IMPROVED CONVERGENCE OF LARGE STEPSIZE GRADIENT DESCENT FOR LOGISTIC REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study gradient descent (GD) with a large constant stepsize for logistic regression on linearly separable data. Existing analysis shows an accelerated rate of $\widetilde{O}(1/\sqrt{\epsilon})$ to reach loss ϵ with an aggressive stepsize, although the loss may initially oscillate. Tighter control of the oscillatory dynamics has been available only for two-dimensional data. We prove a substantially faster rate in arbitrary dimension: GD with a large stepsize $\eta = 1/\epsilon$ reaches loss ϵ within $O(\ln^p(1/\epsilon))$ steps, where p depends only on the margin and the rank of the data. Our proof improves the bound on the transition time of GD from the oscillatory to the stable phase, after which the loss decreases monotonically. We split the oscillatory phase into recursively nested intervals. The margin and the rank bound the nesting depth, and a counting argument bounds the number of intervals at each depth, together yielding the polylogarithmic step complexity.

1 INTRODUCTION

Gradient descent (GD) and its variants are the most widely used optimization methods in modern machine learning. To minimize an objective $L : \mathbb{R}^d \to \mathbb{R}$, GD starts from an initialization $\mathbf{w}_0 \in \mathbb{R}^d$ and iterates

$$\mathbf{w}_{t+1} := \mathbf{w}_t - \eta_t \nabla L(\mathbf{w}_t), \qquad t = 0, 1, 2, \ldots, \tag{GD}$$

where $\mathbf{w}_t \in \mathbb{R}^d$ is the trainable parameter and $(\eta_t)_{t\geq 0}$ are the stepsizes, which may in general vary across iterations. In this work we study GD with a constant stepsize $\eta_t \equiv \eta$ that is *large*, by which we mean a stepsize so large that the loss $L(\mathbf{w}_t)$ need not decrease monotonically along the iterates.

Classical optimization theory studies GD through the *descent lemma*: if L is β-smooth and $\eta_t < 2/\beta$, then every step of GD decreases the loss, and convergence rates follow by summing the per-step decrease (Nesterov, 2018). The vast majority of convergence guarantees for GD, for convex and nonconvex objectives alike, are built on this monotone decrease and therefore require a small stepsize (Nesterov, 2018; Lan, 2020). The practice of deep learning looks different. Neural networks are typically trained with stepsizes so large that the training loss does not decrease monotonically but oscillates locally, while still converging in the long run (Wu et al., 2018; Xing et al., 2018; Lewkowycz et al., 2020; Cohen et al., 2021), and such large stepsizes are often necessary for good performance. Cohen et al. (2021) named this regime the *edge of stability* (EoS): the sharpness of the loss—the largest eigenvalue of its Hessian—rises during training until it reaches the stability threshold $2/\eta$ and then hovers just above it, so that training proceeds precisely where the descent lemma no longer applies. Understanding why GD converges, and why it can even converge faster, without monotone descent has become a central question in the theory of deep learning optimization.

A recent line of work shows that stepsizes beyond the descent lemma can be provably beneficial, in that GD converges faster with stepsizes that classical theory does not cover. For smooth convex optimization, stepsize schedules with occasional long steps accelerate GD beyond the $O(1/t)$ rate of the descent regime (Altschuler & Parrilo, 2025; Grimmer et al., 2025; Zhang et al., 2025). For logistic regression on linearly separable data, Wu et al. (2024) showed that GD with a large constant stepsize η converges despite an initial *unstable phase* in which the loss oscillates. GD exits this phase within $O(\eta)$ steps and then enters a *stable phase* in which the loss decreases monotonically at rate $\widetilde{O}(1/(\eta t))$. Consequently, with a budget of T steps, the aggressive choice $\eta = \Theta(T)$ balances the two phases and yields an accelerated $\widetilde{O}(1/T^2)$ loss. Zhang et al. (2026) subsequently showed

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.

that large stepsizes that adapt to the current loss reach an arbitrarily small loss after a short burn-in period, which is minimax optimal among first-order methods, and Crawshaw & Liu (2026) gave a fine-grained analysis of the oscillatory dynamics of constant stepsize GD in two dimensions, proving that the unstable phase ends within a number of steps that is independent of η, which is tight up to logarithmic factors.

Among these works, Wu et al. (2024) and Crawshaw & Liu (2026) are the closest to ours, as we consider the same problem of GD with a large constant stepsize for logistic regression on linearly separable data. In the analysis of Wu et al. (2024), the length of the unstable phase is bounded only linearly in the stepsize. This is the bottleneck of their rate, since it caps the stepsize that can be used with a given budget, and they conjectured that the bound is improvable. Crawshaw & Liu (2026) confirmed this in two dimensions, where GD with a sufficiently large stepsize consequently reaches any target loss within a number of steps that is independent of the target. Their analysis, however, tracks the oscillation of the iterates across the max-margin direction, which is a one-dimensional motion only in the plane, and whether a comparable improvement is possible in arbitrary dimension has remained open. This question motivates our work.

In this paper, we consider logistic regression for binary classification. Given a dataset $(\mathbf{x}_i, y_i)_{i=1}^n \subset \mathbb{R}^d \times \{\pm 1\}$, the empirical logistic loss is

$$L(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})\right), \tag{1}$$

and we run (GD) on L with a large constant stepsize $\eta > 0$ from the initialization $\mathbf{w}_0 := \mathbf{0}$, as in Wu et al. (2024) and Crawshaw & Liu (2026). We show that GD with a large constant stepsize achieves a substantially faster convergence rate for logistic regression on linearly separable data in arbitrary dimension. Our contributions are as follows.

- **An improved convergence rate.** We prove that GD with a large constant stepsize $\eta = 1/\epsilon$ reaches loss ϵ within $O(\ln^p(1/\epsilon))$ steps, where the exponent p depends only on the margin and the rank of the data. For fixed problem dependent parameters, this improves the $\widetilde{O}(1/\sqrt{\epsilon})$ step complexity of Wu et al. (2024) to polylogarithmic in $1/\epsilon$ (see Table 1). At the core of this result is a new bound on the transition time from the unstable to the stable phase. We show that the unstable phase, if GD ever enters it, ends within $O((\ln \eta)^p)$ steps, in contrast to the $O(\eta)$ bound of Wu et al. (2024).
- **New proof techniques.** Our proof is based on a counting argument that differs from both existing analyses. It neither relies on the time-averaged gradient potential as in Wu et al. (2024), which leads to a bound linear in η, nor tracks the oscillations of the iterates as in Crawshaw & Liu (2026), which restricts the analysis to two dimensions. Specifically, (i) we split the unstable phase into recursively nested intervals whose average update directions are pairwise separated; (ii) a geometric capacity lemma shows that only a limited number of pairwise separated directions can coexist, which bounds the nesting depth in terms of the margin and the rank; and (iii) a counting argument bounds the number of intervals at each depth with a factorial saving, which yields the polylogarithmic bound.

Notation. We use $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ to denote the Euclidean inner product and norm, respectively. We write $\operatorname{conv} S$ for the convex hull of a set S, and $A \,\dot\cup\, B$ for the union of disjoint sets A and B. We use the standard big-O and big-Ω notation, with $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ hiding polylogarithmic factors.

1.1 Related Work

Logistic regression. With a small stepsize obeying the descent lemma, GD on linearly separable data converges in direction to the max-margin classifier (Soudry et al., 2018; Ji & Telgarsky, 2019), where the latter also covers non-separable data; these results have been extended to other optimization geometries and to stochastic gradient descent (Gunasekar et al., 2018; Nacson et al., 2019b). Faster margin maximization is possible with stepsizes that grow with the inverse loss (Nacson et al., 2019a; Ji & Telgarsky, 2021) or with momentum (Ji et al., 2021; Wang et al., 2023).

