Seed 20260920	Seed 20260921	Mean
ADAM	6.5458	6.7471	6.6465
RMSPROP	7.0204	7.0559	7.0381
SGD	7.2695	7.3533	7.3114

Its architecture-block Spearman correlations between median effective step and inverse estimated curvature are -0.118 and -0.127. In the distinct output-class partition, second-moment/curvature correlations are 0.986 and 0.981. Rank correlations with inverse curvature and inverse square-root curvature cannot distinguish these exponents, because both transformations preserve the same ordering. Thus the output-class correlation supports a local association, not identification of a particular preconditioning power or a uniform global scale guarantee.

E AMPLITUDE COMPARISON FOR THE OUTPUT-CLASS DIAGNOSTIC

The inverse-curvature and inverse-square-root predictors have identical ranks, so rank correlations cannot distinguish them. We instead regress log effective step against log empirical curvature across the 58 output classes measured at step 300. Each fixed-exponent comparison fits its own intercept; the reported error is root mean squared log residual. These are descriptive within-run comparisons, not tests over independent classes or a certification of a global smoothness assumption. The square-root

Table 2: Output-class amplitude fits. Seed labels 1 and 2 denote 20260920 and 20260921.

Optimizer	Seed	Fitted slope	RMSE: $-1/2$	RMSE: -1
Adam	1	-0.591	0.609	1.436
Adam	2	-0.663	0.803	1.227
Rmsprop	1	-0.625	0.819	1.486
Rmsprop	2	-0.745	1.076	1.100

predictor has lower residual error in these four runs, but fitted slopes range from -0.591 to -0.745. In the second RMSProp run the two fixed-exponent errors are particularly close. This supports a limited descriptive comparison, rather than an exact square-root law or a causal account of the observed optimization advantage. The empirical curvature diagnostic and effective steps both depend on the training trajectory, and the classes share a model and dataset.

F SENSITIVITY TO CONSERVATIVE SCALE ESTIMATION

We supplement the exact-scale quadratic example with 360 finite-dimensional cases: eight two-dimensional blocks, L_k geometrically spaced from 1 to $\mathcal{R}_H \in \{10, 100, 1000\}$, anisotropy $\theta \in \{0.05, 0.2, 0.8\}$, and ten fixed random draws of $\widehat{L}_k/L_k \in [1, q]$ for $q \in \{1, 2, 4, 8\}$. Each eigendirection starts with equal loss. Exact contraction factors give the first integer iteration reducing loss or squared ordinary gradient norm to 10^{-6} of its own initial value; no iterative stopping approximation is used. These constructed cases test sensitivity within the supplied majorizer model, not estimation of neural-network smoothness. Conservative estimation preserves the contraction

Table 3: Median GD/Block-PGD iteration ratio across the 30 anisotropy–seed combinations in each cell. Columns report loss / squared-gradient thresholds.

q	$\mathcal{R}_H = 10$	$\mathcal{R}_H = 100$	$\mathcal{R}_H = 1000$
1	9.1 / 7.9	91.9 / 62.0	920.2 / 430.6
2	7.9 / 6.9	81.2 / 53.1	813.9 / 359.7
4	6.5 / 5.5	61.3 / 44.0	613.9 / 317.6
8	6.3 / 5.3	52.8 / 38.3	527.8 / 284.5

bracket in every generated case, while its finite-time cost depends on the realized errors. This numerical comparison is not a new universal complexity bound. As an explicit boundary check, setting $\widehat{L}_k = L_k/4$ gives contraction factor -3 in a high-curvature direction and divergence. Such underestimation violates the sufficient condition; the guarantee should not be used when an estimated scale can fall arbitrarily below its true value.