ADAM AS MAGNITUDE-WEIGHTED SIGN DESCENT: DIMENSION-FREE CONVERGENCE UNDER WEAK ASSUMPTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop a convergence analysis of fully bias-corrected Adam through its *magnitude-weighted stochastic sign* representation. The key step controls the accumulated wrong-sign loss while retaining the joint randomness of Adam's update magnitude, momentum, and gradient noise. Under L-smoothness and conditional p-affine second moments, we prove that the expected average ℓ_2 gradient norm is at most $CT^{-1/4}$, where C has no explicit dimension factor and the rate has no extra $\log T$; under global point-pair (L_0, L_1, q)-smoothness, a stopped analysis yields the same exponent with high probability. We further instantiate the theory for fan-in-scaled LayerNorm/RMSNorm networks with fixed depth and increasing hidden width. For a concrete fixed-affine family, explicit width–horizon conditions guarantee that the Adam iterates remain in a width-uniform parameter region throughout training, yielding a convergence constant independent of the total parameter count. To our knowledge, this is the first Adam convergence-rate result establishing dimension-free optimization for an explicit growing family of normalized networks.

1 INTRODUCTION

Adam (Kingma & Ba, 2015) has established itself as the predominant optimizer for training Transformers (Vaswani et al., 2017), serving as the engine behind state-of-the-art Large Language Models (LLMs) (Brown et al., 2020; Touvron et al., 2023; Agarwal et al., 2025; Guo et al., 2025) and Large Vision Models (Radford et al., 2021; Kirillov et al., 2023; Siméoni et al., 2025). Notably, Adam's utility extends beyond Transformers to modern Convolutional Neural Networks (CNNs), such as ConvNeXt (Liu et al., 2022; Woo et al., 2023), where it has become the *de facto* choice. This marks a shift from the traditional preference for Stochastic Gradient Descent (SGD) (Krizhevsky et al., 2017; He et al., 2016), which was historically considered the gold standard for CNN training.

However, the theoretical convergence analysis of Adam lags significantly behind its practical success. The original proof provided by Kingma & Ba (2015) relied on convexity assumptions but was later shown to be flawed (Reddi et al., 2018). While AMSGrad (Reddi et al., 2018) was proposed to rectify this, its analysis still hinged on convexity. Chen et al. (2018) were the first to demonstrate that a class of Adam-type optimizers (including AMSGrad and AdaFom) converge to stationary points in non-convex settings. Subsequently, Défossez et al. (2020) provided a simplified convergence analysis for vanilla Adam and AdaGrad; however, their proof required $\beta_1 < \beta_2$ and the resulting bounds depended explicitly on the model dimensionality d. Similarly, Chen et al. (2022) introduced practical conditions for global convergence, yet their rates also suffered from dimensionality dependence.

Furthermore, most prior analyses (Chen et al., 2018; Défossez et al., 2020; Chen et al., 2022) assumed bounded stochastic gradients and uniform L-smoothness. Recent studies indicate that L-smoothness is inadequate for complex architectures like LSTMs and Transformers (Zhang et al., 2019; Crawshaw et al., 2022), advocating instead for the relaxed non-uniform (L_0, L_1)-smoothness condition. While Zhang et al. (2022) and Wang et al. (2023c) analyzed Random-Reshuffling Adam under weaker affine variance and (L_0, L_1)-smoothness assumptions, their results were bound by epoch complexity and depended on the total sample size. Li et al. (2023) give an Adam bound with no explicit d, under bounded or sub-Gaussian noise. Our sign-descent analysis instead uses conditional p-affine

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.

Table 1: Comparison of Adam and Adam-type convergence analyses. FCT denotes full bias correction. T counts iterations, E epochs, n samples, and d parameters. Rates are at the ℓ_2 gradient-norm scale unless indicated otherwise, with problem parameters and confidence fixed. Generalized-smoothness definitions are distinguished below.

References	FCT	Noise condition	Smoothness	Guarantee	Conv. rate
(Chen et al., 2018)	No	Bounded Grad.	L-smooth	Expectation	$\mathcal{O}\left(\frac{d^{1/2}\ln T}{T^{1/4}}\right)$
(Défossez et al., 2020)	No	Bounded Grad.	L-smooth	Expectation	$\mathcal{O}\left(\frac{d^{1/2}\ln(T)}{T^{1/4}}\right)$
(Chen et al., 2022)	No	Bounded Grad.	L-smooth	Expectation	$\mathcal{O}\left(\frac{d^{3/4}\ln T}{T^{1/4}}\right)$
(Zhang et al., 2022)	No	Finite-sum Affine Var.	L-smooth	Expectation	$\mathcal{O}\left(\frac{n^{1/2}d^{3/4}\ln E}{E^{1/4}}\right)$
(Wang et al., 2023c)	No	Bounded Var.	(L_0, L_1)-smooth	Expectation	$\mathcal{O}\left(\frac{n^{1/2}d^{1/2}\ln E}{E^{1/4}}\right)$
(Li et al., 2023)	Yes	Sub-Gaussian noise	(L_0, L_1, q) Hessian/local	High probability	$\mathcal{O}\left(\frac{(\ln T)^{13/16}}{T^{1/4}}\right)$
(Hong & Lin, 2025)	Yes	Affine Variance	(L_0, L_1, q)-smooth	High probability	$\mathcal{O}\left(\frac{d\ln T}{T^{1/4}}\right)$
(Li et al., 2025)	No	Bounded Var.	L-smooth	Expectation	$\mathcal{O}(\frac{d^{1/2}}{T^{1/4}})$ in ℓ_1-norm
(Jin et al., 2026)	No	Bounded Var.	L-smooth	High probability	$\mathcal{O}(\frac{d^{3/4}}{T^{1/4}})$
Corollary 4.3	Yes	p-affine variance	L-smooth	Expectation	$\mathcal{O}(\frac{1}{T^{1/4}})$; no explicit d
Corollary 4.4	Yes	p-affine variance	(L_0, L_1, q) global all-pairs	High probability	$\mathcal{O}(\frac{1}{T^{1/4}})$; no explicit d

Li et al.'s sub-Gaussian result is listed for its broader noise coverage, converting their $O(\varepsilon_{\text{stat}}^{-4}\log^{13/4}(1/\varepsilon_{\text{stat}}))$ complexity to the norm scale; their bounded-noise result has no extra logarithmic factor. For Jin et al. (v8, Theorem 3), their squared-gradient bound is converted by Cauchy–Schwarz. Algorithmic differences are discussed in Section 2.

second moments and yields a direct expected ℓ_2 bound under L-smoothness. A separate high-probability extension uses global point-pair B.3. Both have rate $\mathcal{O}(CT^{-1/4})$ without an extra $\log T$. Section 5 then bounds the full problem-dependent constant as hidden widths grow in a specified normalized-network family. Table 1 compares the settings.

A common approach treats Adam as preconditioned SGD with momentum, as originally described in Kingma & Ba (2015):

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{\gamma}{\sqrt{\hat{\boldsymbol{v}}_t} + \epsilon} \circ \hat{\boldsymbol{m}}_t.$$

The effective learning rate is not necessarily monotone, and the first and second moments depend on the same gradient history. A convergence proof must account for this dependence. In contrast, accumulating empirical evidence suggests that Adam's effectiveness may stem primarily from its *sign-like* properties (Balles & Hennig, 2018). Kunstner et al. (2023) demonstrated that sign descent with momentum achieves performance comparable to Adam on Transformers. Similarly, Chen et al. (2023b) discovered Lion, an optimizer resembling signSGD via AutoML, which outperforms Adam across various benchmarks. Kunstner et al. (2024) attributed Adam's superior performance on language models to its sign-like behavior, noting its advantages in handling heavy-tailed class imbalance. Furthermore, the recent Muon optimizer (Jordan et al., 2024), which can be treated as an extended matrix-sign algorithm (Chen et al., 2025), has shown significant potential for LLM training (Liu et al., 2025; Shah et al., 2025). This empirical evidence motivates an analysis that tracks the update magnitude and the cost of incorrect signs together.

We use Adam's equivalent representation

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \gamma \underbrace{\frac{|\hat{\boldsymbol{m}}_t|}{\sqrt{\hat{\boldsymbol{v}}_t} + \epsilon}}_{\boldsymbol{u}_t} \circ \mathrm{Sign}(\boldsymbol{m}_t).$$

The sign–magnitude interpretation is known (Balles & Hennig, 2018); our contribution is to turn it into a convergence analysis of fully bias-corrected Adam. The realized magnitude $\boldsymbol{u}_t$ remains coupled to the momentum and gradient noise throughout the analysis.

- **Adam convergence through weighted sign descent.** We control the accumulated magnitude-weighted wrong-sign loss by first-order momentum tracking. A joint bound on this loss, update magnitudes, and gradient drift absorbs the error into the ideal sign descent, without assuming their independence or bounded stochastic gradients.

- **Direct ℓ_2 guarantees under second-moment noise.** Under L-smoothness and conditional p-affine second moments, the expected average gradient norm has rate $\mathcal{O}(T^{-1/4})$ with no explicit d or extra $\log T$. The analysis retains full bias correction and allows any fixed $\beta_2 \in [0, 1)$ without coupling it to β_1. A stopped argument gives the same exponent under (L_0, L_1, q)-smoothness in high probability.

- **Dimension-free convergence for a growing network family.** For fan-in-scaled LayerNorm/RMSNorm networks with fixed depth and increasing hidden width, we establish width-uniform bounds for all problem-dependent constants in the convergence rate. For a fixed-affine subfamily, an explicit width–horizon condition guarantees that the Adam iterates remain in the width-uniform region throughout training, so the complete $O(T^{-1/4})$ constant is independent of parameter count. To our knowledge, this is the first Adam convergence-rate theorem establishing dimension-free optimization for an explicit growing normalized-network family. The conclusion is qualitatively consistent with empirical large-model scaling (Touvron et al., 2023; Bai et al., 2023; Kaplan et al., 2020), without claiming that widening always accelerates training.

2 Related Work

A significant body of literature addresses the theoretical analysis of stochastic descent algorithms. In this section, we review the references most relevant to our study and distinguish our contributions from existing work.

Convergence with Relaxed Assumptions. Bertsekas & Tsitsiklis (2000) provided the first theoretical analysis of SGD under the assumption of affine variance, deriving an asymptotic convergence result. Much later, Bottou et al. (2018) established the non-asymptotic convergence rate of SGD for the gradient norm $\|\nabla F(\boldsymbol{x})\|_2$, achieving a rate of $\tilde{\mathcal{O}}(T^{-1/4})$, which matches the provable rate under the standard bounded variance condition.

Regarding adaptive optimizers, Faw et al. (2022) investigated AdaGrad-Norm under affine variance, proving it could achieve $\tilde{\mathcal{O}}(T^{-1/4})$ provided that $\sigma_1 = \mathcal{O}(T^{-1/2})$. Subsequently, Wang et al. (2023a) showed that AdaGrad-Norm achieves a similar rate without restrictions on σ_1, and further demonstrated that vanilla AdaGrad could achieve the same rate under the stronger assumption of coordinate-wise affine variance. Meanwhile, Attia & Koren (2023) provided a probabilistic convergence rate for AdaGrad-Norm. It is worth noting that Shi et al. (2021) and Zhang et al. (2022) proved that Random-Shuffled AMSProp and Adam, respectively, converge to a neighborhood of stationary points under affine growth conditions (equivalent to affine variance). These are finite-sum, epoch-based results, and the general affine-growth guarantee may include a noise-dependent neighborhood. They are not directly comparable to iteration-based expectation bounds with independently sampled gradients. Recently, Hong & Lin (2023) analyzed vanilla Adam from a high-probability perspective, but their results rely on the stronger coordinate-wise affine variance assumption.

Regarding smoothness, Zhang et al. (2019) first introduced the non-uniform (L_0, L_1)-smoothness condition to explain the acceleration of Clipped-SGD, empirically verifying that local smoothness varies with the gradient norm during DNN training. Zhang et al. (2020) posited that this is equivalent to an affine form of the gradient norm for first-order differentiable functions. This relaxed assumption was subsequently extended to analyses of Clipping-SGD with momentum (Zhang et al., 2020), Distributionally Robust Optimization (Jin et al., 2021), Normalized SGDM (Hübler et al., 2024), and generalized signSGD (Crawshaw et al., 2022). Recently, Wang et al. (2023c) analyzed Random-Shuffled Adam under this condition, but similar to Zhang et al. (2022), their convergence rate is bound by epoch complexity $\tilde{\mathcal{O}}(E^{-1/4})$. Li et al. (2023) analyze Adam under Hessian-type generalized smoothness. Their Theorem 4.1 assumes almost-surely bounded noise and has no extra logarithmic factor; Theorem 4.2 allows the weaker sub-Gaussian norm condition, with logarithmic factors. Table 1 reports the latter for its broader noise coverage. Our C.3 requires only conditional second moments,

but our B.3 is a global all-pairs condition rather than their Hessian/local condition. The improvement in noise assumptions should therefore not be read as a uniform relaxation of every assumption.

Some recent works have attempted to combine these assumptions. Faw et al. (2023); Wang et al. (2023b) analyzed AdaGrad under both affine variance and non-uniform (L_0, L_1)-smoothness, obtaining a rate of $\tilde{\mathcal{O}}(T^{-1/4})$. Wang et al. (2024) provided a high-probability convergence proof for a simplified Adam-Norm under similar assumptions. Most recently, Hong & Lin (2025) established a convergence rate for vanilla Adam under both affine variance and (L_0, L_1)-smoothness; their stated rate retains explicit dependence on the model dimensionality d.

Jin et al. (2026) obtain a high-probability bound under L-smoothness and bounded variance, with sharper confidence dependence: $\widetilde{\mathcal{O}}(\rho^{-1/2}T^{-1/2})$ for the average squared gradient norm. Their formal bound has no extra $\log T$ at fixed confidence, but retains dimension-dependent constants. It uses no bias correction, positive second-moment initialization, and $\beta_2 = 1 - 1/T$. Our result instead gives a direct expectation bound under p-affine second moments, with full bias correction and arbitrary fixed β_2. These are different guarantees; we do not claim sharper confidence dependence.