For GD with a large constant stepsize, Wu et al. (2023) proved convergence and asymptotic margin maximization for arbitrarily large stepsizes on non-degenerate data, with constants exponential in η, and Wu et al. (2024) removed the exponential dependence and established the two-phase analysis

and the $\widetilde{O}(1/T^2)$ acceleration discussed above. Their analysis has since been extended to two-layer networks (Cai et al., 2024), regularized logistic regression (Wu et al., 2025), local GD (Crawshaw et al., 2025), and adaptive stepsizes, for which Zhang et al. (2026) proved a $1/\gamma^2$ step complexity that is minimax optimal among batch first-order methods (see also Kornowski & Shamir, 2025); for non-separable data, GD with a large stepsize need not converge to the minimizer and can instead settle into a cycle (Meng et al., 2024; 2025); Tyurin (2025) related GD with very large stepsizes to the perceptron algorithm. Most relevant to us, Crawshaw & Liu (2026) tightened the two-phase analysis of Wu et al. (2024) in two dimensions and proved a nearly matching lower bound. Our work is directly motivated by Wu et al. (2024) and Crawshaw & Liu (2026); we improve the arbitrary-dimensional analysis of the former and complement the two-dimensional analysis of the latter.

Edge of stability. Neural networks are commonly trained with stepsizes at which the loss oscillates yet converges in the long run (Wu et al., 2018; Xing et al., 2018; Lewkowycz et al., 2020), a phenomenon Cohen et al. (2021) named the edge of stability; Cohen et al. (2022) observed analogous behavior for adaptive methods. A large body of theory explains the mechanism of EoS in concrete models (Zhu et al., 2023; Ahn et al., 2023; Kreisler et al., 2023; Chen et al., 2024; Lyu et al., 2022; Even et al., 2023; Andriushchenko et al., 2023; Lu et al., 2024; Wang et al., 2022; Chen & Bruna, 2023) and characterizes the behavior of GD under instability and how the sharpness evolves during training (Kong & Tao, 2020; Ahn et al., 2022; Arora et al., 2022; Ma et al., 2022; Li et al., 2022; Damian et al., 2023; Cohen et al., 2025). Instead of explaining the mechanism of EoS in deep networks, we study the optimization benefits of GD operating in the unstable regime, in a natural learning problem and under standard assumptions.

Aggressive stepsizes in convex optimization. For smooth convex optimization, a recent line of work shows that stepsize schedules with occasional long steps, which may temporarily increase the loss, provably improve the convergence of GD. Long-step schedules first gave constant-factor gains (Altschuler, 2018; Grimmer, 2024) or gains for objectives with multiple scales (Kelner et al., 2022), and the silver stepsize schedule and its successors accelerate GD beyond the $O(1/t)$ rate of the descent regime (Altschuler & Parrilo, 2025; Grimmer et al., 2025; Zhang et al., 2025). These works design stepsize sequences for function classes whose minimizers are attained, and the acceleration is driven by the schedule. In contrast, we study a large constant stepsize on a problem whose infimum is not attained, where the acceleration arises from the dynamics of GD itself.

2 Main Results

We assume that the data are bounded and linearly separable (Novikoff, 1962; Wu et al., 2024).

Assumption 2.1 (Bounded and separable data)**.** Assume the training data $(\mathbf{x}_i, y_i)_{i=1}^n$ satisfies

A. for every $i = 1, \dots, n$, $\|\mathbf{x}_i\| \le 1$ and $y_i \in \{\pm 1\}$;
B. there is a margin $\gamma \in (0, 1]$ and a unit vector $\mathbf{w}_*$ such that $\langle y_i \mathbf{x}_i, \mathbf{w}_* \rangle \ge \gamma$ for every $i = 1, \dots, n$.

We write $r := \dim \operatorname{span}\{y_1\mathbf{x}_1, \dots, y_n\mathbf{x}_n\} \le \min\{n, d\}$ for the rank of the data.

Under Assumption 2.1, the logistic loss (1) is convex and β-smooth with $\beta = 1/4$. For stepsizes $\eta < 2/\beta = 8$, the descent lemma guarantees monotone decrease of the loss, and classical theory shows that GD reaches loss ϵ within $\widetilde{O}(1/\epsilon)$ steps for fixed n and γ (Nesterov, 2018; Ji & Telgarsky, 2019). With a larger stepsize the loss may initially oscillate, yet for suitable choices GD converges even faster. Wu et al. (2024) showed that GD with a large stepsize $\eta = \Theta(1/\sqrt{\epsilon})$ reaches loss ϵ within $\widetilde{O}(1/\sqrt{\epsilon})$ steps. In this work, we show that GD with a large stepsize $\eta = 1/\epsilon$ reaches loss ϵ within $O(\ln^p(1/\epsilon))$ steps, where p depends only on the margin and the rank of the data.

Following Wu et al. (2024), we say that GD is in the *stable phase* at step t if the loss decreases monotonically from step t onwards, and in the *unstable phase* otherwise. If $L(\mathbf{w}_s) \le 1/\eta$ at some step s, then GD is in the stable phase at step s (Wu et al., 2024, Theorem 1). We therefore define the *phase transition time*

$$\tau := \min\{t \ge 0 : \ L(\mathbf{w}_t) \le 1/\eta\}, \tag{2}$$

so that the loss may oscillate before step τ and decreases monotonically afterwards.

To state the theorem, define the depth ν, the exponent p, the prefactor C, and the stepsize threshold η_0 (whose roles in the proof are explained in Section 3) by

$$\nu := \max\left\{0,\ \min\left\{r,\ \left\lfloor \gamma^{-2} + \frac{\gamma^{-2}-1}{h-1} \right\rfloor\right\} - 2\right\}, \quad \text{where } h := \lceil \ln(16\gamma^{-2}) \rceil, \tag{3}$$

$$p := \begin{cases} \nu + \left\lceil \dfrac{\nu \ln(18h\gamma^{-2}/\nu)}{\ln(\max\{2,\nu\})} \right\rceil & \text{if } \nu \geq 1, \\ 0 & \text{if } \nu = 0, \end{cases} \tag{4}$$

$$C := 80\gamma^{-4}h\big(n + h\ln(4\gamma^{-2})\big), \tag{5}$$

$$\eta_0 := \big\lceil 128\, K(1+\ln K) \big\rceil, \quad \text{where } K := \exp(4) + h\gamma^{-2} + p + \ln(2\gamma^{-2}C). \tag{6}$$

We now present an improved analysis for GD with a large constant stepsize for logistic regression on linearly separable data. Our main theorem improves the dependence of the phase transition time on η from linear (Wu et al., 2024) to polylogarithmic in arbitrary dimension.

Theorem 2.2. *Consider (GD) with a constant stepsize $\eta > 0$ and initialization $\mathbf{w}_0 = \mathbf{0}$ for logistic regression (1) under Assumption 2.1, and let p, C, and η_0 be as in (3)–(6).*

(i) ***Phase transition.*** *If $0 < \gamma < 1/\sqrt{2}$ and $\eta \geq \eta_0$, then*

$$\tau \leq C\,(\ln \eta)^p; \tag{7}$$

if $\gamma \geq 1/\sqrt{2}$ and $\eta \geq 64n^2$, then $\tau \leq 1$.

(ii) ***The stable phase*** *(Wu et al., 2024, Theorem 1). The loss decreases monotonically from step τ onwards, that is, $L(\mathbf{w}_{t+1}) \leq L(\mathbf{w}_t)$ for every $t \geq \tau$.*

The proof of Theorem 2.2 is deferred to Section 3 and Appendices A and B. Theorem 2.2 describes both phases of GD with a large constant stepsize. Part (i) bounds the length of the unstable phase. Once the stepsize exceeds the threshold η_0, GD reaches loss $1/\eta$ within $C(\ln \eta)^p$ steps, and when $\gamma \geq 1/\sqrt{2}$ a single step suffices. This improves the linear dependence on η of the previous phase transition time bound (Wu et al., 2024) to a polylogarithmic one in arbitrary dimension. Part (ii) describes the stable phase and restates the analysis of Wu et al. (2024). From step τ onwards the loss decreases monotonically, and Appendix A.4 recalls the rate at which it continues to decrease.

Parameter dependence. For $0 < \gamma < 1/\sqrt{2}$, the quantities in (3)–(6) satisfy

$$C = O\big(\gamma^{-4}\ln(\gamma^{-2})\,[\,n + \ln^2(\gamma^{-2})\,]\big), \quad \eta_0 = \widetilde{O}\big(\gamma^{-2} + \ln n\big), \quad p = O\big(\min\{r, \gamma^{-2}\}\ln(\gamma^{-2})\big).$$

Thus the prefactor is polynomial in n and $1/\gamma$, while the threshold η_0 is polynomial in $1/\gamma$ and only polylogarithmic in n. When $r \leq 2$, we have $\nu = p = 0$, and the bound (7) becomes $\tau \leq C$, which does not depend on η. When the rank satisfies the condition (40) in Appendix A.2, the exponent satisfies $p = (1 + o(1))\gamma^{-2}$ as $\gamma \to 0$. All hidden constants are absolute, and these bounds are proved in Proposition A.2.

Comparisons with prior results. Wu et al. (2024) proved the phase transition time bound

$$\tau \leq \frac{60}{\gamma^2}\max\left\{\eta,\ n,\ e,\ \frac{\eta+n}{\eta}\ln\frac{\eta+n}{\eta}\right\} \tag{8}$$

for every $\eta > 0$ in arbitrary dimension. The bound (8) grows linearly with the stepsize, whereas the bound (7) grows only polylogarithmically, and does not grow at all when $r \leq 2$; for large stepsizes, (7) is therefore smaller by a factor almost linear in η. The price is the stepsize threshold, as (7) holds only for $\eta \geq \eta_0$ while (8) holds for every $\eta > 0$. This threshold is mild, however, and the two-dimensional bound of Crawshaw & Liu (2026) requires $\eta \geq n$.

The threshold $1/\sqrt{2}$ in Theorem 2.2(i) marks a change in the data geometry. For $\gamma \geq 1/\sqrt{2}$ the signed examples have pairwise nonnegative inner products, and every GD step adds a nonnegative combination of them, so no margin ever decreases and the loss decreases monotonically from initialization for every $\eta > 0$. There is no unstable phase in this regime, and the second case of Theorem 2.2(i) adds that a single step with $\eta \geq 64n^2$ already brings the loss below $1/\eta$. Crawshaw & Liu (2026) observe the same effect and exclude this regime from their lower bound.

Table 1: Step complexity of constant stepsize GD for logistic regression on linearly separable data. Each complexity holds for all sufficiently small ϵ.

	Complexity	Stepsize	Setting
Wu et al. (2024, Corollary 2)	$\widetilde{O}\big(1/(\gamma^2\sqrt{\epsilon})\big)$	$\eta = \Theta(1/\sqrt{\epsilon})$	any d
Crawshaw & Liu (2026, Theorem 2.1)	$O\big(n/\gamma + \ln(1/\gamma)/\gamma^2\big)$	$\eta \geq \Omega\big(1/(\epsilon(\gamma n + 1))\big)$	$d = 2$
Corollary 2.3	$O\big(n\gamma^{-4}\ln^3(e/\gamma)\ln^p(1/\epsilon)\big)$	$\eta = 1/\epsilon$	any d

The phase transition bound in part (i) and the monotonicity in part (ii) of Theorem 2.2 combine to give the step complexity of GD with a large constant stepsize, in terms of a target loss ϵ and of a budget of T steps. The proof is deferred to Appendix A.5.

Corollary 2.3 (Complexity of GD with a large constant stepsize)**.** *Under the setting of Theorem 2.2, set $C := 1$ and $p := 0$ if $\gamma \geq 1/\sqrt{2}$, and define the thresholds*

$$\epsilon_0 := \begin{cases} 1/\eta_0 & \text{if } \gamma < 1/\sqrt{2}, \\ 1/(64n^2) & \text{if } \gamma \geq 1/\sqrt{2}, \end{cases} \qquad T_0 := \big\lceil C(\ln \eta_0)^p \big\rceil.$$

(i) ***Target loss.*** *For every $0 < \epsilon \leq \epsilon_0$, GD with the constant stepsize $\eta = 1/\epsilon$ satisfies $L(\mathbf{w}_t) \leq \epsilon$ for some $t \leq C\ln^p(1/\epsilon)$, and the loss decreases monotonically from step t onwards.*

(ii) ***Budget.*** *If $p \geq 1$, then for every budget $T \geq T_0$, GD with the constant stepsize $\eta = \exp\big((T/C)^{1/p}\big)$ satisfies $L(\mathbf{w}_T) \leq \exp\big(-(T/C)^{1/p}\big)$. If $p = 0$, then for every $0 < \epsilon \leq \epsilon_0$, GD with the constant stepsize $\eta = 1/\epsilon$ satisfies $L(\mathbf{w}_T) \leq \epsilon$ for every $T \geq C$.*

Corollary 2.3 shows that, for fixed n and γ, the step complexity improves from $\widetilde{O}(1/\sqrt{\epsilon})$ (Wu et al., 2024) to $O(\ln^p(1/\epsilon))$, that is, from polynomial to polylogarithmic in $1/\epsilon$. See Table 1 for a comparison with prior results. When $p = 0$, in particular when $r \leq 2$ or $\gamma \geq 1/\sqrt{2}$, the complexity does not depend on ϵ, as Crawshaw & Liu (2026) showed for $d = 2$. Given a budget of T steps, the loss improves from $\widetilde{O}(1/T^2)$ (Wu et al., 2024) to $\exp(-(T/C)^{1/p})$ when $p \geq 1$.

Limitations. Our bounds are not tight in several respects. First, the prefactor C is not optimized; in two dimensions it is weaker than the nearly tight bound of Crawshaw & Liu (2026). Second, our results apply only above the stepsize threshold η_0, and we claim nothing for moderate stepsizes between the descent regime and η_0. Third, we prove no lower bounds, and this paper does not determine whether the exponent p or the dependence on η in (7) is optimal.

3 Analysis

In this section we prove the phase transition bound of Theorem 2.2(i) for $0 < \gamma < 1/\sqrt{2}$; the large-margin case is elementary and is proved in Appendix A.3, and part (ii) is recalled in Appendix A.4. Throughout, $\mathbf{z}_i := y_i\mathbf{x}_i$ are the signed examples, $\eta \geq \eta_0$, $\Lambda := \ln\eta$, and h, ν, p, C, and η_0 are as in (3)–(6). Rather than averaging a potential along the trajectory or tracking the oscillations of the iterates, we count steps. Steps at which some signed margin is negative are few. The remaining steps are organized into nested intervals (see **Figure 1** in Appendix B.2 for an illustration), each carrying reference directions that constrain every update in it. The margin and the rank bound the depth of this hierarchy, and the cumulative update coefficients of sibling intervals are separated on a logarithmic scale, which lets us count the intervals at each depth with a factorial saving. Adding up the lengths of the intervals proves the bound.