Convergence of Sign Descent. Sign-based algorithms, which exploit only the sign of gradients, date back to Rprop (Riedmiller & Braun, 1993). Seide et al. (2014); Ström (2015) proposed 1-bit SGD, demonstrating strong empirical performance with reduced communication costs. The non-stochastic convergence of signSGD was first analyzed by Karimi et al. (2016) under the Polyak-Łojasiewicz condition. Bernstein et al. (2018) later systematically established the convergence rate of signSGD in stochastic non-convex scenarios, though it required a batch size increasing to $\mathcal{O}(\sqrt{n})$ (where n is sample size). Sun et al. (2023) subsequently proved that momentum ensures signSGD convergence without increasing batch size. More recently, Chen et al. (2023b) employed AutoML to discover Lion, an effective optimizer resembling signSGD with momentum, which outperforms Adam on various tasks. While Chen et al. (2023a) theoretically analyzed Lion's efficacy, they did not provide a full convergence proof. Historically, Adam and RMSProp (Hinton et al., 2012) evolved from sign-based Rprop, and Balles & Hennig (2018) noted the deep connection between sign descent and Adam. Kunstner et al. (2023; 2024) empirically showed that Adam's sign-like property is a primary driver of its performance in DNNs and heavy-tailed scenarios. Recently, Muon (Jordan et al., 2024), an extended matrix-sign optimizer (Chen et al., 2025), has shown significant potential for training LLMs (Liu et al., 2025; Shah et al., 2025). Balles & Hennig (2018) use the sign–magnitude view to interpret Adam, while Sun et al. (2023) give a descent bound for diagonally scaled sign updates in their signSGD analysis. Neither result alone controls Adam's cumulative error: its diagonal magnitude depends on the same gradient history as the momentum error. We develop the coupled weighted-error estimates needed to turn this representation into a convergence proof for Algorithm 1.

Dimension dependence. Li et al. (2025) prove an expected $\mathcal{O}(\sqrt{d}\,T^{-1/4})$ ℓ_1-stationarity bound for AdamW under L-smoothness and bounded coordinate-wise variance. Although $\|g\|_2 \le \|g\|_1 \le \sqrt{d}\|g\|_2$, this ℓ_1 upper bound implies only an $\mathcal{O}(\sqrt{d}\,T^{-1/4})$ ℓ_2 upper bound: the second inequality cannot be reversed to remove $\sqrt{d}$. It therefore does not, by itself, establish a dimension-free $\mathcal{O}(T^{-1/4})$ ℓ_2 guarantee. Li et al. further provide empirical evidence that, for some neural networks, $\|\nabla f(x_k)\|_1 = \Theta(\sqrt{d})\|\nabla f(x_k)\|_2$, offering useful practical motivation for interpreting their ℓ_1 result. Since this relation is observed for specific architectures and experimental settings rather than established under general theoretical assumptions, whether it extends to broader classes of objectives and networks remains an interesting open question. Their analysis omits bias correction and couples the stability parameter to the step size; our contribution is a direct ℓ_2 analysis with full bias correction and gradient-dependent second-moment noise bounds, not a faster iteration exponent. Li et al. (2023) obtain an Adam rate with no explicit factor of d, but their bound retains problem-dependent constants such as generalized-smoothness coefficients L_0, L_1, a noise scale σ, and the initial gap Δ. They do not determine how these constants scale across a growing family of normalized DNNs. Thus, a bound with *no explicit dimension dependence* does not by itself establish *dimension-free convergence under model scaling*. For increasing hidden widths at fixed depth and output dimension, Theorem 5.1 constructs admissible $L_0, L_1, \sigma_0, \sigma_1$ and bounds Δ independently of the parameter count. For the concrete fixed-affine normalized-network family in Theorem 5.2, explicit width–horizon conditions guarantee that the Adam iterates remain in the width-uniform region and that the resulting convergence constant is independent of parameter count.

3 PRELIMINARY

3.1 NOTATION

Throughout this paper, we use boldface letters to denote vectors, *i.e.*, $\boldsymbol{x} \in \mathbb{R}^d$, with $\boldsymbol{x}^{(j)}$ representing the j-th coordinate of the vector $\boldsymbol{x}$. $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the ℓ_1-norm and ℓ_2-norm, respectively. $\circ$ denotes the Hadamard product (element-wise product). $\mathbb{E}_{\xi_t}[\cdot]$ is the expectation operator at the time t in which the random variable ξ_t is independently and identically sampled from the dataset, and $\mathbb{E}[\cdot] = \mathbb{E}_{\xi_1}\mathbb{E}_{\xi_2}...\mathbb{E}_{\xi_t}[\cdot]$ is the total expectation operator.

For adaptive iterates, let $\mathscr{F}_{t-1} = \sigma(\xi_1, \ldots, \xi_{t-1})$, with $\boldsymbol{x}_1$ deterministic. We interpret $\mathbb{E}_{\xi_t}[\cdot]$ as $\mathbb{E}[\cdot \mid \mathscr{F}_{t-1}]$, and set $\mathrm{Sign}(0) = 0$. All vector quotients below are coordinate-wise.

In this paper, the optimizer aims to minimize the empirical risk loss of a model on a dataset, *i.e.*,

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) = \mathbb{E}_{\xi}[f(\boldsymbol{x};\xi)] = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x};\omega_i), \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is the optimization variable and ξ is independently sampled from the dataset $\{\omega_i : \omega_i \in \mathcal{D}, 1 \le i \le n\}$. For simplicity, we sometimes use $\boldsymbol{g} = \nabla f(\boldsymbol{x})$.

3.2 DETAILS OF ADAM

To facilitate the analysis of Adam, we provide the details of Adam in Algorithm 1. As shown in Table 1, the previous studies (Chen et al., 2018; Défossez et al., 2020; Zhang et al., 2022; Wang et al., 2023c) have omitted the bias correction in Line 6-7 of Algorithm 1 when analyzing the convergence rate, while we will provide the theoretical analysis of vanilla Adam with bias correction.

Algorithm 1 Adam with bias correction

1 **Input:** $\boldsymbol{m}_0 = 0$, $\boldsymbol{v}_0 = 0$, numerical-stability constant ϵ, exponential-moving-average coefficients β_1, β_2, and learning rate γ.
2 **for** $t = 1, \ldots, T$ **do**
3 $\quad \boldsymbol{g}_t \leftarrow \nabla f(\boldsymbol{x}_t; \xi_t)$ — stochastic gradient
4 $\quad \boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t$ — first moment
5 $\quad \boldsymbol{v}_t \leftarrow \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^2$ — second moment
6 $\quad \hat{\boldsymbol{m}}_t \leftarrow \boldsymbol{m}_t/(1-\beta_1^t)$ — bias correction
7 $\quad \hat{\boldsymbol{v}}_t \leftarrow \boldsymbol{v}_t/(1-\beta_2^t)$ — bias correction
8 $\quad \boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \gamma\, \hat{\boldsymbol{m}}_t/(\sqrt{\hat{\boldsymbol{v}}_t} + \epsilon)$ — parameter update
9 **end for**

3.3 ASSUMPTIONS AND CONDITIONS

To analyze the convergence rate of Adam, we list the main assumption as follows.

Assumption A [Bounded Infimum]. *There exists a constant $F^* > -\infty$, and the objective function follows $F(\boldsymbol{x}) \ge F^*$ for any $\boldsymbol{x} \in \mathbb{R}^d$.*

Assumption B.1 [L-Smoothness] *There exists a constants $L \ge 0$, and then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, the gradient of the objective function follows*

$$\|\nabla F(\boldsymbol{y}) - \nabla F(\boldsymbol{x})\|_2 \le L\|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{2}$$

Assumption B.2 [(L_0, L_1)-Smoothness] *There exist constants $L_0, L_1 \ge 0$, and then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, the gradient of the objective function follows*

$$\|\nabla F(\boldsymbol{y}) - \nabla F(\boldsymbol{x})\|_2 \le (L_0 + L_1\|\nabla F(\boldsymbol{x})\|_2)\|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{3}$$

Assumption B.3 [(L_0, L_1, q)-Smoothness] *There exist constants $L_0, L_1 > 0$ and $0 \le q \le 1$, and then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, the gradient of the objective function follows*

$$\|\nabla F(\boldsymbol{y}) - \nabla F(\boldsymbol{x})\|_2 \le (L_0 + L_1\|\nabla F(\boldsymbol{x})\|_2^q)\|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{4}$$

When $q = 1$, B.3 becomes B.2. When $q = 0$ (or $L_1 = 0$ is allowed), it reduces to B.1. Generalized smoothness was introduced through Hessian and local gradient-difference conditions (Zhang et al., 2019; 2020) and has been used to analyze Adam (Li et al., 2023; Wang et al., 2024; Hong & Lin, 2025). B.3 as written here is an all-pairs condition; it should not be identified with a local condition without checking its range. We retain B.2–B.3 for the one-step analysis; the corresponding descent lemma is stated and proved in Appendix A. The expectation-rate theorem uses B.1; Corollary 4.4 gives a separate high-probability extension under B.3.

Assumption C.1 [Bounded Variance]. *There exists $\sigma_0 \geq 0$ such that, for every deterministic $\boldsymbol{x} \in \mathbb{R}^d$, the stochastic gradient is unbiased and has bounded variance:*

$$\begin{aligned} &\mathbb{E}[\nabla f(\boldsymbol{x};\xi)] = \nabla F(\boldsymbol{x}), \\ &\mathbb{E}[\|\nabla f(\boldsymbol{x};\xi) - \nabla F(\boldsymbol{x})\|_2^2] \leq \sigma_0^2. \end{aligned} \tag{5}$$

Assumption C.2 [Affine Variance]. *There exist $\sigma_0, \sigma_1 \geq 0$ such that, for every deterministic $\boldsymbol{x} \in \mathbb{R}^d$, the stochastic gradient is unbiased and satisfies*

$$\begin{aligned} &\mathbb{E}[\nabla f(\boldsymbol{x};\xi_t)] = \nabla F(\boldsymbol{x}), \\ &\mathbb{E}[\|\nabla f(\boldsymbol{x};\xi) - \nabla F(\boldsymbol{x})\|_2^2] \leq \sigma_0^2 + \sigma_1^2\|\nabla F(\boldsymbol{x})\|_2^2. \end{aligned} \tag{6}$$

Assumption C.3 [p-Affine Variance]. *There exist $\sigma_0, \sigma_1 \geq 0$ and $0 \leq p \leq 2$ such that, for every deterministic $\boldsymbol{x} \in \mathbb{R}^d$, the stochastic gradient is unbiased and satisfies*

$$\begin{aligned} &\mathbb{E}[\nabla f(\boldsymbol{x};\xi_t)] = \nabla F(\boldsymbol{x}), \\ &\mathbb{E}[\|\nabla f(\boldsymbol{x};\xi) - \nabla F(\boldsymbol{x})\|_2^2] \leq \sigma_0^2 + \sigma_1^2\|\nabla F(\boldsymbol{x})\|_2^p. \end{aligned} \tag{7}$$

When $p = 2$, the p-affine variance is reduced to the affine variance. When $\sigma_1 = 0$ or $p = 0$, the p-affine variance (Assumption C.3) is reduced to the bounded variance (Assumption C.1). The affine variance (Assumption C.2) was originally studied in (Bertsekas & Tsitsiklis, 2000) to analyze the convergence behavior of SGD. It was later applied to analyze AdaGrad (Faw et al., 2022; Wang et al., 2023a) and a simplified Adam (Wang et al., 2024).

4 Main Results: Convergence without Explicit Dimension Dependence

In this section, we establish the direct dimension-free ℓ_2 stationarity guarantee under Assumptions A, B.1, and C.3. Corollary 4.4 gives a separate high-probability extension under B.3, and Section 5 verifies that the full problem-dependent constant remains $O(1)$ in a specified model-scaling regime of normalized DNNs.

Conceptual and technical point. The sign–magnitude identity is the starting point, not the convergence argument. The main step is to bound the accumulated wrong-sign loss for Adam's correlated random magnitudes, absorb the associated drift, and convert weighted descent to the usual ℓ_2 gradient norm. Lemma 4.1 supplies the pathwise identities used in this argument.

We begin with the pathwise sign decomposition; its one-step form also applies under B.3.

Lemma 4.1 (Magnitude-weighted sign descent). *Let $\boldsymbol{u}_t = |\hat{\boldsymbol{m}}_t|/(\sqrt{\hat{\boldsymbol{v}}_t}+\epsilon)$ be generated by Algorithm 1, with $\epsilon > 0$, and write $\boldsymbol{e}_t = \hat{\boldsymbol{m}}_t - \nabla F(\boldsymbol{x}_t)$. For each coordinate, let*

$$\mathcal{B}_{t,j} = \{\mathrm{Sign}(\boldsymbol{m}_t^{(j)}) \neq \mathrm{Sign}(\nabla_j F(\boldsymbol{x}_t))\}.$$

Define

$$\begin{aligned} \mathcal{T}_{1,t} &= \langle |\nabla F(\boldsymbol{x}_t)|, \boldsymbol{u}_t \rangle, \\ \mathcal{T}_{2,t} &= 2\sum_{j=1}^{d} |\nabla_j F(\boldsymbol{x}_t)| u_t^{(j)} \mathbf{1}_{\mathcal{B}_{t,j}}. \end{aligned} \tag{8}$$

Then, on every sample path,

$$\begin{aligned} \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{u}_t \circ \mathrm{Sign}(\boldsymbol{m}_t) \rangle &= \mathcal{T}_{1,t} - \mathcal{T}_{2,t}, \\ \mathcal{T}_{2,t} &\leq 2\|\boldsymbol{u}_t\|_2 \|\boldsymbol{e}_t\|_2, \\ \epsilon\|\boldsymbol{u}_t\|_2^2 &\leq \mathcal{T}_{1,t} + \|\boldsymbol{u}_t\|_2 \|\boldsymbol{e}_t\|_2. \end{aligned} \tag{9}$$

The same realized magnitude therefore appears in both the ideal descent and its tracking-error correction.

The next theorem controls the ideal sign descent and the first-order momentum tracking error. The proof keeps their joint dependence on the random update magnitude.