3.1 Update geometry before the phase transition

For each step, define the logistic weights, the update coefficient, and the update direction

$$q_{i,t} := \frac{1}{1 + \exp(\langle \mathbf{z}_i, \mathbf{w}_t \rangle)}, \qquad Q_t := \sum_{i=1}^{n} q_{i,t}, \qquad \mu_t := \frac{\eta Q_t}{n}, \qquad \mathbf{v}_t := \frac{1}{Q_t}\sum_{i=1}^{n} q_{i,t}\mathbf{z}_i. \tag{9}$$

In this notation, the update (GD) becomes

$$\mathbf{w}_{t+1} - \mathbf{w}_t = \mu_t \mathbf{v}_t, \qquad \mathbf{v}_t \in \operatorname{conv}\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}, \qquad \|\mathbf{v}_t\| \le 1, \qquad \langle \mathbf{v}_t, \mathbf{w}_* \rangle \ge \gamma. \tag{10}$$

Thus μ_t is the coefficient of the step and $\mathbf{v}_t$ its direction, an average of the signed examples. The trajectory estimates below hold on the finite horizon

$$\widehat{T} := \left\lfloor \frac{\gamma^2 \exp(\eta/4)}{\eta} \right\rfloor, \qquad U := \min\{\tau, \widehat{T}\}, \tag{11}$$

and we count the steps $0, 1, \ldots, U-1$; at the end of the proof we show that $U < \widehat{T}$, which forces $U = \tau$. We split these steps according to the signs of the margins, into $\mathcal{T}_- := \{t < U : \langle \mathbf{z}_i, \mathbf{w}_t \rangle < 0 \text{ for some } i\}$ and $\mathcal{T}_+ := \{0, 1, \ldots, U-1\} \setminus \mathcal{T}_-$. The next lemma bounds the total update coefficient before U, which shows that $\mathcal{T}_-$ is small, and states two facts about the steps in $\mathcal{T}_+$: each of them carries a coefficient of at least $1/2$, and the iterate is not yet far along its own update direction.

Lemma 3.1 (Trajectory estimates)**.** *We have*

$$\sum_{t<U} \mu_t \le 2\eta\gamma^{-2}, \qquad |\mathcal{T}_-| < 4n\gamma^{-2}, \tag{12}$$

and at every $t \in \mathcal{T}_+$,

$$\mu_t \ge \frac{\eta}{2} L(\mathbf{w}_t) > \frac{1}{2}, \qquad 0 \le \langle \mathbf{w}_t, \mathbf{v}_t \rangle < \Lambda. \tag{13}$$

Proof. The first bound in (12) follows from the split optimization technique introduced by Wu et al. (2024) on the horizon (11), for which $\eta \ge 4$ suffices (Appendix B.1). At $t \in \mathcal{T}_-$ some weight exceeds $1/2$, so $\mu_t > \eta/(2n)$, and the second bound follows from the first. Now let $t \in \mathcal{T}_+$, write $\ell(u) := \ln(1 + e^{-u})$, and let $m_i := \langle \mathbf{z}_i, \mathbf{w}_t \rangle \ge 0$. For $u \ge 0$ and $a := e^{-u} \in (0, 1]$,

$$\frac{1}{1+e^u} = \frac{a}{1+a} \ge \frac{a}{2} \ge \frac{1}{2}\ln(1+a) = \frac{1}{2}\ell(u),$$

so $Q_t/n \ge L(\mathbf{w}_t)/2$, which gives $\mu_t \ge \eta L(\mathbf{w}_t)/2$; this exceeds $1/2$ because $t < \tau$. For the radial bound, convexity of ℓ and $\ell'(m_i) = -q_{i,t}$ give $\ell(\Lambda) \ge \ell(m_i) + q_{i,t}(m_i - \Lambda)$, and summing over i,

$$Q_t\big(\langle \mathbf{w}_t, \mathbf{v}_t \rangle - \Lambda\big) = \sum_{i=1}^{n} q_{i,t}(m_i - \Lambda) \le n\ell(\Lambda) - nL(\mathbf{w}_t) < 0,$$

because $\ell(\Lambda) = \ln(1 + 1/\eta) < 1/\eta < L(\mathbf{w}_t)$. Nonnegativity holds because every $m_i \ge 0$. □

3.2 A BOUNDED-DEPTH HIERARCHY OF INTERVALS

This section builds the hierarchy of intervals and bounds its depth. See Figure 1 in Appendix B.2 for an illustration. Fix $\theta := \gamma^2/h$. Every interval below is a set of consecutive steps in $\mathcal{T}_+$. We say that vectors $\mathbf{a}_1, \ldots, \mathbf{a}_j$ are *reference directions* for an interval I if

$$\mathbf{a}_i \in \operatorname{conv}\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}, \qquad \langle \mathbf{a}_i, \mathbf{a}_k \rangle \le \theta \ \ (i \ne k), \qquad \langle \mathbf{a}_i, \mathbf{v}_t \rangle \le \theta \ \ (t \in I). \tag{14}$$

A reference direction is an earlier average update direction from which every update in I has turned away. These constraints are local to one branch of the construction.

Both constructions in the next lemma freeze a running average of update directions and start a new interval whenever the average has moved by more than $\theta/2$. Roots use the average from the origin, $\boldsymbol{\xi}_t := \mathbf{w}_t/S_t$ $(t \ge 1)$ with $S_t := \sum_{k<t} \mu_k$; refinement uses the average since the start of the parent.

Lemma 3.2 (Root partition and refinement)**.** *The following statements hold with $R := 1 + \lceil 4\Lambda/\theta \rceil$.*

(i) *The steps in $\mathcal{T}_+ \setminus \{0\}$ can be partitioned into at most $|\mathcal{T}_-| + 1 + \lceil 4\ln(4\gamma^{-2})/\theta \rceil$ consecutive intervals, each carrying one reference direction.*

(ii) *Every interval $I = \{a, \ldots, b\} \subseteq \mathcal{T}_+$ carrying j reference directions admits a partition*

$$I = I_0 \,\dot\cup\, I_1 \,\dot\cup\, \cdots \,\dot\cup\, I_m, \qquad |I_0| \le R, \qquad m \le R, \tag{15}$$

into a prefix I_0 and consecutive children $I_1, \ldots, I_m$. Each child inherits the j reference directions of I and gains one new reference direction satisfying (14). If $b_i := \min I_i$ and $\sigma_i := \sum_{t=a}^{b_i - 1} \mu_t$, then

$$\sigma_i \ge \frac{2\Lambda}{\theta} \quad (1 \le i \le m), \qquad \ln \frac{\sigma_{i+1}}{\sigma_i} > \frac{\theta}{4} \quad (1 \le i < m). \tag{16}$$

The case $m = 0$ is allowed, with $I_0 = I$.

Proof sketch. Root partition. The vector $\boldsymbol{\xi}_{t+1}$ is a convex combination of $\boldsymbol{\xi}_t$ and $\mathbf{v}_t$ with weights S_t/S_{t+1} and μ_t/S_{t+1}, and both vectors lie in the convex hull of the signed examples, so

$$\|\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t\| \le \frac{2\mu_t}{S_{t+1}} \le 2\ln\frac{S_{t+1}}{S_t}, \qquad \sum_{t=1}^{U-1}\|\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t\| \le 2\ln\frac{S_U}{S_1} \le 2\ln(4\gamma^{-2}), \tag{17}$$

by $S_1 = \mu_0 = \eta/2$ and (12). The set $\mathcal{T}_+ \setminus \{0\}$ consists of at most $|\mathcal{T}_-| + 1$ maximal runs of consecutive steps. Cut each run whenever $\boldsymbol{\xi}_t$ has moved by more than $\theta/2$ from its value at the start of the current piece; by (17) there are at most $4\ln(4\gamma^{-2})/\theta$ cuts in total. A piece starting at step c receives $\mathbf{a}_1 := \boldsymbol{\xi}_c$, which lies in the convex hull, and for every t in the piece, by (13), $S_t \ge S_1 = \eta/2$, and the stepsize condition $4\Lambda \le \eta\theta$,

$$\langle \mathbf{a}_1, \mathbf{v}_t\rangle \le \frac{\langle \mathbf{w}_t, \mathbf{v}_t\rangle}{S_t} + \|\boldsymbol{\xi}_c - \boldsymbol{\xi}_t\| < \frac{2\Lambda}{\eta} + \frac{\theta}{2} \le \theta.$$

Refinement. Let $S_t^{(a)} := \sum_{k=a}^{t-1}\mu_k$ and, for $t > a$, let $\boldsymbol{\zeta}_t := (\mathbf{w}_t - \mathbf{w}_a)/S_t^{(a)}$ be the average update direction since step a. Every signed margin at a is nonnegative and $\mathbf{v}_t$ is a convex combination of signed examples, so $\langle \mathbf{w}_a, \mathbf{v}_t\rangle \ge 0$, and (13) gives

$$\langle \boldsymbol{\zeta}_t, \mathbf{v}_t\rangle \le \frac{\langle \mathbf{w}_t, \mathbf{v}_t\rangle}{S_t^{(a)}} < \frac{\Lambda}{S_t^{(a)}} \le \frac{\theta}{2} \qquad \text{once } S_t^{(a)} \ge \frac{2\Lambda}{\theta}. \tag{18}$$

The prefix I_0 consists of the steps of I at which $S_t^{(a)} < 2\Lambda/\theta$; since $\mu_t > 1/2$ on $\mathcal{T}_+$, it has at most $\lceil 4\Lambda/\theta\rceil$ steps, and $\sigma_i \ge 2\Lambda/\theta$ by construction. On the rest of I we cut a new child whenever $\boldsymbol{\zeta}_t$ has moved by more than $\theta/2$ from its value at the start of the current child. As in (17), $\|\boldsymbol{\zeta}_{t+1} - \boldsymbol{\zeta}_t\| \le 2\ln(S_{t+1}^{(a)}/S_t^{(a)})$, and the total movement after the prefix is at most $2\ln(\eta\gamma^{-2}\theta/\Lambda) < 2\Lambda$ by (12), so $m < 1 + 4\Lambda/\theta$. The child I_i receives $\mathbf{a}_{j+1} := \boldsymbol{\zeta}_{b_i}$, a convex combination of the $\mathbf{v}_k$ with $a \le k < b_i$. The inner-product bounds (14) for the old reference directions pass through this average exactly, and the bound for the new direction follows from (18) and the cut rule:

$$\begin{aligned} \langle \mathbf{a}_s, \mathbf{a}_{j+1}\rangle &= \frac{1}{\sigma_i}\sum_{k=a}^{b_i-1}\mu_k\langle \mathbf{a}_s, \mathbf{v}_k\rangle \le \theta \qquad (1 \le s \le j), \\ \langle \mathbf{a}_{j+1}, \mathbf{v}_t\rangle &\le \langle \boldsymbol{\zeta}_t, \mathbf{v}_t\rangle + \|\boldsymbol{\zeta}_{b_i} - \boldsymbol{\zeta}_t\| \le \frac{\theta}{2} + \frac{\theta}{2} = \theta \qquad (t \in I_i). \end{aligned} \tag{19}$$

Finally, a completed child moves the average by more than $\theta/2$, and the movement over the child is at most $2\ln(\sigma_{i+1}/\sigma_i)$, which gives (16). Appendix B.2 gives the cut points and the counts. □

Lemma 3.2 shows where the powers of $\Lambda = \ln\eta$ come from. The roots use the average from the origin, whose total movement (17) does not depend on η, so the first reference direction costs no factor of Λ. Each refinement costs one factor of Λ through R, because the average since the start of the parent must first accumulate coefficient $2\Lambda/\theta$. Since the reference directions are convex averages, every inherited inner-product bound is preserved exactly, so the same θ serves at every depth. The depth of the hierarchy is bounded by the following capacity argument.

Lemma 3.3. *Let* $\mathbf{b}_1, \dots, \mathbf{b}_m \in \operatorname{conv}\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ *satisfy* $\langle \mathbf{b}_i, \mathbf{b}_k\rangle \le \theta$ *for all* $i \ne k$. *Then*

$$m \le \min\left\{r,\ \left\lfloor \frac{1-\theta}{\gamma^2 - \theta}\right\rfloor\right\} \le \nu + 2. \tag{20}$$

Proof. Each $\mathbf{b}_i$ satisfies $\|\mathbf{b}_i\| \le 1$ and $\langle \mathbf{b}_i, \mathbf{w}_*\rangle \ge \gamma$. Suppose $\sum_i c_i\mathbf{b}_i = \mathbf{0}$ with the c_i not all zero. Coefficients of both signs occur, since a nonzero nonnegative combination has positive inner product with $\mathbf{w}_*$. Let $\mathbf{u} := \sum_{c_i>0} c_i\mathbf{b}_i = \sum_{c_k<0}(-c_k)\mathbf{b}_k$, $a := \sum_{c_i>0} c_i$, and $b := \sum_{c_k<0}(-c_k)$. The separator gives $\|\mathbf{u}\|^2 \ge \gamma^2\max\{a^2, b^2\} \ge \gamma^2 ab$, while pairing the two representations gives $\|\mathbf{u}\|^2 \le \theta ab < \gamma^2 ab$, a contradiction. Hence the $\mathbf{b}_i$ are linearly independent vectors in the span of the data, and $m \le r$. Next,

$$m^2\gamma^2 \le \langle \textstyle\sum_i \mathbf{b}_i, \mathbf{w}_*\rangle^2 \le \left\|\textstyle\sum_i \mathbf{b}_i\right\|^2 \le m + m(m-1)\theta,$$

so $m(\gamma^2 - \theta) \le 1 - \theta$. With $\theta = \gamma^2/h$ the floor in (20) equals $\lfloor \gamma^{-2} + (\gamma^{-2} - 1)/(h-1)\rfloor$, and (3) gives the last inequality. □

By Lemma 3.3, at most $\nu+2$ vectors of the convex hull can have pairwise inner products at most θ. An interval carrying $\nu+1$ reference directions therefore leaves room for only one more such vector, so any two of its update directions have inner product above θ. The next lemma shows that such an interval is short, because $1/(\eta L(\mathbf{w}_t))$ grows by a fixed amount at each of its steps.

Lemma 3.4. *If an interval $I \subseteq \mathcal{T}_+$ carries $\nu+1$ reference directions, then $|I| \le \lceil 4/\theta \rceil$.*

Proof. If two distinct steps $s,t \in I$ had $\langle \mathbf{v}_s, \mathbf{v}_t \rangle \le \theta$, then $\mathbf{v}_s$, $\mathbf{v}_t$, and the $\nu+1$ reference directions would be $\nu+3$ vectors in the convex hull with pairwise inner products at most θ, contradicting Lemma 3.3. Hence $\langle \mathbf{v}_t, \mathbf{v}_{t+1} \rangle > \theta$ for consecutive steps in I. Convexity of L at $\mathbf{w}_{t+1}$, the identity $\nabla L(\mathbf{w}_{t+1}) = -(\mu_{t+1}/\eta)\mathbf{v}_{t+1}$, and (13) give

$$L(\mathbf{w}_t) - L(\mathbf{w}_{t+1}) \ge \langle \nabla L(\mathbf{w}_{t+1}), \mathbf{w}_t - \mathbf{w}_{t+1} \rangle = \frac{\mu_t \mu_{t+1}}{\eta} \langle \mathbf{v}_t, \mathbf{v}_{t+1} \rangle \ge \frac{\theta \eta}{4} L(\mathbf{w}_t) L(\mathbf{w}_{t+1}),$$

that is, $1/(\eta L(\mathbf{w}_{t+1})) - 1/(\eta L(\mathbf{w}_t)) \ge \theta/4$. Before τ every reciprocal $1/(\eta L(\mathbf{w}_t))$ is below one, so summing over the $|I|-1$ consecutive pairs gives $(|I|-1)\theta/4 < 1$, that is, $|I| \le \lceil 4/\theta \rceil$. □

Each root carries one reference direction and each refinement adds one, so an interval reached after ν refinements carries $\nu+1$ reference directions. We call such an interval *terminal*. It is not refined further, and Lemma 3.4 bounds its length. This is the role of the 2 subtracted in (3), since $\nu+1$ reference directions and two update directions are one vector more than Lemma 3.3 allows. When $\nu = 0$, in particular when $r \le 2$, every root is already terminal and no refinement is needed, which is why $p = 0$ there.