Theorem 4.2. *Let the sequences $\{\hat{\boldsymbol{m}}_t\}_{t=1}^{T}$, $\{\hat{\boldsymbol{v}}_t\}_{t=1}^{T}$, and $\{\boldsymbol{x}_t\}_{t=1}^{T+1}$ be generated by Algorithm 1 from a deterministic $\boldsymbol{x}_1$. Suppose that Assumptions A, B.1, and C.3 hold, with a positive smoothness upper bound L. For a given $T \geq 2$, choose*

$$\begin{aligned} 1-\beta_1 = T^{-1/2}, \qquad \gamma = \frac{C_2}{\sqrt{T}}, \\ 0 < C_2 \leq \frac{\epsilon}{L\max\{32, 60p\sigma_1^2\}}, \qquad 0 \leq \beta_2 < 1. \end{aligned} \tag{10}$$

Both momentum coefficients are constant during the run. Set

$$\begin{aligned} \Delta &= F(\boldsymbol{x}_1) - F^*, \\ \hat{\sigma}^2 &= \sigma_0^2 + (1-p/2)\sigma_1^2, \\ \overline{\Delta} &= 2\Delta + \frac{60C_2\hat{\sigma}^2}{\epsilon}, \\ \Sigma^2 &= \hat{\sigma}^2 + pL\sigma_1^2\overline{\Delta}, \end{aligned} \tag{11}$$

and define the following constants, all independent of T:

$$\begin{aligned} C_W &= \frac{2\Delta}{C_2} + \frac{60\Sigma^2}{\epsilon}, \\ C_U &= \frac{5C_W}{4\epsilon} + \frac{15\Sigma^2}{\epsilon^2}, \\ C_E &= \sqrt{3}\,\Sigma + C_2 L\sqrt{C_U}. \end{aligned} \tag{12}$$

Then

$$\begin{aligned} &\max_{1\leq t\leq T+1} \mathbb{E}[F(\boldsymbol{x}_t) - F^*] \leq \overline{\Delta}, \\ &\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\langle |\nabla F(\boldsymbol{x}_t)|, \boldsymbol{u}_t\rangle] \leq \frac{C_W}{\sqrt{T}}, \\ &\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\hat{\boldsymbol{m}}_t - \nabla F(\boldsymbol{x}_t)\|_2 \leq \frac{C_E}{T^{1/4}}. \end{aligned} \tag{13}$$

Theorem 4.2 allows arbitrary $\beta_2 \in [0, 1)$. The choice of C_2 uses L, p, σ_1, and the theorem does not require a coupling condition between β_1 and β_2.

Corollary 4.3 (p-affine variance: direct dimension-free ℓ_2 stationarity)**.** *Under the assumptions and parameter choices of Theorem 4.2, let*

$$D_* := \epsilon + \sqrt{2L\overline{\Delta} + \Sigma^2}, \qquad C := 2C_E + \sqrt{2D_* C_W}.$$

Then

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla F(\boldsymbol{x}_t)\|_2 \leq CT^{-1/4}. \tag{14}$$

The constant C contains no explicit factor of the parameter dimension d: the proof is closed directly in vector ℓ_2 quantities rather than through coordinate-wise aggregation. The proof is given in Appendix D.

A separate stopped analysis gives the following high-probability extension under B.3. Its proof is in Appendix E.

Corollary 4.4 (Generalized smoothness: high-probability bound)**.** *Suppose Assumptions A, B.3, and C.3 hold, with B.3 required for all pairs $\boldsymbol{x}, \boldsymbol{y}$ as stated in Section 3, and with C.3 interpreted conditionally on $\mathscr{F}_{t-1}$ at each iterate. Let $\Delta = F(\boldsymbol{x}_1) - F^*$ for deterministic $\boldsymbol{x}_1$. Fix $\epsilon > 0$, $\rho \in (0,1)$, and $T \geq 2$. Define*

$$\begin{aligned} Q &:= 1 + \frac{4\Delta}{\rho}, \qquad H_\rho := \max\left\{\sqrt{4QL_0},\ (4QL_1)^{1/(2-q)}\right\}, \\ L_\rho &:= L_0 + L_1 H_\rho^q, \qquad \Sigma_\rho^2 := \sigma_0^2 + \sigma_1^2 H_\rho^p, \\ \eta_\rho &:= \min\left\{\frac{1}{32}, \frac{\rho Q L_\rho}{120\Sigma_\rho^2}\right\}, \qquad A_\rho := \Delta + \frac{30\eta_\rho \Sigma_\rho^2}{L_\rho}. \end{aligned} \tag{15}$$

If $\Sigma_\rho = 0$, take $\eta_\rho = 1/32$. Let Algorithm 1 use any fixed $\beta_2 \in [0,1)$ and

$$1 - \beta_1 = T^{-1/2}, \qquad \gamma = \frac{\eta_\rho \epsilon}{L_\rho \sqrt{T}}.$$

Set

$$\begin{aligned} C_{W,\rho} &:= \frac{2A_\rho L_\rho}{\eta_\rho \epsilon}, \qquad C_{U,\rho} := \frac{5C_{W,\rho}}{4\epsilon} + \frac{15\Sigma_\rho^2}{\epsilon^2}, \\ C_{E,\rho} &:= \sqrt{3}\,\Sigma_\rho + \eta_\rho \epsilon \sqrt{C_{U,\rho}}, \\ K_{2,\rho} &:= 2C_{E,\rho} + \sqrt{2\left(\epsilon + \sqrt{H_\rho^2 + \Sigma_\rho^2}\right) C_{W,\rho}}. \end{aligned} \tag{16}$$

Then

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\boldsymbol{x}_t)\|_2 \leq \frac{2K_{2,\rho}}{\rho} T^{-1/4}\right) \geq 1 - \rho. \tag{17}$$

All displayed constants are independent of T for each fixed confidence level ρ.

Corollary 4.3 is the main expectation guarantee: under conditional p-affine second moments, the stationarity measure is controlled directly in ℓ_2 and the bound contains no explicit d.

What the sign representation makes possible. For $N \leq T$, write $W_N = \sum_{t \leq N} \mathbb{E}\mathcal{T}_{1,t}$, $B_N = \sum_{t \leq N} \mathbb{E}[\|\boldsymbol{u}_t\|_2 \|\boldsymbol{e}_t\|_2]$, and $U_N = \sum_{t \leq N} \mathbb{E}\|\boldsymbol{u}_t\|_2^2$. With $K_N = N(1-\beta_1) + \sum_{t \leq N} 1/t$ and a second-moment noise bound S^2, Appendix C proves

$$\epsilon U_N \leq W_N + B_N, \qquad B_N \leq S\sqrt{K_N U_N} + \frac{\gamma L}{1-\beta_1} U_N.$$

These coupled inequalities absorb the wrong-sign loss and smoothness remainder into at most half the ideal descent, apart from a noise term. The proof then derives a bound on S from the prefix inequalities under C.3, rather than assuming bounded gradients. The final conversion in Appendix D uses a denominator moment bound, not a positive lower bound on $\boldsymbol{u}_t$.

Parameter scope. The bounds use the horizon-dependent choices in (10). They are dimension-free for fixed $\epsilon > 0$, not uniform as $\epsilon \to 0$: the admissible learning-rate constant is proportional to ϵ. These restrictions are distinct from the absence of a β_1–β_2 coupling condition.

5 Dimension-Free Convergence under Model Scaling

Section 4 establishes the rate

$$\mathcal{O}(CT^{-1/4}), \qquad C = C(L_0, L_1, \sigma_0, \sigma_1, \Delta). \tag{18}$$

Here C denotes the appropriate constant: under B.1, $L_0 = L$ and L_1 is unused; under B.3, $C = 2K_{2,\rho}/\rho$. The exponents, algorithmic settings (including ϵ and a fixed positive fraction of the admissible step-size bound), mini-batch size, and confidence level are held fixed. We establish dimension-free convergence under model scaling by proving that these problem quantities remain bounded for a family of fan-in-scaled networks with increasing hidden widths and fixed depth and output dimension. All bounds below concern parameter dimension within this family, not arbitrary depth scaling.

Theorem 5.1 (Dimension-free problem constants under model scaling)**.** *Consider the network family in (71), with increasing hidden widths and fixed depth and output dimension, on convex regions Θ satisfying Assumption F.1. Let $\boldsymbol{g}_t$ average B_{mb} fresh independent sample gradients. For softmax cross-entropy, there are $G_*, L_* \geq 1$, independent of d within this family, such that on each Θ,*

$$\|\nabla F(\boldsymbol{\theta})\|_2 \leq G_*, \tag{19}$$

$$\mathbb{E}[\|\boldsymbol{g}_t - \nabla F(\boldsymbol{\theta}_t)\|_2^2 \mid \mathscr{F}_{t-1}] \leq G_*^2/B_{\mathrm{mb}}, \tag{20}$$

$$\|\nabla F(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta}')\|_2 \leq L_*\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2. \tag{21}$$

For any $p \in [0,2]$ and $q \in [0,1]$, admissible coefficients for the B.3 and C.3 inequalities on Θ are

$$L_0 = L_*, \qquad L_1 = 1, \qquad \sigma_0 = G_*/\sqrt{B_{\mathrm{mb}}}, \qquad \sigma_1 = 1. \tag{22}$$

If the initial logits $z_1 = a_K(\boldsymbol{\theta}_1; \xi)$ satisfy $B_{\mathrm{init}} := \mathbb{E}_\xi\|z_1\|_\infty = O(1)$ as hidden width grows, then

$$0 \leq \Delta \leq \log n_K + 2B_{\mathrm{init}}, \qquad L_0, L_1, \sigma_0, \sigma_1, \Delta = O(1). \tag{23}$$

These are admissible, nonunique coefficients. The network result proves the stronger regional bounds of L-smoothness and bounded variance; the positive L_1, σ_1 do not assert that gradient-dependent growth is necessary.

Theorem 5.2 (Width-uniform Adam convergence for normalized networks)**.** *Consider (71) at fixed depth and input/output dimensions, with fixed-affine regularized LayerNorm/RMSNorm blocks and smooth activations with bounded first two derivatives. Train all W_ℓ by Algorithm 1. Assume width-uniform initial logits and scaled spectral norms for $W_2, \ldots, W_K$ with margin $r > 0$, and set $n_{\min} := \min_{\ell<K} n_\ell$. Use (10) with sufficiently small fixed C_2. There are width- and T-independent $c_{\mathrm{det}}, c_{\mathrm{hp}}, C_{\mathrm{sq}} > 0$ such that:*

1. *if $n_{\min} \geq c_{\mathrm{det}} T$, the Adam iterates remain in the width-uniform region almost surely for T steps and*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla F(\boldsymbol{\theta}_t)\|_2^2 \leq C_{\mathrm{sq}} T^{-1/2}; \tag{24}$$

2. *for any $\rho \in (0,1)$, if $n_{\min} \geq c_{\mathrm{hp}}\sqrt{T}/\rho$, then with probability at least $1-\rho$ the iterates remain in that region and*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\boldsymbol{\theta}_t)\|_2 \leq \sqrt{2C_{\mathrm{sq}}/\rho}\, T^{-1/4}. \tag{25}$$

All constants are independent of total parameter count. The $\Omega(\sqrt{T}/\rho)$ case uses the $O(\sqrt{T})$ accumulated update energy from the sign analysis; the first-layer spectral norm is not restricted. See Appendix F.6.

Interpretation. The theorem gives an explicit width–horizon regime in which Adam achieves a convergence constant independent of hidden width and total parameter count. To our knowledge, this is the first such convergence-rate guarantee for an explicit growing family of normalized networks. The result is qualitatively consistent with RMSNorm models trained with AdamW across scales and with reported scaling-law sample-efficiency gains (Touvron et al., 2023; Bai et al., 2023; Kaplan et al., 2020). Its scope is the network family specified above; it does not claim a theory of full Llama/Qwen training or that widening universally accelerates optimization.

6 Conclusion

We turn Adam's magnitude-weighted sign representation into a convergence analysis that retains the coupling among update magnitudes, momentum, and gradient noise. By controlling the accumulated wrong-sign loss, we obtain a direct expected ℓ_2 rate of $\mathcal{O}(CT^{-1/4})$ under L-smoothness and conditional p-affine second moments, without an explicit d or extra $\log T$. We then establish width-uniform problem constants for fan-in-scaled LayerNorm/RMSNorm networks. For a concrete fixed-affine family, explicit width–horizon conditions ensure that the Adam iterates remain in the width-uniform region and achieve a convergence constant independent of parameter count. To our knowledge, this is the first Adam convergence-rate theorem establishing dimension-free optimization for an explicit growing normalized-network family.

AI USE STATEMENT

In preparing this work, we used generative AI tools to polish the writing, identify and retrieve relevant literature, summarize and analyze prior work, and assist in refining proofs. We reviewed all AI-assisted material and conducted a manual literature survey to check AI-suggested research ideas for potential overlap with prior work. The authors take full responsibility for the final manuscript, including any text, claims, or artifacts developed with AI assistance.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. Gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Amit Attia and Tomer Koren. SGD with AdaGrad stepsizes: Full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *International Conference on Machine Learning*, pp. 1147–1171. PMLR, 2023.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. URL https://arxiv.org/abs/1607.06450.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. URL https://arxiv.org/abs/2309.16609.

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pp. 404–413, 2018.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pp. 560–569, 2018.

Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Congliang Chen, Li Shen, Fangyu Zou, and Wei Liu. Towards practical Adam: Non-convexity, convergence theory, and mini-batch acceleration. *Journal of Machine Learning Research*, 23(229): 1–47, 2022.

Lizhang Chen, Bo Liu, Kaizhao Liang, and Qiang Liu. Lion secretly solves constrained optimization: As lyapunov predicts. *arXiv preprint arXiv:2310.05898*, 2023a.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023b.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of Adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018.

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

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of Adam and AdaGrad. *arXiv preprint arXiv:2003.02395*, 2020.

Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in SGD: Self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, pp. 313–355. PMLR, 2022.

Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive SGD. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 89–160. PMLR, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.

Yusu Hong and Junhong Lin. High probability convergence of Adam under unbounded gradients and affine variance noise. *arXiv preprint arXiv:2311.02000*, 2023.

Yusu Hong and Junhong Lin. On convergence of Adam for stochastic optimization under relaxed assumptions. *Advances in Neural Information Processing Systems*, 2025.

Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *International Conference on Artificial Intelligence and Statistics*, pp. 4861–4869. PMLR, 2024.

Jikai Jin, Bohang Zhang, Haiyang Wang, and Liwei Wang. Non-convex distributionally robust optimization: Non-asymptotic analysis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2771–2782, 2021.

Ruinan Jin, Yingbin Liang, and Shaofeng Zou. Why Adam can beat SGD: Second-moment normalization yields sharper tails, 2026. URL https://arxiv.org/abs/2603.03099v8. arXiv:2603.03099v8.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. *https://kellerjordan.github.io/posts/muon/*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML-PKDD-2016*, pp. 795–811, 2016.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

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. *arXiv preprint arXiv:2304.13960*, 2023.

Frederik Kunstner, Robin Yadav, Alan Milligan, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why Adam outperforms gradient descent on language models. *arXiv preprint arXiv:2402.19449*, 2024.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of Adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36, 2023.