3.3 Counting intervals

This section counts the intervals of the hierarchy at each depth and bounds the total length of one tree. The construction produces one rooted tree per root interval. A node is an interval together with its reference directions, the children of a nonterminal interval are the $I_1, \dots, I_m$ of Lemma 3.2(ii), and terminal intervals have no children. Let $D_j(I)$ denote the number of descendants of I at depth j below I, with $D_0(I) = 1$. Each nonterminal interval has at most $R \le 6\Lambda/\theta$ children, as $\Lambda > 1 > \theta$, so $D_j(I) \le (6\Lambda/\theta)^j$ for every j. This count ignores that the children of one interval share its coefficient budget. By Lemma 3.2(ii), the children start at coefficients $\sigma_1 < \sigma_2 < \cdots$ measured from the start of the parent, and (16) keeps their logarithms $\theta/4$ apart. Counting the siblings jointly along this axis yields a factorial saving.

To this end, for a nonempty interval I define its coefficient and its logarithmic size

$$\mu(I) := \sum_{t \in I} \mu_t, \qquad \chi(I) := \max\left\{0,\ \ln \frac{\theta \mu(I)}{2\Lambda}\right\}. \tag{21}$$

By (12), $0 \le \chi(I) \le \max\{0, \ln(\eta/(h\Lambda))\} \le \Lambda$. The next lemma uses the truncated exponential $E_k(s) := \sum_{i=0}^{k} s^i/i!$ and, for $\nu \ge 1$, the parameter $\alpha := 2\nu/\Lambda$.

Lemma 3.5. *Let $\nu \ge 1$. For every interval I of the construction and every $0 \le j \le \nu$,*

$$D_j(I) \le 2\,\frac{(6\Lambda/\theta)^j}{j!}. \tag{22}$$

Proof. If I has no children, (24) below is trivial. Otherwise let $I_1, \dots, I_m$ be the children of I and $\sigma_1 < \cdots < \sigma_m$ their starting coefficients, so that $\mu(I_i) = \sigma_{i+1} - \sigma_i \le \sigma_{i+1}$ for $i < m$ and $\mu(I_m) \le \mu(I)$. The points $u_i := \ln(\theta\sigma_i/(2\Lambda))$, $2 \le i \le m$, are nonnegative because $\sigma_1 \ge 2\Lambda/\theta$, are at most $\chi(I)$, and are more than $\theta/4$ apart by (16). Hence, for every nondecreasing $f \ge 0$, comparing the disjoint intervals $[u_i, u_i + \theta/4]$ with an integral gives

$$\sum_{i=1}^{m} f(\chi(I_i)) \le f(\chi(I)) + \frac{4}{\theta} \int_0^{\chi(I)+\theta/4} f(u)\,du, \tag{23}$$

where the first term accounts for the last child. Take $f(u) = E_k(\alpha u)$. Since $\int_0^v E_k(\alpha u)\,du = (E_{k+1}(\alpha v) - 1)/\alpha$ and $E_{k+1}(u+v) \le e^v E_{k+1}(u)$ for $u, v \ge 0$, the right side of (23) is at most

$(1 + 4e^{\alpha\theta/4}/(\alpha\theta))E_{k+1}(\alpha\chi(I))$, and $\alpha\theta < 4/(h\Lambda) < 1/5$ gives

$$\sum_{i=1}^{m} E_k\big(\alpha\chi(I_i)\big) \le \frac{6}{\alpha\theta} E_{k+1}\big(\alpha\chi(I)\big) \qquad (k \ge 0). \tag{24}$$

By induction on j, simultaneously for all intervals, (24) gives $D_j(I) \le (6/(\alpha\theta))^j E_j(\alpha\chi(I))$: the case $j = 0$ is trivial, and $D_{j+1}(I) = \sum_i D_j(I_i)$. Finally, $\chi(I) \le \Lambda$ and $\alpha\Lambda = 2\nu \ge 2j$, so the terms of $E_j(\alpha\Lambda)$, read from the top, decrease by factors of at least two, whence $E_j(\alpha\chi(I)) \le 2(\alpha\Lambda)^j/j!$. Substituting proves (22). □

It remains to bound the length of a root interval J. Every step of J lies either in the prefix of a nonterminal interval or in a terminal interval. Prefixes have at most R steps and occur at depths $0, \dots, \nu - 1$, and terminal intervals have at most $\lceil 4/\theta \rceil \le 6/\theta$ steps and occur at depth ν. Let $\nu \ge 1$ and $z := 6\Lambda/\theta$. Since $z/\nu > 18$, the sum $\sum_{j<\nu} z^j/j!$ is dominated by its last term, so Lemma 3.5 gives $\sum_{j<\nu} D_j(J) < 3z^{\nu-1}/(\nu-1)!$ and $D_\nu(J) \le 2z^\nu/\nu!$. With $R \le z$ this yields

$$|J| \le R\sum_{j=0}^{\nu-1} D_j(J) + \left\lceil \frac{4}{\theta} \right\rceil D_\nu(J) \le \left(3\nu + \frac{12}{\theta}\right) \frac{(6\Lambda/\theta)^\nu}{\nu!}, \tag{25}$$

and the same bound holds when $\nu = 0$, where J itself is terminal. Each level of prefixes costs one factor of Λ and the terminal intervals cost none, so the degree in Λ is ν and not $\nu + 1$.

3.4 Completing the proof

The stepsize threshold η_0 in (6) is chosen so that every $\eta \ge \eta_0$ satisfies

$$\eta \ge 4, \qquad \Lambda > 10, \qquad 4\Lambda \le \eta\theta, \qquad C\Lambda^p < \widehat{T} \tag{26}$$

(Appendix B.4). The first three conditions were used above, and the last one is used at the end of the proof. We now add up the steps before U. Besides step zero and the fewer than $4n\gamma^{-2}$ steps of $\mathcal{T}_-$, there are the root intervals of Lemma 3.2(i), whose number is below $5\gamma^{-2}(n + h\ln(4\gamma^{-2}))$ by (12). Applying (25) to each of them and using $3\nu + 12h\gamma^{-2} \le 15h\gamma^{-2}$ gives

$$U \le C\,\frac{(6h\gamma^{-2}\Lambda)^\nu}{\nu!}, \qquad C = 80\gamma^{-4}h\big(n + h\ln(4\gamma^{-2})\big). \tag{27}$$

The factorial in (27) is the saving of Section 3.3. It remains to absorb the fixed coefficient $(6h\gamma^{-2})^\nu/\nu!$ into the exponent, which is how p in (4) arises. If $\nu = 0$, then $p = 0$ and (27) reads $U \le C$. If $\nu \ge 1$, let $b := \max\{2, \nu\}$. From $\nu! \ge (\nu/e)^\nu$, $e < 3$, and the definition of p,

$$\frac{(6h\gamma^{-2})^\nu}{\nu!} \le \left(\frac{18h\gamma^{-2}}{\nu}\right)^\nu \le b^{p-\nu}. \tag{28}$$

If $\Lambda \ge b$, then (27) and (28) give $U \le C\Lambda^\nu b^{p-\nu} \le C\Lambda^p$. If instead $10 < \Lambda < b$, then $b = \nu$ and $\Lambda < \nu$; here (12) and $\mu_t > 1/2$ on $\mathcal{T}_+$ give the crude count $U = |\mathcal{T}_-| + |\mathcal{T}_+| < C + 4\gamma^{-2}e^\Lambda \le C(1 + e^\Lambda)$, and $e^\Lambda \le \Lambda^\Lambda \le \Lambda^\nu$ together with $p \ge \nu + 1$ gives $U \le 2C\Lambda^\nu < C\Lambda^p$. The threshold η_0 therefore never needs $\Lambda \ge \nu$, which would be an exponential stepsize requirement. In every case $U \le C\Lambda^p$, and by (26) the right side is smaller than $\widehat{T}$. Since $U = \min\{\tau, \widehat{T}\}$, this forces $U = \tau$, and $\tau \le C(\ln\eta)^p$ is (7). This completes the proof of Theorem 2.2(i) for $0 < \gamma < 1/\sqrt{2}$.

Remark. Since $U = \tau$, (27) itself gives the bound $\tau \le C(6h\gamma^{-2}\ln\eta)^\nu/\nu!$ on the phase transition time, which is at least as sharp as (7) whenever $\ln\eta \ge \max\{2, \nu\}$ (Proposition B.13). For $\nu \ge 1$ it has the smaller degree ν, but its coefficient $(6h\gamma^{-2})^\nu/\nu!$ can be superpolynomial in $1/\gamma$. The exponent p absorbs this coefficient and keeps the prefactor C polynomial in n and $1/\gamma$.

4 Conclusion

We consider gradient descent (GD) with a constant stepsize for logistic regression on linearly separable data. We show that GD with a large stepsize $\eta = 1/\epsilon$ achieves an $O(\ln^p(1/\epsilon))$ step complexity, where p depends only on the margin and the rank of the data. This improves over the prior rate in arbitrary dimension for fixed problem dependent parameters. We leave it to future work to tighten the constants in our bounds and to establish new lower bounds.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting and polishing the writing, developing parts of the proof strategy, and reviewing the proofs. The finite-horizon argument and the tightened interval-counting argument were developed through iterative collaboration with GPT-5.6 Sol. GPT-5.6 Sol was also used to simplify the proof and improve its organization and presentation. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 247–257, 2022.

Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via the "edge of stability". In *Advances in Neural Information Processing Systems*, volume 36, pp. 19540–19569, 2023.

Jason Altschuler. Greed, hedging, and acceleration in convex optimization. Master's thesis, Massachusetts Institute of Technology, 2018.

Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging: Silver stepsize schedule for smooth convex optimization. *Mathematical Programming*, 213(1–2):1105–1118, 2025.

Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. SGD with large step sizes learns sparse features. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 903–925, 2023.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 948–1024, 2022.

Yuhang Cai, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter L. Bartlett. Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 71306–71351, 2024.

Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4330–4391, 2023.

Xuxing Chen, Krishnakumar Balasubramanian, Promit Ghosal, and Bhavya Agrawalla. From stability to chaos: Analyzing gradient descent dynamics in quadratic regression. *Transactions on Machine Learning Research*, 2024.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

Jeremy Cohen, Alex Damian, Ameet Talwalkar, J. Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows. In *International Conference on Learning Representations*, 2025.

Jeremy M. Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E. Dahl, and Justin Gilmer. Adaptive gradient methods at the edge of stability. *arXiv preprint arXiv:2207.14484*, 2022.

Michael Crawshaw and Mingrui Liu. Tight bounds for logistic regression with large stepsize gradient descent in low dimension. In *Proceedings of the 39th Annual Conference on Learning Theory (COLT)*, volume 336 of *Proceedings of Machine Learning Research*, pp. 1575–1610, 2026. arXiv:2602.12471.

Michael Crawshaw, Blake Woodworth, and Mingrui Liu. Constant stepsize local GD for logistic regression: Acceleration by instability. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11465–11492, 2025.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations*, 2023.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (S)GD over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 29406–29448, 2023.

Benjamin Grimmer. Provably faster gradient descent via long steps. *SIAM Journal on Optimization*, 34(3):2588–2608, 2024.

Benjamin Grimmer, Kevin Shu, and Alex L. Wang. Accelerated objective gap and gradient norm convergence for gradient descent via long steps. *INFORMS Journal on Optimization*, 7(2):156–169, 2025.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841, 2018.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Proceedings of the 32nd Annual Conference on Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1772–1798, 2019.

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory (ALT)*, volume 132 of *Proceedings of Machine Learning Research*, pp. 772–804, 2021.

Ziwei Ji, Nathan Srebro, and Matus Telgarsky. Fast margin maximization via dual acceleration. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4860–4869, 2021.

Jonathan Kelner, Annie Marsden, Vatsal Sharan, Aaron Sidford, Gregory Valiant, and Honglin Yuan. Big-step-little-step: Efficient gradient methods for objectives with multiple scales. In *Proceedings of the 35th Annual Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2431–2540, 2022.

Lingkai Kong and Molei Tao. Stochasticity of deterministic gradient descent: Large learning rate for multiscale objective function. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2625–2638, 2020.

Guy Kornowski and Ohad Shamir. The oracle complexity of simplex-based matrix games: Linear separability and Nash equilibria. In *Proceedings of the 38th Annual Conference on Learning Theory (COLT)*, volume 291 of *Proceedings of Machine Learning Research*, pp. 3327–3353, 2025.

Itai Kreisler, Mor Shpigel Nacson, Daniel Soudry, and Yair Carmon. Gradient descent monotonically decreases the sharpness of gradient flow solutions in scalar networks and beyond. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17684–17744, 2023.

Guanghui Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer, 2020.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: The catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

Zhouzi Li, Zixuan Wang, and Jian Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9983–9994, 2022.

Miao Lu, Beining Wu, Xiaodong Yang, and Difan Zou. Benign oscillation of stochastic gradient descent with large learning rate. In *International Conference on Learning Representations*, 2024.

Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34689–34708, 2022.

Chao Ma, Daniel Kunin, Lei Wu, and Lexing Ying. Beyond the quadratic approximation: The multiscale structure of neural network loss landscapes. *Journal of Machine Learning*, 1(3):247–267, 2022.

Si Yi Meng, Antonio Orvieto, Daniel Yiming Cao, and Christopher De Sa. Gradient descent on logistic regression with non-separable data and large step sizes. *arXiv preprint arXiv:2406.05033*, 2024.

Si Yi Meng, Baptiste Goujaud, Antonio Orvieto, and Christopher De Sa. Gradient descent on logistic regression: Do large step-sizes work with data on the sphere? *arXiv preprint arXiv:2507.11228*, 2025.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3420–3428, 2019a.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3051–3059, 2019b.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2nd edition, 2018.

Albert B. J. Novikoff. On convergence proofs for perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pp. 615–622, 1962.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018.

Alexander Tyurin. From logistic regression to the perceptron algorithm: Exploring gradient descent with large step sizes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20938–20946, 2025.

Guanghui Wang, Rafael Hanashiro, Etash Kumar Guha, and Jacob Abernethy. On accelerated perceptrons and beyond. In *International Conference on Learning Representations*, 2023.

Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large learning rate tames homogeneity: Convergence and balancing effect. In *International Conference on Learning Representations*, 2022.

Jingfeng Wu, Vladimir Braverman, and Jason D. Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74229–74256, 2023.

Jingfeng Wu, Peter L. Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. In *Proceedings of the 37th Annual Conference on Learning Theory (COLT)*, volume 247 of *Proceedings of Machine Learning Research*, pp. 5019–5073, 2024.

Jingfeng Wu, Pierre Marion, and Peter L. Bartlett. Large stepsizes accelerate gradient descent for regularized logistic regression. In *Advances in Neural Information Processing Systems*, volume 38, pp. 104485–104525, 2025.