Huan Li, Yiming Dong, and Zhouchen Lin. On the $O(\frac{\sqrt{d}}{K^{1/4}})$ convergence rate of AdamW measured by ℓ_1 norm. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

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

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763, 2021.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

Martin Riedmiller and Heinrich Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *IEEE international conference on neural networks*, pp. 586–591. IEEE, 1993.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Conference of the International Speech Communication Association*, volume 2014, pp. 1058–1062. Singapore, 2014.

Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, et al. Practical efficiency of muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025.

Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. RMSProp converges with proper hyperparameter. In *International Conference on Learning Representation*, 2021.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Nikko Ström. Scalable distributed DNN training using commodity GPU cloud computing. In *Conference of the International Speech Communication Association*, 2015.

Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of signSGD under weaker assumptions. In *International Conference on Machine Learning*, pp. 33077–33099, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

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.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of AdaGrad for non-convex objectives: Simple proofs and relaxed assumptions. In *Conference on Learning Theory*, pp. 161–190. PMLR, 2023a.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of AdaGrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 161–190. PMLR, 2023b.

Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, Tie-Yan Liu, Zhi-Quan Luo, and Wei Chen. Provable adaptivity of Adam under non-uniform smoothness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2960–2969, 2023c.

Bohan Wang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, and Wei Chen. On the convergence of Adam under non-uniform smoothness: Separability from sgdm and beyond. *arXiv preprint arXiv:2403.15146*, 2024.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *arXiv preprint arXiv:1910.07467*, 2019. URL https://arxiv.org/abs/1910.07467.

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

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*, 2019.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. *Advances in neural information processing systems*, 35: 28386–28399, 2022.

Appendix

A Descent lemma under generalized smoothness

Lemma A.1 (Descent lemma under B.3)**.** *Under Assumption B.3, for any* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$F(\boldsymbol{y}) \le F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L_0 + L_1 \|\nabla F(\boldsymbol{x})\|_2^q}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \tag{26}$$

Proof. By the fundamental theorem of calculus,

$$\begin{aligned} F(\boldsymbol{y}) - F(\boldsymbol{x}) &= \int_0^1 \langle \nabla F(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \rangle \, dt \\ &= \langle \nabla F(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \\ &\quad + \int_0^1 \langle \nabla F(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - \nabla F(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \, dt. \end{aligned}$$

Assumption B.3 gives

$$\|\nabla F(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - \nabla F(\boldsymbol{x})\|_2 \le t\big(L_0 + L_1 \|\nabla F(\boldsymbol{x})\|_2^q\big) \|\boldsymbol{y} - \boldsymbol{x}\|_2.$$

Applying Cauchy–Schwarz inside the integral and using $\int_0^1 t \, dt = 1/2$ proves (26). □

B Proof of Lemma 4.1

Proof. Fix a coordinate j and a sample path. Bias correction preserves the sign of $\boldsymbol{m}_t^{(j)}$. If this coordinate of $\boldsymbol{m}_t$ is zero, then $u_t^{(j)} = 0$. With $\mathrm{Sign}(0) = 0$, it follows in all cases that

$$\nabla_j F(\boldsymbol{x}_t) u_t^{(j)} \mathrm{Sign}(\boldsymbol{m}_t^{(j)}) = |\nabla_j F(\boldsymbol{x}_t)| u_t^{(j)} - 2 |\nabla_j F(\boldsymbol{x}_t)| u_t^{(j)} \mathbf{1}_{\mathcal{B}_{t,j}}. \tag{27}$$

On $\mathcal{B}_{t,j}$, $|\nabla_j F(\boldsymbol{x}_t)| \le |e_t^{(j)}|$. Summing (27) and applying Cauchy–Schwarz proves

$$\mathcal{T}_{2,t} \le 2 \langle |\boldsymbol{e}_t|, \boldsymbol{u}_t \rangle \le 2 \|\boldsymbol{u}_t\|_2 \|\boldsymbol{e}_t\|_2. \tag{28}$$

Finally, $D_t^{(j)} := \sqrt{\hat{\boldsymbol{v}}_t^{(j)}} + \epsilon \ge \epsilon$ and $D_t^{(j)} u_t^{(j)} = |\hat{\boldsymbol{m}}_t^{(j)}|$. Hence

$$\epsilon \|\boldsymbol{u}_t\|_2^2 \le \sum_j D_t^{(j)} (u_t^{(j)})^2 = \langle |\hat{\boldsymbol{m}}_t|, \boldsymbol{u}_t \rangle \le \mathcal{T}_{1,t} + \|\boldsymbol{u}_t\|_2 \|\boldsymbol{e}_t\|_2. \tag{29}$$

All statements are pathwise. □

C Proof of Theorem 4.2

Proof. All expectations in this proof are over the original sampling process. We use B.1 with $L > 0$; no stopping event is introduced. Write $\delta = 1 - \beta_1$ and $\boldsymbol{e}_t = \hat{\boldsymbol{m}}_t - \nabla F(\boldsymbol{x}_t)$.

Finiteness of the moments. For each finite horizon, the moments used below are finite. Indeed, Lipschitz continuity of the gradient gives $\|\nabla F(\boldsymbol{x})\|_2 \le \|\nabla F(\boldsymbol{x}_1)\|_2 + L\|\boldsymbol{x} - \boldsymbol{x}_1\|_2$. Since $p \le 2$, C.3 then implies a bound of the form $\mathbb{E}[\|\boldsymbol{g}_t\|_2^2 \mid \mathscr{F}_{t-1}] \le c(1 + \|\boldsymbol{x}_t - \boldsymbol{x}_1\|_2^2)$, for a finite c. The vector $\hat{\boldsymbol{m}}_t$ is a convex combination of past gradients, and $\|\boldsymbol{u}_t\|_2 \le \|\hat{\boldsymbol{m}}_t\|_2/\epsilon$. Induction in the update rule proves finiteness of the second moments of $\boldsymbol{x}_t$, $\boldsymbol{g}_t$, $\hat{\boldsymbol{m}}_t$, and $\boldsymbol{u}_t$. The quadratic upper bound on F from B.1 and its lower bound from A also give $\mathbb{E}|F(\boldsymbol{x}_t)| < \infty$.

Ideal sign descent and the two error terms. Following the same single-step descent argument as in Lemma A.1, now with the constant L of B.1, we have

$$F(\boldsymbol{x}_{t+1}) \le F(\boldsymbol{x}_t) - \gamma\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{u}_t \circ \mathrm{Sign}(\boldsymbol{m}_t)\rangle + \frac{L\gamma^2}{2}\|\boldsymbol{u}_t\|_2^2. \tag{30}$$

By Lemma 4.1, the inner product is $\mathcal{T}_{1,t} - \mathcal{T}_{2,t}$. The error term satisfies $\mathcal{T}_{2,t} \le 2\|\boldsymbol{u}_t\|_2\|\boldsymbol{e}_t\|_2$. We keep this magnitude weight instead of replacing it by a uniform upper bound.

For any $N \le T$, define

$$\begin{aligned} W_N &:= \sum_{t=1}^{N} \mathbb{E}\mathcal{T}_{1,t}, \\ B_N &:= \sum_{t=1}^{N} \mathbb{E}[\|\boldsymbol{u}_t\|_2\|\boldsymbol{e}_t\|_2], \\ U_N &:= \sum_{t=1}^{N} \mathbb{E}\|\boldsymbol{u}_t\|_2^2, \end{aligned} \tag{31}$$

and let $K_N = N\delta + \sum_{t=1}^{N} 1/t$. Taking expectations and summing (30) gives

$$\mathbb{E}[F(\boldsymbol{x}_{N+1}) - F^*] + \gamma W_N \le \Delta + 2\gamma B_N + \frac{L\gamma^2}{2}U_N, \qquad \epsilon U_N \le W_N + B_N. \tag{32}$$

The second inequality is the sum of the last inequality in (9).

Momentum expansion. Put $\boldsymbol{\zeta}_t = \boldsymbol{g}_t - \nabla F(\boldsymbol{x}_t)$. Recalling $\boldsymbol{m}_t = \beta_1\boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t$, for $t \ge 2$ we obtain

$$\begin{aligned} \boldsymbol{m}_t - \nabla F(\boldsymbol{x}_t) &= \beta_1(\boldsymbol{m}_{t-1} - \nabla F(\boldsymbol{x}_{t-1})) + (1-\beta_1)\boldsymbol{\zeta}_t - \beta_1(\nabla F(\boldsymbol{x}_t) - \nabla F(\boldsymbol{x}_{t-1})), \\ \boldsymbol{m}_t - \nabla F(\boldsymbol{x}_t) &= -\beta_1^t \nabla F(\boldsymbol{x}_1) + (1-\beta_1)\sum_{k=1}^{t}\beta_1^{t-k}\boldsymbol{\zeta}_k - \sum_{k=2}^{t}\beta_1^{t-k+1}(\nabla F(\boldsymbol{x}_k) - \nabla F(\boldsymbol{x}_{k-1})). \end{aligned} \tag{33}$$

The initial identity is $\boldsymbol{m}_1 - \nabla F(\boldsymbol{x}_1) = -\beta_1\nabla F(\boldsymbol{x}_1) + \delta\boldsymbol{\zeta}_1$. To handle bias correction without discarding this initial term, define

$$p_{t,k} := \frac{\delta\beta_1^{t-k}}{1-\beta_1^t}, \qquad c_{t,s} := \frac{\beta_1^{t-s}(1-\beta_1^s)}{1-\beta_1^t} \le \beta_1^{t-s}. \tag{34}$$

Since $\sum_{k=1}^{t} p_{t,k} = 1$, the bias-corrected error has the exact decomposition

$$\boldsymbol{e}_t = \underbrace{\sum_{k=1}^{t} p_{t,k}\boldsymbol{\zeta}_k}_{\boldsymbol{n}_t} - \underbrace{\sum_{s=1}^{t-1} c_{t,s}(\nabla F(\boldsymbol{x}_{s+1}) - \nabla F(\boldsymbol{x}_s))}_{\boldsymbol{d}_t}. \tag{35}$$

This follows either by combining (33) with the bias correction, or by writing $\nabla F(\boldsymbol{x}_k) - \nabla F(\boldsymbol{x}_t) = -\sum_{s=k}^{t-1}(\nabla F(\boldsymbol{x}_{s+1}) - \nabla F(\boldsymbol{x}_s))$. In particular, there is no undefined iterate $\boldsymbol{x}_0$ in the drift sum.

The noise term. For now, suppose S^2 bounds the unconditional noise second moments up to time N. We will derive such a bound, uniform in T, below. Conditional unbiasedness makes the martingale differences orthogonal, even though the iterates are adaptive. Thus

$$\begin{aligned} \mathbb{E}\|\boldsymbol{n}_t\|_2^2 &= \sum_{k=1}^{t} p_{t,k}^2 \mathbb{E}\|\boldsymbol{\zeta}_k\|_2^2, \\ \sum_{k=1}^{t} p_{t,k}^2 &= \frac{\delta}{1+\beta_1}\frac{1+\beta_1^t}{1-\beta_1^t} \le \frac{\delta}{1-\beta_1^t} \le \delta + \frac{1}{t}. \end{aligned} \tag{36}$$

For the last inequality, $(1-\delta)^{-t} \geq 1+t\delta$ implies $1-(1-\delta)^t \geq t\delta/(1+t\delta)$. Summing (36) and using Cauchy–Schwarz jointly over time and randomness gives

$$\sum_{t=1}^{N} \mathbb{E}[\|\boldsymbol{u}_t\|_2 \|\boldsymbol{n}_t\|_2] \leq \sqrt{U_N \sum_{t=1}^{N} \mathbb{E}\|\boldsymbol{n}_t\|_2^2} \leq S\sqrt{K_N U_N}. \tag{37}$$

Only the noise second moment is used here; we do not form a recursion for $\mathbb{E}\|\boldsymbol{e}_t\|_2^2$.

The gradient drift. By B.1 and the update rule, $\|\nabla F(\boldsymbol{x}_{s+1}) - \nabla F(\boldsymbol{x}_s)\|_2 \leq L\gamma\|\boldsymbol{u}_s\|_2$. For $a_t = \|\boldsymbol{u}_t\|_2$, the geometric sum obeys the pathwise bound

$$\sum_{s<t\leq N} \beta_1^{t-s} a_s a_t \leq \frac{1}{2} \sum_{s<t\leq N} \beta_1^{t-s}(a_s^2 + a_t^2) \leq \frac{1}{\delta} \sum_{t=1}^{N} a_t^2. \tag{38}$$

Consequently,

$$\sum_{t=1}^{N} \mathbb{E}[\|\boldsymbol{u}_t\|_2 \|\boldsymbol{d}_t\|_2] \leq \frac{\gamma L}{\delta} U_N, \qquad B_N \leq S\sqrt{K_N U_N} + \frac{\gamma L}{\delta} U_N. \tag{39}$$

The same expansion, without the magnitude weight, also yields

$$\frac{1}{N} \sum_{t=1}^{N} \mathbb{E}\|\boldsymbol{e}_t\|_2 \leq S\sqrt{\frac{K_N}{N}} + \frac{\gamma L}{\delta}\sqrt{\frac{U_N}{N}}. \tag{40}$$

Indeed, the noise contribution is at most $\sqrt{N\sum_t \mathbb{E}\|\boldsymbol{n}_t\|_2^2}$ before division by N, and the drift is at most $(\gamma L/\delta)\sum_t \mathbb{E}\|\boldsymbol{u}_t\|_2$.

Absorbing the sign loss and the smoothness remainder. Let $\mathcal{R}_N = 2B_N + (\gamma L/2)U_N$. Since $\gamma L/\delta \leq \epsilon/32$ and $\delta \leq 1$, (39) implies

$$\begin{aligned} \mathcal{R}_N &\leq 2S\sqrt{K_N U_N} + \gamma L\left(\frac{2}{\delta} + \frac{1}{2}\right) U_N \\ &\leq 2S\sqrt{K_N U_N} + \frac{\epsilon}{8} U_N \leq \frac{\epsilon}{4} U_N + \frac{8S^2 K_N}{\epsilon}. \end{aligned} \tag{41}$$

On the other hand, (32) gives $\epsilon U_N \leq W_N + B_N \leq W_N + \mathcal{R}_N/2$. Combining the two inequalities, we obtain

$$\mathcal{R}_N \leq \frac{1}{4} W_N + \frac{1}{8}\mathcal{R}_N + \frac{8S^2 K_N}{\epsilon}, \qquad \mathcal{R}_N \leq \frac{2}{7} W_N + \frac{64 S^2 K_N}{7\epsilon} \leq \frac{1}{2} W_N + \frac{10 S^2 K_N}{\epsilon}. \tag{42}$$

Thus the two error terms consume at most half the ideal sign descent, apart from the noise contribution. Substitution into (32) proves, for every prefix N,

$$\mathbb{E}[F(\boldsymbol{x}_{N+1}) - F^*] + \frac{\gamma}{2} W_N \leq \Delta + \frac{10\gamma S^2 K_N}{\epsilon}, \qquad \epsilon U_N \leq \frac{5}{4} W_N + \frac{5 S^2 K_N}{\epsilon}. \tag{43}$$

Closing the expectation bound under C.3. Let $\Delta_T = \max_{1\leq t\leq T+1} \mathbb{E}[F(\boldsymbol{x}_t) - F^*]$, which is finite by the initial moment argument. For any $\boldsymbol{x}$, taking $\boldsymbol{y} = \boldsymbol{x} - \nabla F(\boldsymbol{x})/L$ in the smoothness inequality and using $F(\boldsymbol{y}) \geq F^*$ gives

$$\|\nabla F(\boldsymbol{x})\|_2^2 \leq 2L(F(\boldsymbol{x}) - F^*). \tag{44}$$

Young's inequality $a^p \leq 1 - p/2 + (p/2)a^2$ now gives

$$\mathbb{E}\|\boldsymbol{\zeta}_t\|_2^2 \leq \hat{\sigma}^2 + \frac{p}{2}\sigma_1^2 \mathbb{E}\|\nabla F(\boldsymbol{x}_t)\|_2^2 \leq \hat{\sigma}^2 + pL\sigma_1^2 \Delta_T =: S_T^2. \tag{45}$$

This is a finite bound for the given horizon, not yet a uniform bound. Apply (43) with $S = S_T$ to each $N \leq T$. Since $\delta = T^{-1/2}$,

$$K_N \leq K_T = \sqrt{T} + \sum_{t=1}^{T} \frac{1}{t} \leq 3\sqrt{T}.$$

Taking the maximum over the prefixes therefore yields

$$\Delta_T \le \Delta + \frac{30C_2}{\epsilon}(\hat{\sigma}^2 + pL\sigma_1^2\Delta_T). \tag{46}$$

The coefficient of Δ_T on the right is at most $1/2$ by (10). Hence

$$\Delta_T \le 2\Delta + \frac{60C_2\hat{\sigma}^2}{\epsilon} = \overline{\Delta}, \qquad \mathbb{E}\|\boldsymbol{\zeta}_t\|_2^2 \le \Sigma^2. \tag{47}$$

No bound on the function values or gradients was assumed; the expectation bounds in (47) follow from the prefix inequalities.

Finally, use $S = \Sigma$ and $N = T$ in (43), then in (40). This gives

$$\begin{aligned} \frac{W_T}{T} &\le \left(\frac{2\Delta}{C_2} + \frac{60\Sigma^2}{\epsilon}\right) T^{-1/2} = C_W T^{-1/2}, \\ \frac{U_T}{T} &\le \left(\frac{5C_W}{4\epsilon} + \frac{15\Sigma^2}{\epsilon^2}\right) T^{-1/2} = C_U T^{-1/2}, \\ \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\boldsymbol{e}_t\|_2 &\le \left(\sqrt{3}\,\Sigma + C_2 L\sqrt{C_U}\right) T^{-1/4} = C_E T^{-1/4}. \end{aligned} \tag{48}$$

Together with (47), these are the claimed bounds. □

D Proof of Corollary 4.3

Proof. We convert the ideal sign descent to the usual gradient norm only at this last step. Write $\boldsymbol{e}_t = \hat{\boldsymbol{m}}_t - \nabla F(\boldsymbol{x}_t)$ and $D_t^{(j)} = \sqrt{\hat{\boldsymbol{v}}_t^{(j)}} + \epsilon$.

Reliable signs and coordinates dominated by tracking error. For a fixed coordinate, suppose first that $|\nabla_j F(\boldsymbol{x}_t)| > 2|e_t^{(j)}|$. Then $\hat{\boldsymbol{m}}_t^{(j)}$ has the correct sign and $|\hat{\boldsymbol{m}}_t^{(j)}| \ge |\nabla_j F(\boldsymbol{x}_t)|/2$. Thus

$$|\nabla_j F(\boldsymbol{x}_t)| \le \sqrt{2D_t^{(j)}|\nabla_j F(\boldsymbol{x}_t)|u_t^{(j)}}.$$

In the complementary case, $|\nabla_j F(\boldsymbol{x}_t)| \le 2|e_t^{(j)}|$. Combining the cases, including zero coordinates, gives

$$|\nabla_j F(\boldsymbol{x}_t)| \le 2|e_t^{(j)}| + \sqrt{2D_t^{(j)}|\nabla_j F(\boldsymbol{x}_t)|u_t^{(j)}}. \tag{49}$$

Let $D_{t,\max} = \max_j D_t^{(j)}$. Taking the Euclidean norm of (49) and applying the triangle inequality yields

$$\|\nabla F(\boldsymbol{x}_t)\|_2 \le 2\|\boldsymbol{e}_t\|_2 + \sqrt{2\sum_j D_t^{(j)}|\nabla_j F(\boldsymbol{x}_t)|u_t^{(j)}} \le 2\|\boldsymbol{e}_t\|_2 + \sqrt{2D_{t,\max}\mathcal{T}_{1,t}}. \tag{50}$$

The denominator moment. Conditional unbiasedness and (47) imply

$$\mathbb{E}\|\boldsymbol{g}_t\|_2^2 = \mathbb{E}\|\nabla F(\boldsymbol{x}_t)\|_2^2 + \mathbb{E}\|\boldsymbol{\zeta}_t\|_2^2 \le 2L\overline{\Delta} + \Sigma^2. \tag{51}$$

Since the coefficients in $\hat{\boldsymbol{v}}_t$ are nonnegative and sum to one,

$$\mathbb{E}\|\hat{\boldsymbol{v}}_t\|_1 = \frac{1-\beta_2}{1-\beta_2^t}\sum_{k=1}^{t}\beta_2^{t-k}\mathbb{E}\|\boldsymbol{g}_k\|_2^2 \le 2L\overline{\Delta} + \Sigma^2. \tag{52}$$

For $\beta_2 = 0$, this formula is understood as $\hat{\boldsymbol{v}}_t = \boldsymbol{g}_t^2$. As $D_{t,\max} \le \epsilon + \sqrt{\|\hat{\boldsymbol{v}}_t\|_1}$, Jensen's inequality gives

$$\mathbb{E}D_{t,\max} \le \epsilon + \sqrt{2L\overline{\Delta} + \Sigma^2} = D_*. \tag{53}$$

Only this expectation bound is needed, not a pathwise bound on $\boldsymbol{g}_t$ or $\hat{\boldsymbol{v}}_t$.

Expected average gradient norm. Apply Cauchy–Schwarz jointly over the time index and the sampling randomness in (50). By Theorem 4.2 and (53),

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\boldsymbol{x}_t)\|_2 &\le \frac{2C_E}{T^{1/4}} + \sqrt{2\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}D_{t,\max}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\mathcal{T}_{1,t}\right)} \\ &\le \left(2C_E + \sqrt{2D_* C_W}\right) T^{-1/4}.\end{aligned} \tag{54}$$

This is a full-expectation bound obtained directly in the vector ℓ_2 norm. □

E Proof of Corollary 4.4

Lemma E.1 (Weighted first-order momentum error)**.** *Write $w_t := \mathcal{T}_{1,t} = \langle|\nabla F(\boldsymbol{x}_t)|, \boldsymbol{u}_t\rangle$ and $b_t := \|\boldsymbol{u}_t\|_2\|\boldsymbol{e}_t\|_2$. Let $\delta = 1-\beta_1 \in (0,1]$ and $\boldsymbol{e}_t = \hat{\boldsymbol{m}}_t - \nabla F(\boldsymbol{x}_t)$. Let $I_t \in \{0,1\}$ be $\mathscr{F}_{t-1}$-measurable and nonincreasing in t; the choice $I_t \equiv 1$ is allowed. Assume the following quantities are finite. Suppose, for $k \le n$,*

$$\mathbb{E}[\boldsymbol{\zeta}_k \mid \mathscr{F}_{k-1}] = 0, \qquad \mathbb{E}[I_k\|\boldsymbol{\zeta}_k\|_2^2] \le \Sigma^2, \qquad \boldsymbol{\zeta}_k = \boldsymbol{g}_k - \nabla F(\boldsymbol{x}_k),$$

and that, whenever $I_t = 1$ and $s < t$,

$$\|\nabla F(\boldsymbol{x}_{s+1}) - \nabla F(\boldsymbol{x}_s)\|_2 \le \gamma\mathcal{L}\|\boldsymbol{u}_s\|_2$$

for a deterministic constant $\mathcal{L} > 0$. Set

$$\mathcal{U}_n := \sum_{t=1}^{n}\mathbb{E}[I_t\|\boldsymbol{u}_t\|_2^2], \quad \mathcal{B}_n := \sum_{t=1}^{n}\mathbb{E}[I_t\|\boldsymbol{u}_t\|_2\|\boldsymbol{e}_t\|_2], \quad N_n := n\delta + \sum_{t=1}^{n}\frac{1}{t}.$$

Then

$$\mathcal{B}_n \le \Sigma\sqrt{N_n\mathcal{U}_n} + \frac{\gamma\mathcal{L}}{\delta}\mathcal{U}_n, \tag{55}$$

and

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[I_t\|\boldsymbol{e}_t\|_2] \le \Sigma\sqrt{\frac{N_n}{n}} + \frac{\gamma\mathcal{L}}{\delta}\sqrt{\frac{\mathcal{U}_n}{n}}. \tag{56}$$

Proof. We follow the separation into noise and gradient drift. Recalling $\boldsymbol{m}_t = \beta_1\boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t$, we first have, for $t \ge 2$,

$$\begin{aligned}\boldsymbol{m}_t - \nabla F(\boldsymbol{x}_t) = \beta_1(\boldsymbol{m}_{t-1} - \nabla F(\boldsymbol{x}_{t-1})) + (1-\beta_1)(\boldsymbol{g}_t - \nabla F(\boldsymbol{x}_t)) \\ - \beta_1(\nabla F(\boldsymbol{x}_t) - \nabla F(\boldsymbol{x}_{t-1})).\end{aligned} \tag{57}$$

Since $\boldsymbol{m}_0 = 0$, expanding this recursion gives

$$\begin{aligned}\boldsymbol{m}_t - \nabla F(\boldsymbol{x}_t) = &-\beta_1^t\nabla F(\boldsymbol{x}_1) + (1-\beta_1)\sum_{k=1}^{t}\beta_1^{t-k}(\boldsymbol{g}_k - \nabla F(\boldsymbol{x}_k)) \\ &- \sum_{k=2}^{t}\beta_1^{t-k+1}(\nabla F(\boldsymbol{x}_k) - \nabla F(\boldsymbol{x}_{k-1})).\end{aligned} \tag{58}$$

Dividing by $1-\beta_1^t$ and collecting the gradient terms yields the exact bias-corrected identity

$$\boldsymbol{e}_t = \boldsymbol{n}_t - \sum_{s=1}^{t-1}c_{t,s}(\nabla F(\boldsymbol{x}_{s+1}) - \nabla F(\boldsymbol{x}_s)), \quad \boldsymbol{n}_t = \sum_{k=1}^{t}p_{t,k}\boldsymbol{\zeta}_k, \tag{59}$$

where

$$p_{t,k} = \frac{\delta\beta_1^{t-k}}{1-\beta_1^t}, \qquad c_{t,s} = \frac{\beta_1^{t-s}(1-\beta_1^s)}{1-\beta_1^t} \le \beta_1^{t-s}.$$

In particular, $\boldsymbol{e}_1 = \boldsymbol{\zeta}_1$; no initial bias term has been discarded.

Noise term. Because $I_t = 1$ implies $I_k = 1$ for $k \le t$,

$$\mathbb{E}[I_t\|\boldsymbol{n}_t\|_2^2] \le \mathbb{E}\left\|\sum_{k=1}^{t} p_{t,k} I_k \boldsymbol{\zeta}_k\right\|_2^2 = \sum_{k=1}^{t} p_{t,k}^2 \mathbb{E}[I_k\|\boldsymbol{\zeta}_k\|_2^2].$$

The equality follows from conditional mean zero and predictability of I_k. The geometric sum satisfies

$$\sum_{k=1}^{t} p_{t,k}^2 = \frac{\delta}{1+\beta_1}\frac{1+\beta_1^t}{1-\beta_1^t} \le \frac{\delta}{1-\beta_1^t} \le \delta + \frac{1}{t}. \tag{60}$$

For $0 < \delta < 1$, the last inequality follows from $(1-\delta)^{-t} \ge 1 + t\delta$; the case $\delta = 1$ is immediate. Thus

$$\sum_{t=1}^{n} \mathbb{E}[I_t\|\boldsymbol{n}_t\|_2^2] \le \Sigma^2 N_n.$$

Cauchy–Schwarz over time and samples gives

$$\sum_{t=1}^{n} \mathbb{E}[I_t\|\boldsymbol{u}_t\|_2\|\boldsymbol{n}_t\|_2] \le \Sigma\sqrt{N_n \mathcal{U}_n}, \qquad \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[I_t\|\boldsymbol{n}_t\|_2] \le \Sigma\sqrt{\frac{N_n}{n}}. \tag{61}$$

Gradient drift. On an active prefix, the drift term in (59) is bounded by $\gamma\mathcal{L}\sum_{s<t}\beta_1^{t-s}\|\boldsymbol{u}_s\|_2$. Put $a_t = I_t\|\boldsymbol{u}_t\|_2$. Since $I_t \le I_s$ for $s < t$, the weighted drift sum is at most

$$\begin{aligned}\gamma\mathcal{L}\,\mathbb{E}\sum_{s<t\le n}\beta_1^{t-s} a_s a_t &\le \frac{\gamma\mathcal{L}}{2}\mathbb{E}\sum_{s<t\le n}\beta_1^{t-s}(a_s^2 + a_t^2) \\ &\le \frac{\gamma\mathcal{L}}{\delta}\mathcal{U}_n.\end{aligned} \tag{62}$$

For the unweighted sum, interchange the two finite sums and use $I_t \le I_s$ to obtain

$$\frac{\gamma\mathcal{L}}{n}\sum_{t=1}^{n}\sum_{s<t}\beta_1^{t-s}\mathbb{E}[I_t\|\boldsymbol{u}_s\|_2] \le \frac{\gamma\mathcal{L}}{\delta n}\sum_{s=1}^{n}\mathbb{E}[a_s] \le \frac{\gamma\mathcal{L}}{\delta}\sqrt{\frac{\mathcal{U}_n}{n}}. \tag{63}$$

Combining (59), (61), and (62) proves (55); using (63) proves (56). □

Lemma E.2 (Absorbing sign errors and the smoothness remainder)**.** *Under the conditions of Lemma E.1, let*

$$\mathcal{T}_1 = \mathcal{W}_n := \sum_{t=1}^{n}\mathbb{E}[I_t w_t], \qquad \mathcal{T}_2 := 2\sum_{t=1}^{n}\mathbb{E}[I_t b_t].$$

Suppose the cumulative smoothness remainder $\mathcal{T}_3$ is at most $\gamma\mathcal{L}\mathcal{U}_n/2$. If $\gamma\mathcal{L}/\delta \le \epsilon/32$, then

$$\mathcal{T}_2 + \mathcal{T}_3 \le \frac{1}{2}\mathcal{W}_n + \frac{10\Sigma^2 N_n}{\epsilon}, \qquad \mathcal{U}_n \le \frac{5\mathcal{W}_n}{4\epsilon} + \frac{5\Sigma^2 N_n}{\epsilon^2}. \tag{64}$$

If also $I_t = \mathbf{1}_{\{t<\tau\}}$ for a stopping time τ and, on active steps,

$$F(\boldsymbol{x}_{t+1}) \le F(\boldsymbol{x}_t) - \gamma w_t + 2\gamma b_t + \frac{\mathcal{L}\gamma^2}{2}\|\boldsymbol{u}_t\|_2^2,$$

then

$$\mathbb{E}[F(\boldsymbol{x}_{(n+1)\wedge\tau}) - F^*] + \frac{\gamma}{2}\mathcal{W}_n \le \Delta + \frac{10\gamma\Sigma^2 N_n}{\epsilon}. \tag{65}$$

The convention $\tau = \infty$ gives the unconditional version.

Proof. Lemma 4.1 yields $\mathcal{T}_2 \le 2\mathcal{B}_n$ and $\epsilon\mathcal{U}_n \le \mathcal{W}_n + \mathcal{B}_n$. Define only for this calculation $\mathcal{R}_n = 2\mathcal{B}_n + \gamma\mathcal{L}\mathcal{U}_n/2$. Since $\delta \le 1$, the step restriction and (55) give

$$\begin{aligned}\mathcal{R}_n &\le 2\Sigma\sqrt{N_n\mathcal{U}_n} + \gamma\mathcal{L}\left(\frac{2}{\delta} + \frac{1}{2}\right)\mathcal{U}_n \\ &\le 2\Sigma\sqrt{N_n\mathcal{U}_n} + \frac{\epsilon}{8}\mathcal{U}_n \le \frac{\epsilon}{4}\mathcal{U}_n + \frac{8\Sigma^2 N_n}{\epsilon}.\end{aligned} \tag{66}$$

The last inequality is Young's inequality. As $\mathcal{B}_n \le \mathcal{R}_n/2$,

$$\mathcal{R}_n \le \frac{1}{4}\mathcal{W}_n + \frac{1}{8}\mathcal{R}_n + \frac{8\Sigma^2 N_n}{\epsilon}.$$

Hence

$$\mathcal{T}_2 + \mathcal{T}_3 \le \mathcal{R}_n \le \frac{2}{7}\mathcal{W}_n + \frac{64\Sigma^2 N_n}{7\epsilon} \le \frac{1}{2}\mathcal{W}_n + \frac{10\Sigma^2 N_n}{\epsilon}.$$

Also $\mathcal{B}_n \le \mathcal{W}_n/4 + 5\Sigma^2 N_n/\epsilon$, which, together with $\epsilon\mathcal{U}_n \le \mathcal{W}_n + \mathcal{B}_n$, proves the second inequality in (64). Finally, summing the pathwise descent inequality after multiplication by I_t uses the exact identity

$$\sum_{t=1}^{n} I_t\big(F(\boldsymbol{x}_{t+1}) - F(\boldsymbol{x}_t)\big) = F(\boldsymbol{x}_{(n+1)\wedge\tau}) - F(\boldsymbol{x}_1).$$

The first inequality in (64) then gives (65). □

Proof. Write $z_t = F(\boldsymbol{x}_t) - F^*$ and define

$$\tau := \inf\{t \ge 1 : z_t > Q\}, \qquad I_t := \mathbf{1}_{\{t<\tau\}}, \qquad \mathcal{E}_T := \{\tau > T+1\}.$$

The indicators I_t are nonincreasing and $\mathscr{F}_{t-1}$-measurable. They are used only in the analysis; Algorithm 1 is not stopped or modified.

Local coefficients. For $h = \|\nabla F(\boldsymbol{x})\|_2$, apply Lemma A.1 at $\boldsymbol{x} - \nabla F(\boldsymbol{x})/(L_0 + L_1 h^q)$. The global point-pair form of B.3 and $F \ge F^*$ give

$$h^2 \le 2(L_0 + L_1 h^q)(F(\boldsymbol{x}) - F^*).$$

When $F(\boldsymbol{x}) - F^* \le Q$, at least one of $h^2 \le 4QL_0$ or $h^{2-q} \le 4QL_1$ holds. Thus $\|\nabla F(\boldsymbol{x}_t)\|_2 \le H_\rho$ on $I_t = 1$. Both the descent coefficient and the gradient-drift coefficient on an active prefix are at most L_ρ, and

$$\mathbb{E}[I_t\|\boldsymbol{\zeta}_t\|_2^2] = \mathbb{E}\left[I_t\mathbb{E}[\|\boldsymbol{\zeta}_t\|_2^2 \mid \mathscr{F}_{t-1}]\right] \le \Sigma_\rho^2.$$

The stopped quantities have finite moments. For example, on an active prefix, $\hat{\boldsymbol{m}}_t = \sum_{k\le t} p_{t,k} I_k \boldsymbol{g}_k$ and $\mathbb{E}[I_k\|\boldsymbol{g}_k\|_2^2] \le H_\rho^2 + \Sigma_\rho^2$; $\|\boldsymbol{u}_t\|_2 \le \|\hat{\boldsymbol{m}}_t\|_2/\epsilon$ then gives the required second moments. The exiting function value is integrable by the descent bound anchored at the last active iterate. No moment bound on the unrestricted post-exit trajectory is used.

Stopped sign descent. Apply Lemmas E.1 and E.2 with $\mathcal{L} = L_\rho$, $\Sigma = \Sigma_\rho$, and these predictable indicators. Since $\gamma L_\rho/\delta = \eta_\rho\epsilon \le \epsilon/32$,

$$\mathbb{E}[z_{(T+1)\wedge\tau}] + \frac{\gamma}{2}\mathcal{W}_T \le \Delta + \frac{10\gamma\Sigma_\rho^2 N_T}{\epsilon} \le A_\rho. \tag{67}$$

Here $\mathcal{W}_T = \sum_t \mathbb{E}[I_t w_t]$ and $N_T \le 3\sqrt{T}$. The choices in (15) ensure $\Delta \le \rho Q/4$ and $30\eta_\rho\Sigma_\rho^2/L_\rho \le \rho Q/4$. On $\{\tau \le T+1\}$, $z_{(T+1)\wedge\tau} = z_\tau > Q$, so

$$\mathbb{P}(\mathcal{E}_T^c) \le \frac{\mathbb{E}[z_{(T+1)\wedge\tau}]}{Q} \le \frac{\rho}{2}. \tag{68}$$

The same lemmas yield

$$\frac{\mathcal{W}_T}{T} \le C_{W,\rho}T^{-1/2}, \qquad \frac{\mathcal{U}_T}{T} \le C_{U,\rho}T^{-1/2}, \qquad \frac{1}{T}\sum_t \mathbb{E}[I_t\|\boldsymbol{e}_t\|_2] \le C_{E,\rho}T^{-1/4}. \tag{69}$$

Unweighted gradient on the good event. For $k \le t$, $I_t \le I_k$. Thus $\mathbb{E}[I_t\|\boldsymbol{g}_k\|_2^2] \le H_\rho^2 + \Sigma_\rho^2$. The convex-combination definition of $\hat{\boldsymbol{v}}_t$ gives

$$\mathbb{E}[I_t\|\hat{\boldsymbol{v}}_t\|_1] \le H_\rho^2 + \Sigma_\rho^2.$$

For $D_{t,\max} = \max_j(\sqrt{\hat{v}_t^{(j)}} + \epsilon)$, it follows that

$$\frac{1}{T}\sum_t \mathbb{E}[\mathbf{1}_{\mathcal{E}_T} D_{t,\max}] \le \epsilon + \sqrt{H_\rho^2 + \Sigma_\rho^2}.$$

Using the same reliable-sign split as in (49),

$$\|\nabla F(\boldsymbol{x}_t)\|_2 \le 2\|\boldsymbol{e}_t\|_2 + \sqrt{2D_{t,\max} w_t}.$$

Cauchy–Schwarz and (69) therefore imply

$$\mathbb{E}\left[\mathbf{1}_{\mathcal{E}_T} \frac{1}{T}\sum_t \|\nabla F(\boldsymbol{x}_t)\|_2\right] \le K_{2,\rho} T^{-1/4}. \tag{70}$$

The future-dependent indicator $\mathbf{1}_{\mathcal{E}_T}$ is used only at this final conversion, not to assert conditional mean zero of the noise. Markov's inequality gives

$$\mathbb{P}\left(\mathcal{E}_T,\ \frac{1}{T}\sum_t \|\nabla F(\boldsymbol{x}_t)\|_2 > \frac{2K_{2,\rho}}{\rho} T^{-1/4}\right) \le \frac{\rho}{2}.$$

Combining this with (68) proves (17).

□

Scope of the extension. The constant $K_{2,\rho}$ and the learning-rate constant depend on ρ. Thus (70) is not an unconditional expectation bound for the original trajectory, nor can (17) be integrated over ρ as if it concerned one fixed parameter choice. The expectation results in Section 4 instead follow from the separate, unstopped argument under B.1.

F Proofs for the normalized-network results

F.1 Network family and precise conditions

We use the Euclidean norm on the complete parameter vector, with the Frobenius norm on each matrix block. For a bilinear map $\mathcal{H}$, write $\|\mathcal{H}\|_{\mathrm{op}} = \sup_{\|u\|_2=\|v\|_2=1} \|\mathcal{H}[u,v]\|_2$. The input $h_0 = h_0(\xi)$ is independent of the trainable parameters. A network with K linear maps has the form

$$\begin{aligned} a_\ell &= \frac{W_\ell h_{\ell-1}}{\sqrt{n_{\ell-1}}}, && 1 \le \ell \le K, \\ h_\ell &= \Psi_\ell(a_\ell, \boldsymbol{\vartheta}_\ell), && 1 \le \ell < K, \\ f(\boldsymbol{\theta};\xi) &= \log\left(\sum_{j=1}^{n_K} e^{a_K^{(j)}}\right) - a_K^{(y)}, && \boldsymbol{\theta} = (W_1,\dots,W_K,\boldsymbol{\vartheta}_1,\dots,\boldsymbol{\vartheta}_{K-1}). \end{aligned} \tag{71}$$

Here $W_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$, y is the sample label, and K, n_K are fixed as the hidden widths increase. A block Ψ_ℓ may combine an activation with LayerNorm or RMSNorm. Any trainable gain or bias in that block is included in $\boldsymbol{\vartheta}_\ell$; a parameter-free block has an empty $\boldsymbol{\vartheta}_\ell$. The blocks have distinct parameters. Thus

$$d = \sum_{\ell=1}^{K} n_\ell n_{\ell-1} + \sum_{\ell=1}^{K-1} \dim(\boldsymbol{\vartheta}_\ell).$$

This is the explicit fan-in-scaled parameterization of the theorem: the optimizer acts on W_ℓ, not on $W_\ell/\sqrt{n_{\ell-1}}$.

Assumption F.1 (Bounds independent of hidden width)**.** Let Θ be an open convex parameter region. Each Ψ_ℓ is C^2 in its joint arguments, and the sample losses are twice continuously differentiable on a neighborhood of Θ. For every $\boldsymbol{\theta} \in \Theta$ and almost every example, assume

$$\begin{aligned} \frac{\|h_\ell\|_2}{\sqrt{n_\ell}} &\le A_\ell, && 0 \le \ell < K, \\ \frac{\|W_\ell\|_{\mathrm{op}}}{\sqrt{n_{\ell-1}}} &\le M_\ell, && 1 \le \ell \le K, \\ \|D\Psi_\ell(a_\ell,\boldsymbol{\vartheta}_\ell)\|_{\mathrm{op}} &\le J_\ell, \qquad \|D^2\Psi_\ell(a_\ell,\boldsymbol{\vartheta}_\ell)\|_{\mathrm{op}} \le H_\ell, && 1 \le \ell < K. \end{aligned} \tag{72}$$

All displayed constants are independent of the widths. The last two derivatives are with respect to the *joint* argument (a_ℓ, ϑ_ℓ), including mixed derivatives when block parameters are trained. The bounds hold on a common full-probability set of examples. We assume the population loss is finite at one point of Θ; for the finite-data objective in Section 3 this integrability condition is automatic.

The distinction between derivatives in the input alone and in the joint argument matters. A bound on $D_{a_\ell}\Psi_\ell$ does not by itself control gradients or curvature with respect to trainable normalization parameters. Section F.4 gives sufficient conditions for both versions.

F.2 Total gradient variance

Proof of the noise bound in Theorem 5.1. For softmax cross-entropy, put $p = \mathrm{softmax}(a_K)$. The output gradient is $p - e_y$, so

$$\|\delta_K\|_2^2 = \|p - e_y\|_2^2 \le 2, \qquad \delta_\ell := \nabla_{a_\ell} f(\boldsymbol{\theta}; \xi).$$

For $\ell < K$, differentiating the block composition gives

$$\delta_\ell = D_a \Psi_\ell(a_\ell, \vartheta_\ell)^\top \frac{W_{\ell+1}^\top}{\sqrt{n_\ell}} \delta_{\ell+1}. \tag{73}$$

Since the partial Jacobian norm is at most the joint Jacobian norm, define

$$B_\ell := \sqrt{2} \prod_{r=\ell}^{K-1} J_r M_{r+1}. \qquad \text{Then} \qquad \|\delta_\ell\|_2 \le B_\ell. \tag{74}$$

The empty product is one. For each weight matrix,

$$\nabla_{W_\ell} f = \frac{\delta_\ell h_{\ell-1}^\top}{\sqrt{n_{\ell-1}}}, \qquad \|\nabla_{W_\ell} f\|_F^2 = \frac{\|\delta_\ell\|_2^2 \|h_{\ell-1}\|_2^2}{n_{\ell-1}} \le A_{\ell-1}^2 B_\ell^2. \tag{75}$$

For a trainable block parameter,

$$\nabla_{\vartheta_\ell} f = D_{\vartheta} \Psi_\ell(a_\ell, \vartheta_\ell)^\top \frac{W_{\ell+1}^\top}{\sqrt{n_\ell}} \delta_{\ell+1}, \qquad \|\nabla_{\vartheta_\ell} f\|_2 \le J_\ell M_{\ell+1} B_{\ell+1} = B_\ell.$$

Let $\chi_\ell = 1$ if ϑ_ℓ is trained and $\chi_\ell = 0$ otherwise. Summing the squared norms of *all* parameter blocks yields

$$\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; \xi)\|_2^2 \le G_*^2, \qquad G_*^2 := 1 + \sum_{\ell=1}^{K} A_{\ell-1}^2 B_\ell^2 + \sum_{\ell=1}^{K-1} \chi_\ell B_\ell^2. \tag{76}$$

The harmless added 1 makes $G_* > 0$. Since depth and the bounds in Assumption F.1 are fixed, G_* does not grow with d. Jensen's inequality also gives

$$\|\nabla F(\boldsymbol{\theta})\|_2 = \|\mathbb{E}_\xi \nabla f(\boldsymbol{\theta}; \xi)\|_2 \le G_*, \tag{77}$$

which proves (19).

The derivative bound permits differentiation under the data expectation on Θ. For $G(\xi) = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; \xi)$ and $\bar{G} = \nabla F(\boldsymbol{\theta})$,

$$\mathbb{E}_\xi \|G(\xi) - \bar{G}\|_2^2 = \mathbb{E}_\xi \|G(\xi)\|_2^2 - \|\bar{G}\|_2^2 \le G_*^2. \tag{78}$$

At an adaptive iterate $\boldsymbol{\theta}_t$, condition on $\mathscr{F}_{t-1}$ and draw B_{mb} fresh independent examples. For their averaged gradient, conditional unbiasedness and the vanishing cross terms give

$$\mathbb{E}\big[\|\boldsymbol{g}_t - \nabla F(\boldsymbol{\theta}_t)\|_2^2 \mid \mathscr{F}_{t-1}\big] = \frac{1}{B_{\mathrm{mb}}} \mathbb{E}_\xi \|G(\xi) - \bar{G}\|_2^2 \le \frac{G_*^2}{B_{\mathrm{mb}}}. \tag{79}$$

This proves (20), which bounds the trace of the conditional gradient covariance. Set $\sigma_0 = G_*/\sqrt{B_{\mathrm{mb}}}$ and choose $\sigma_1 = 1$. For every $p \in [0, 2]$, (79) implies

$$\mathbb{E}[\|\boldsymbol{g}_t - \nabla F(\boldsymbol{\theta}_t)\|_2^2 \mid \mathscr{F}_{t-1}] \le \sigma_0^2 \le \sigma_0^2 + \sigma_1^2 \|\nabla F(\boldsymbol{\theta}_t)\|_2^p. \tag{80}$$

Thus C.3 admits coefficients $\sigma_0, \sigma_1 = O(1)$ for this mini-batch oracle; their boundedness is not an extra oracle assumption. Moreover, (77) gives

$$\sigma_0^2 + \sigma_1^2 \|\nabla F(\boldsymbol{\theta})\|_2^p \le \sigma_0^2 + \sigma_1^2 G_*^p =: \Sigma_{\mathrm{D}}^2 = O(1). \tag{81}$$

These are admissible, nonunique upper-bound coefficients. Any fixed positive σ_1 works with the same σ_0. The construction concerns the averaged sample-gradient oracle above; it does not bound the coefficients of an unrelated oracle that adds arbitrary dimension-dependent noise. □

F.3 PARAMETER-SPACE SMOOTHNESS

Proof of the smoothness bound in Theorem 5.1. We bound derivatives with respect to the full parameter vector, rather than only the input of a layer. Write a perturbation as $V = (V_1, \dots, V_K, v_1, \dots, v_{K-1})$, with $\|V\|_2^2 = \sum_\ell \|V_\ell\|_F^2 + \sum_\ell \|v_\ell\|_2^2$. Set $R_0 = Q_0 = 0$. For $1 \le \ell \le K$, define

$$S_\ell := A_{\ell-1} + M_\ell R_{\ell-1}, \qquad T_\ell := 2R_{\ell-1} + M_\ell Q_{\ell-1}, \tag{82}$$

and, for $\ell < K$, set

$$R_\ell := J_\ell \sqrt{S_\ell^2 + \chi_\ell}, \qquad Q_\ell := H_\ell(S_\ell^2 + \chi_\ell) + J_\ell T_\ell. \tag{83}$$

These are scalar recursions with fixed depth and coefficients independent of width.

We prove inductively that

$$\begin{aligned} \|Da_\ell[V]\|_2 &\le S_\ell \|V\|_2, \quad \|D^2 a_\ell[V,U]\|_2 \le T_\ell \|V\|_2 \|U\|_2, \\ \|Dh_\ell[V]\|_2 &\le R_\ell \|V\|_2, \quad \|D^2 h_\ell[V,U]\|_2 \le Q_\ell \|V\|_2 \|U\|_2. \end{aligned} \tag{84}$$

The statements for h_ℓ are needed only when $\ell < K$. Since h_0 is parameter-independent, the induction starts with zero derivatives. Differentiating the linear map gives

$$Da_\ell[V] = \frac{V_\ell h_{\ell-1} + W_\ell Dh_{\ell-1}[V]}{\sqrt{n_{\ell-1}}}.$$

The normalized activation bound and the scaled weight bound imply

$$\|Da_\ell[V]\|_2 \le A_{\ell-1}\|V_\ell\|_F + M_\ell R_{\ell-1}\|V\|_2 \le S_\ell \|V\|_2.$$

For a second perturbation $U = (U_1, \dots, U_K, u_1, \dots, u_{K-1})$,

$$D^2 a_\ell[V,U] = \frac{V_\ell Dh_{\ell-1}[U] + U_\ell Dh_{\ell-1}[V] + W_\ell D^2 h_{\ell-1}[V,U]}{\sqrt{n_{\ell-1}}}. \tag{85}$$

Using $n_{\ell-1} \ge 1$ bounds the first two terms by $2R_{\ell-1}\|V\|_2\|U\|_2$; the third is at most $M_\ell Q_{\ell-1}\|V\|_2\|U\|_2$. This proves the claimed bound with T_ℓ.

Let $b_\ell = (a_\ell, \vartheta_\ell)$. Then

$$Db_\ell[V] = (Da_\ell[V], v_\ell), \qquad \|Db_\ell[V]\|_2 \le \sqrt{S_\ell^2 + \chi_\ell}\|V\|_2, \qquad D^2 b_\ell[V,U] = (D^2 a_\ell[V,U], 0).$$

The first- and second-order chain rules give

$$\begin{aligned} Dh_\ell[V] &= D\Psi_\ell(b_\ell) Db_\ell[V], \\ D^2 h_\ell[V,U] &= D^2\Psi_\ell(b_\ell)[Db_\ell[V], Db_\ell[U]] + D\Psi_\ell(b_\ell)(D^2 a_\ell[V,U], 0). \end{aligned}$$

Assumption F.1 now gives exactly the constants R_ℓ, Q_ℓ in (83). This also accounts for the mixed derivatives involving trainable block parameters.

For the output loss $\ell(z,y) = \log \sum_j e^{z_j} - z_y$,

$$\nabla_z \ell = p - e_y, \qquad \nabla_z^2 \ell = \operatorname{diag}(p) - pp^\top, \qquad p = \operatorname{softmax}(z).$$

The first norm is at most $\sqrt{2}$. Each absolute row sum of the Hessian is $2p_i(1-p_i) \le 1/2$, so its operator norm is at most $1/2$. These bounds concern derivatives of the loss, not a bound on its value. Applying the chain rule at $z = a_K$ yields

$$\begin{aligned} |D^2 f(\boldsymbol{\theta};\xi)[V,U]| &\le \frac{1}{2}\|Da_K[V]\|_2 \|Da_K[U]\|_2 + \sqrt{2}\|D^2 a_K[V,U]\|_2 \\ &\le \left(\frac{1}{2}S_K^2 + \sqrt{2}T_K\right)\|V\|_2\|U\|_2. \end{aligned} \tag{86}$$

Thus we may choose the positive constant

$$L_* := 1 + \frac{1}{2}S_K^2 + \sqrt{2}T_K, \qquad \|\nabla_{\boldsymbol{\theta}}^2 f(\boldsymbol{\theta};\xi)\|_{\mathrm{op}} \le L_*. \tag{87}$$

The uniform derivative bounds justify averaging the Hessian over examples, giving $\|\nabla^2 F(\boldsymbol{\theta})\|_{\mathrm{op}} \le L_*$ on Θ. For $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, convexity keeps the segment between them inside Θ, hence

$$\nabla F(\boldsymbol{\theta}') - \nabla F(\boldsymbol{\theta}) = \int_0^1 \nabla^2 F(\boldsymbol{\theta} + s(\boldsymbol{\theta}' - \boldsymbol{\theta}))(\boldsymbol{\theta}' - \boldsymbol{\theta})\, ds.$$

Taking norms proves (21). This is the Lipschitz constant of the *parameter gradient*, not the input-output Lipschitz constant of the network. Set $L_0 = L_*$ and $L_1 = 1$. For any $q \in [0, 1]$ and $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$,

$$\begin{aligned} \|\nabla F(\boldsymbol{\theta}') - \nabla F(\boldsymbol{\theta})\|_2 &\le L_* \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2 \\ &\le \big(L_0 + L_1 \|\nabla F(\boldsymbol{\theta})\|_2^q\big) \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2. \end{aligned} \tag{88}$$

Hence the B.3 inequality admits $L_0, L_1 = O(1)$ on Θ, while B.1 holds there with $L = L_0$. Any fixed positive L_1 is admissible. This follows from the stronger smoothness estimate on the stated network region; it is not an identification of local bounds with global B.3. □

F.4 First and second derivatives of the normalization blocks

LayerNorm and RMSNorm are defined in Ba et al. (2016) and Zhang & Sennrich (2019), respectively. The bounds below are derived directly for their regularized forms. For parameter-free RMSNorm, write

$$s = \frac{\|z\|_2^2}{n}, \qquad r = s + \varepsilon_{\mathrm{N}}, \qquad \mathcal{N}(z) = z r^{-1/2}, \qquad \varepsilon_{\mathrm{N}} > 0.$$

Its normalized energy and Jacobian satisfy

$$\frac{\|\mathcal{N}(z)\|_2^2}{n} = \frac{s}{s + \varepsilon_{\mathrm{N}}} \le 1, \qquad D\mathcal{N}(z) = r^{-1/2} I - \frac{z z^\top}{n r^{3/2}}, \qquad \|D\mathcal{N}(z)\|_{\mathrm{op}} \le \varepsilon_{\mathrm{N}}^{-1/2}. \tag{89}$$

Indeed, the eigenvalues are $r^{-1/2}$ orthogonally to z and $\varepsilon_{\mathrm{N}} r^{-3/2}$ in the direction of z. Differentiating once more gives the bilinear map

$$D^2\mathcal{N}(z)[u, v] = -\frac{\langle z, v\rangle u + \langle z, u\rangle v + \langle u, v\rangle z}{n r^{3/2}} + \frac{3\langle z, u\rangle \langle z, v\rangle z}{n^2 r^{5/2}}. \tag{90}$$

Consequently,

$$\|D^2\mathcal{N}(z)\|_{\mathrm{op}} \le \frac{3}{\sqrt{n}} \left(\frac{\sqrt{s}}{(s + \varepsilon_{\mathrm{N}})^{3/2}} + \frac{s^{3/2}}{(s + \varepsilon_{\mathrm{N}})^{5/2}} \right) \le \frac{6}{\sqrt{n}\, \varepsilon_{\mathrm{N}}} \le \frac{6}{\varepsilon_{\mathrm{N}}}. \tag{91}$$

Each fraction in parentheses is at most $1/(s + \varepsilon_{\mathrm{N}})$. The resulting first- and second-order bounds have constants independent of the layer width.

For LayerNorm, let $P = I - \mathbf{1}\mathbf{1}^\top / n$. It can be written as $\mathcal{N}_{\mathrm{LN}}(z) = \mathcal{N}_{\mathrm{RMS}}(Pz)$. Since $\|P\|_{\mathrm{op}} \le 1$, the same estimates give

$$\frac{\|\mathcal{N}_{\mathrm{LN}}(z)\|_2^2}{n} \le 1, \qquad \|D\mathcal{N}_{\mathrm{LN}}(z)\|_{\mathrm{op}} \le \varepsilon_{\mathrm{N}}^{-1/2}, \qquad \|D^2\mathcal{N}_{\mathrm{LN}}(z)\|_{\mathrm{op}} \le \frac{6}{\sqrt{n}\, \varepsilon_{\mathrm{N}}}. \tag{92}$$

For a coordinate-wise C^2 activation with $|\phi'| \le c_1$ and $|\phi''| \le c_2$, the Euclidean operator norms obey $\|D\phi\|_{\mathrm{op}} \le c_1$ and $\|D^2\phi\|_{\mathrm{op}} \le c_2$. Thus, for $r_\phi(a) = \mathcal{N}(\phi(a))$,

$$\|D r_\phi(a)\|_{\mathrm{op}} \le b_1 := \frac{c_1}{\sqrt{\varepsilon_{\mathrm{N}}}}, \qquad \|D^2 r_\phi(a)\|_{\mathrm{op}} \le b_2 := \frac{6 c_1^2}{\varepsilon_{\mathrm{N}}} + \frac{c_2}{\sqrt{\varepsilon_{\mathrm{N}}}}. \tag{93}$$

The same chain-rule argument covers a smooth activation after normalization. Nonsmooth activations are not included in this C^2 argument.

Fixed and trainable affine normalization parameters. For $\Psi(a) = \alpha \circ r_\phi(a) + b$ with fixed coefficients satisfying $\|\alpha\|_\infty \le \Gamma$ and $\|b\|_2/\sqrt{n} \le A_b$, one can take $A = \Gamma + A_b$, $J = \Gamma b_1$, and $H = \Gamma b_2$. When α, b are trained, a sufficient additional condition for the joint derivative bounds is $\|r_\phi(a)\|_\infty \le R_{\mathrm{N}}$ throughout the region, with R_{N} independent of n. For a direction (u, a', b'),

$$D\Psi[u, a', b'] = \alpha \circ D r_\phi[u] + a' \circ r_\phi + b'.$$

Hence

$$\|D\Psi\|_{\mathrm{op}} \le \sqrt{\Gamma^2 b_1^2 + R_{\mathrm{N}}^2 + 1}.$$

The joint second derivative contains $\alpha \circ D^2 r_\phi[u,v]$, $a' \circ Dr_\phi[v]$, and $a'' \circ Dr_\phi[u]$, giving

$$\|D^2\Psi\|_{\mathrm{op}} \le \Gamma b_2 + 2b_1.$$

These bounds include the mixed gain-input derivatives. The coordinate bound R_{N} is an additional condition for this trained-gain case: the energy bound $\|r_\phi\|_2^2/n \le 1$ alone does not imply it. Likewise, normalization does not by itself bound the scaled weights or guarantee that a training trajectory stays in Θ. BatchNorm is not covered by the independent-example calculation above because its training-time statistics couple examples in the same batch.

F.5 Dimension-free Adam iteration bound

Initial softmax cross-entropy gap. Take $F^* = \inf F \ge 0$, since softmax cross-entropy is nonnegative. With the fixed output dimension n_K, for every logit vector,

$$0 \le \ell(z,y) = \log \sum_{j=1}^{n_K} e^{z_j} - z_y \le \log n_K + 2\|z\|_\infty.$$

For the initial logits in the theorem, $B_{\mathrm{init}} = \mathbb{E}_\xi \|a_K(\boldsymbol{\theta}_1;\xi)\|_\infty$, so

$$0 \le \Delta = F(\boldsymbol{\theta}_1) - F^* \le F(\boldsymbol{\theta}_1) \le \log n_K + 2B_{\mathrm{init}} = O(1). \tag{94}$$

This proves the remaining assertion in (23). The initialization condition bounds the logits, not an independently postulated objective gap. Softmax cross-entropy alone does not bound the loss for arbitrary initial logits.

Regional dimension-free bound under trajectory containment. We justify the Adam argument directly on Θ. By the almost-sure trajectory condition and convexity of Θ, every update segment satisfies the descent and gradient-drift inequalities with coefficient L_*. By Assumption C.3 and (19),

$$\mathbb{E}[\|\boldsymbol{g}_t - \nabla F(\boldsymbol{\theta}_t)\|_2^2 \mid \mathscr{F}_{t-1}] \le \Sigma_{\mathrm{D}}^2 := \sigma_0^2 + \sigma_1^2 G_*^p, \tag{95}$$

and $\Sigma_{\mathrm{D}} = O(1)$ by the coefficient construction (22). This uses admissible C.3 coefficients derived from the stronger bounded-variance estimate, not a separate assumption that the noise grows with the gradient.

Let $\delta = 1 - \beta_1 = T^{-1/2}$ and $\gamma = C_2 T^{-1/2}$. The parameter choice in the corollary means

$$C_2 = \frac{c\epsilon}{L_0 \max\{32, 60p\sigma_1^2\}}, \qquad 0 < c \le 1,$$

where c is fixed across widths. Thus $C_2 L_* = C_2 L_0 \le \epsilon/32$. Applying the weighted tracking and absorption argument of Section 4, with $L = L_0$ and $S = \Sigma_{\mathrm{D}}$, gives, for every prefix $N \le T$,

$$\mathbb{E}[F(\boldsymbol{\theta}_{N+1}) - F^*] + \frac{\gamma}{2} W_N \le \Delta + \frac{10\gamma\Sigma_{\mathrm{D}}^2 K_N}{\epsilon}, \qquad \epsilon U_N \le \frac{5}{4} W_N + \frac{5\Sigma_{\mathrm{D}}^2 K_N}{\epsilon}. \tag{96}$$

Here $\Delta = O(1)$ follows from (94). No global self-bounding step is used: the gradient bound already gives (95), and convexity of Θ keeps every update segment in the region where the smoothness estimate holds.

Define

$$\begin{aligned} C_W^{\mathrm{D}} &:= \frac{2\Delta}{C_2} + \frac{60\Sigma_{\mathrm{D}}^2}{\epsilon}, & C_U^{\mathrm{D}} &:= \frac{5C_W^{\mathrm{D}}}{4\epsilon} + \frac{15\Sigma_{\mathrm{D}}^2}{\epsilon^2}, \\ C_E^{\mathrm{D}} &:= \sqrt{3}\,\Sigma_{\mathrm{D}} + C_2 L_* \sqrt{C_U^{\mathrm{D}}}, & D_{\mathrm{D}} &:= \epsilon + G_*. \end{aligned} \tag{97}$$

Since $K_T \le 3\sqrt{T}$, (96) and the first-order tracking bound give

$$\frac{W_T}{T} \le C_W^{\mathrm{D}} T^{-1/2}, \qquad \frac{U_T}{T} \le C_U^{\mathrm{D}} T^{-1/2}, \qquad \frac{1}{T}\sum_t \mathbb{E}\|\boldsymbol{e}_t\|_2 \le C_E^{\mathrm{D}} T^{-1/4}.$$

Moreover, (76) implies $\|\boldsymbol{g}_t\|_2 \le G_*$ for every sampled mini-batch. Hence $\|\hat{\boldsymbol{v}}_t\|_1 \le G_*^2$ and $D_{t,\max} \le D_{\mathrm{D}}$. The reliable-sign conversion and Cauchy–Schwarz therefore yield

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\boldsymbol{\theta}_t)\|_2 \le \left(2C_E^{\mathrm{D}} + \sqrt{2D_{\mathrm{D}}C_W^{\mathrm{D}}}\right)T^{-1/4} =: C_{\mathrm{DNN}}T^{-1/4}. \tag{98}$$

The connection to (18) is explicit. By (22),

$$L_* = L_0, \qquad G_* = \sqrt{B_{\mathrm{mb}}}\,\sigma_0, \qquad \Sigma_{\mathrm{D}}^2 = \sigma_0^2 + \sigma_1^2(\sqrt{B_{\mathrm{mb}}}\,\sigma_0)^p. \tag{99}$$

Substituting these identities and the displayed choice of C_2 into (97)–(98) shows that, for fixed $p, B_{\mathrm{mb}}, \epsilon, c$, the problem dependence is precisely

$$C_{\mathrm{DNN}} = C(L_0, L_1, \sigma_0, \sigma_1, \Delta),$$

with no use of L_1 in this smooth expectation argument. All five quantities have upper bounds independent of width by Theorem 5.1; also $L_0 \ge 1$ and C_2 is bounded away from zero across the family. Hence every expression in (97) is $O(1)$ and so is C_{DNN}.

For clarity, (18) denotes a dependency pattern, not one numerical constant shared by the expectation and high-probability theorems. In Corollary 4.4, the constant is $2K_{2,\rho}/\rho$, with fixed p, q, ρ, ϵ suppressed. Its formulas (15)–(16) likewise contain no other problem quantity. This observation applies when that corollary's global hypotheses hold; the present DNN expectation proof only needs the stated regional bounds and almost-sure containment.

This corollary concerns the original, unprojected Adam iterates under the stated *almost-sure containment assumption*, not an algorithm that is stopped or projected into Θ. □

F.6 Trajectory certification from initialization

We prove Theorem 5.2. Here the affine normalization parameters are fixed. For regularized LayerNorm/RMSNorm and the smooth activations in Section F.4, the activation-energy and block-derivative bounds are global. The only quantities that must be kept in a width-uniform region are therefore the scaled spectral norms of $W_2, \dots, W_K$. The first layer is different: M_1 is absent from (74), and $R_0 = Q_0 = 0$ gives $S_1 = A_0$ and $T_1 = 0$ in (82). Thus no scaled spectral-norm bound on W_1 is needed.

Assume

$$\frac{\|W_{\ell,1}\|_{\mathrm{op}}}{\sqrt{n_{\ell-1}}} \le M_\ell^{(0)}, \qquad 2 \le \ell \le K,$$

with $M_\ell^{(0)} = O(1)$, and fix $r > 0$ independent of width. Define

$$\Theta_r := \left\{\boldsymbol{\theta} : \frac{\|W_\ell - W_{\ell,1}\|_{\mathrm{op}}}{\sqrt{n_{\ell-1}}} < r,\ 2 \le \ell \le K\right\}, \qquad \Theta_{2r} \text{ analogously.}$$

On Θ_{2r}, the proof of Theorem 5.1 applies with $M_\ell = M_\ell^{(0)} + 2r$ for $\ell \ge 2$ and gives width-independent $G_+, L_+, \Sigma_+ \ge 1$ satisfying

$$\|\nabla f(\boldsymbol{\theta};\xi)\|_2 \le G_+, \quad \|\nabla^2 F(\boldsymbol{\theta})\|_{\mathrm{op}} \le L_+, \quad \mathbb{E}[\|\boldsymbol{\zeta}_t\|_2^2 \mid \mathscr{F}_{t-1}] \le \Sigma_+^2. \tag{100}$$

Choose

$$\delta = 1 - \beta_1 = T^{-1/2}, \qquad \gamma = C_2 T^{-1/2}, \qquad C_2 L_+ \le \epsilon/32, \quad C_2 G_+/\epsilon \le r/2. \tag{101}$$

As $\widehat{m}_t$ is a convex combination of past stochastic gradients, every active history in Θ_{2r} obeys

$$\|\boldsymbol{u}_t\|_2 \le \|\widehat{m}_t\|_2/\epsilon \le G_+/\epsilon. \tag{102}$$

The last condition in (101) ensures that an update starting in Θ_r stays inside Θ_{2r}, including the exit step.

Pathwise containment. For $s \le T+1$, while previous iterates remain in Θ_r,

$$\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_1\|_2 \le \gamma \sum_{t<s} \|\boldsymbol{u}_t\|_2 \le \frac{C_2 G_+}{\epsilon}\sqrt{T}. \tag{103}$$

If

$$n_{\min} \ge c_{\det} T, \qquad c_{\det} := \frac{4C_2^2 G_+^2}{\epsilon^2 r^2}, \tag{104}$$

then the right side of (103) is at most $(r/2)\sqrt{n_{\min}}$. Since $\|A\|_{\mathrm{op}} \le \|A\|_F$, no W_ℓ, $\ell \ge 2$, can reach the boundary of Θ_r. A first-exit contradiction proves almost-sure containment for all T steps.

Stopped update energy and exit probability. Let

$$\tau := \inf\{t \ge 1 : \boldsymbol{\theta}_t \notin \Theta_r\}, \qquad I_t := \mathbf{1}_{\{t<\tau\}}.$$

Lemmas E.1–E.2 apply with $\mathcal{L} = L_+$ and $\Sigma = \Sigma_+$. Put

$$W_T^+ := \sum_{t=1}^{T} \mathbb{E}[I_t T_{1,t}], \qquad U_T^+ := \sum_{t=1}^{T} \mathbb{E}[I_t \|\boldsymbol{u}_t\|_2^2].$$

Since $N_T \le 3\sqrt{T}$, (65) and $F \ge F^*$ give

$$W_T^+ \le C_W^+ \sqrt{T}, \qquad U_T^+ \le C_U^+ \sqrt{T}, \tag{105}$$

where

$$C_W^+ := \frac{2\Delta}{C_2} + \frac{60\Sigma_+^2}{\epsilon}, \qquad C_U^+ := \frac{5C_W^+}{4\epsilon} + \frac{15\Sigma_+^2}{\epsilon^2}. \tag{106}$$

Furthermore,

$$\mathbb{E} \max_{s \le (T+1)\wedge\tau} \|\boldsymbol{\theta}_s - \boldsymbol{\theta}_1\|_2^2 \le \gamma^2 T\, U_T^+ \le C_2^2 C_U^+ \sqrt{T}. \tag{107}$$

Leaving Θ_r requires total parameter displacement at least $r\sqrt{n_{\min}}$, hence

$$\mathbb{P}(\tau \le T+1) \le \frac{C_2^2 C_U^+}{r^2} \frac{\sqrt{T}}{n_{\min}}. \tag{108}$$

Thus this probability is at most $\rho/2$ if $n_{\min} \ge c_{\mathrm{hp}}\sqrt{T}/\rho$, where $c_{\mathrm{hp}} := 2C_2^2 C_U^+ / r^2$.

Squared stationarity. Write $\boldsymbol{e}_t = \boldsymbol{n}_t - \boldsymbol{d}_t$ as in (59). Equation (60) yields

$$\sum_{t=1}^{T} \mathbb{E}[I_t \|\boldsymbol{n}_t\|_2^2] \le 3\Sigma_+^2 \sqrt{T}.$$

For the drift, Young's convolution inequality gives

$$\sum_{t=1}^{T} \mathbb{E}[I_t \|\boldsymbol{d}_t\|_2^2] \le \left(\frac{\gamma L_+}{\delta}\right)^2 U_T^+ \le (C_2 L_+)^2 C_U^+ \sqrt{T}.$$

Consequently,

$$\sum_{t=1}^{T} \mathbb{E}[I_t \|\boldsymbol{e}_t\|_2^2] \le C_{E,2}^+ \sqrt{T}, \qquad C_{E,2}^+ := 6\Sigma_+^2 + 2(C_2 L_+)^2 C_U^+. \tag{109}$$

On $I_t = 1$, every gradient entering $\widehat{v}_t$ is generated before exit, so $D_{t,\max} \le D_+ := \epsilon + G_+$. Squaring the reliable-sign split coordinatewise gives

$$I_t \|\nabla F(\boldsymbol{\theta}_t)\|_2^2 \le 4I_t \|\boldsymbol{e}_t\|_2^2 + 2D_+ I_t T_{1,t}. \tag{110}$$

Combining the preceding estimates,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[I_t \|\nabla F(\boldsymbol{\theta}_t)\|_2^2] \le C_{\mathrm{sq}} T^{-1/2}, \qquad C_{\mathrm{sq}} := 4C_{E,2}^+ + 2D_+ C_W^+. \tag{111}$$

Under (104), $I_t \equiv 1$ almost surely, proving (24). Otherwise, Markov's inequality gives the bound $2C_{\mathrm{sq}}/(\rho\sqrt{T})$ for the active average squared gradient with probability at least $1 - \rho/2$. Intersecting with $\{\tau > T+1\}$ from (108) and applying Cauchy–Schwarz over time proves (25). All constants depend on fixed depth, normalization/activation constants, initialization margins, $B_{\mathrm{mb}}, \epsilon, C_2$, and Δ, but not on hidden width or total parameter count.