Lei Wu, Chao Ma, and Weinan E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with SGD. *arXiv preprint arXiv:1802.08770*, 2018.

Ruiqi Zhang, Jingfeng Wu, Licong Lin, and Peter L. Bartlett. Minimax optimal convergence of gradient descent in logistic regression via large and adaptive stepsizes. *Journal of Machine Learning Research*, 27(124):1–31, 2026.

Zihan Zhang, Jason D. Lee, Simon S. Du, and Yuxin Chen. Anytime acceleration of gradient descent. In *Proceedings of the 38th Annual Conference on Learning Theory (COLT)*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5991–6013, 2025.

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *International Conference on Learning Representations*, 2023.

A Proofs for Section 2

This appendix states the general form of Theorem 2.2(i), in which an integer parameter $h \geq 2$ sets the tolerance $\theta_h = \gamma^2/h$ of the interval construction (Theorem A.1, proved in Appendix B), proves the parameter dependence stated in Section 2 (Proposition A.2) and the large-margin case (Proposition A.3), recalls the stable phase, and proves Corollary 2.3. Throughout, Assumption 2.1 holds, GD is run from $\mathbf{w}_0 = \mathbf{0}$, and $h \geq 2$ is an integer; the restriction $\gamma < 1/\sqrt{2}$ is stated wherever it is needed.

A.1 Explicit constants and the general theorem

Set

$$\theta_h := \frac{\gamma^2}{h}, \qquad \nu_h := \max\left\{0,\ \min\left\{r,\ \left\lfloor \gamma^{-2} + \frac{\gamma^{-2}-1}{h-1} \right\rfloor \right\} - 2 \right\}. \tag{29}$$

Since $h \geq 2$, the difference $\gamma^2 - \theta_h = \gamma^2(1 - 1/h)$ is positive, and multiplying numerator and denominator by h/γ^2 and then writing $h/(h-1) = 1 + 1/(h-1)$ gives, for every integer $h \geq 2$,

$$\frac{1-\theta_h}{\gamma^2 - \theta_h} = \frac{h - \gamma^2}{\gamma^2(h-1)} = \gamma^{-2} + \frac{\gamma^{-2}-1}{h-1}. \tag{30}$$

The quantity ν_h bounds the depth of the recursive interval construction. Increasing h lowers θ_h and does not increase the depth bound ν_h, but it increases the bound on the number of children created at each refinement. Since $\gamma \leq 1$ and $h \geq 2$, we have $0 \leq (\gamma^{-2}-1)/(h-1) \leq \gamma^{-2} - 1$, so the quantity inside the floor in (29) is at most $2\gamma^{-2} - 1$, and therefore

$$0 \leq \nu_h < 2\gamma^{-2}. \tag{31}$$

Moreover, $\nu_h = 0$ whenever $r \leq 2$. For $\nu_h \geq 1$ define the exponent

$$p_h := \nu_h + \left\lceil \frac{\nu_h \ln(18h\gamma^{-2}/\nu_h)}{\ln(\max\{2, \nu_h\})} \right\rceil, \tag{32}$$

and set $p_h := 0$ when $\nu_h = 0$. When $\nu_h \geq 1$, (31) gives $18h\gamma^{-2}/\nu_h > 9h \geq 18$, and $\ln(\max\{2,\nu_h\}) \geq \ln 2 > 0$, so the quotient in (32) is positive and p_h is an integer with $p_h \geq \nu_h + 1$. Finally, put

$$C_h := 80\gamma^{-4}h\big(n + h\ln(4\gamma^{-2})\big), \tag{33}$$

$$K_h := \exp(4) + h\gamma^{-2} + p_h + \ln(2\gamma^{-2}C_h), \qquad \eta_h^0 := \big\lceil 128\, K_h(1 + \ln K_h)\big\rceil. \tag{34}$$

Theorem A.1 (General form of Theorem 2.2(i)). *Suppose Assumption 2.1 holds with $0 < \gamma < 1/\sqrt{2}$. For every integer $h \geq 2$ and every $\eta \geq \eta_h^0$,*

$$\tau \leq C_h \, (\ln \eta)^{p_h}. \tag{35}$$

Theorem A.1 is proved in Appendix B. The low-margin case of Theorem 2.2(i) is Theorem A.1 instantiated at

$$h_* := \lceil \ln(16\gamma^{-2}) \rceil, \qquad C := C_{h_*}, \qquad p := p_{h_*}, \qquad \eta_0 := \eta_{h_*}^0. \tag{36}$$

For $\gamma < 1/\sqrt{2}$ we have $\gamma^{-2} > 2$, hence $\ln(16\gamma^{-2}) > \ln 32 > 3$ and $h_* \geq 4$. The instantiation is therefore admissible, the definitions (3)–(6) of the main text are (29)–(34) at $h = h_*$, and (35) at $h = h_*$ is (7).

A.2 Parameter dependence

The next proposition gives the orders of the selected quantities in n, r, and γ, with absolute constants in $O(\cdot)$; the bounds (37)–(39) hold on the whole range $0 < \gamma < 1/\sqrt{2}$, and only the statements after (40) are asymptotic.

Proposition A.2 (Parameter dependence). *Let $0 < \gamma < 1/\sqrt{2}$, let h_*, C, p, and η_0 be as in (36), and write $\nu := \nu_{h_*}$. Then*

$$C = O\big(n\gamma^{-4} \ln(\gamma^{-2}) + \gamma^{-4} \ln^3(\gamma^{-2})\big), \tag{37}$$

$$p \leq \nu \left(2 + \frac{\ln(18 h_* \gamma^{-2})}{\ln 2} \right) = O\big(\min\{r, \gamma^{-2}\} \ln(\gamma^{-2}) \big), \tag{38}$$

$$\eta_0 = O(A \ln A), \qquad \textit{where } \; A := \gamma^{-2} \ln(\gamma^{-2}) + \ln(en). \tag{39}$$

In particular $p = O(\gamma^{-2} \ln(\gamma^{-2}))$, and $p = 0$ whenever $r \leq 2$. If moreover the rank satisfies

$$r \geq \left\lfloor \frac{h_* - \gamma^2}{\gamma^2 (h_* - 1)} \right\rfloor = \left\lfloor \gamma^{-2} + \frac{\gamma^{-2} - 1}{h_* - 1} \right\rfloor, \tag{40}$$

then, as $\gamma \to 0$,

$$p = \gamma^{-2} + O\left(\gamma^{-2} \frac{\ln \ln(\gamma^{-2})}{\ln(\gamma^{-2})} \right) = (1 + o(1)) \gamma^{-2}, \tag{41}$$

and the threshold on the right of (40) is itself $(1 + o(1))\gamma^{-2}$.

Proof. Put $\omega := \ln(\gamma^{-2})$. Since $\gamma^{-2} > 2$, we have $\ln 2 < \omega < \gamma^{-2}$; in particular every absolute constant is $O(\omega)$.

The parameter h_*. Since $\ln(16\gamma^{-2}) = \omega + 4 \ln 2$ and a ceiling adds less than one,

$$\omega < h_* < \omega + 4 \ln 2 + 1, \tag{42}$$

so $h_* = O(\omega)$.

The prefactor C. Since $\ln(4\gamma^{-2}) = \omega + 2 \ln 2 = O(\omega)$ and $h_* = O(\omega)$, (33) at $h = h_*$ gives $C = O(\gamma^{-4} \omega (n + \omega^2))$, which is (37).

The exponent p. If $r \leq 2$, then $\nu = 0$, as observed after (31), and $p = 0$ by definition. More generally, (38) is trivial when $\nu = 0$, so let $\nu \geq 1$. The ceiling in (32) adds less than one, the logarithm in (32) is positive with $\ln(\max\{2, \nu\}) \geq \ln 2$ below it, and $18 h_* \gamma^{-2} / \nu \leq 18 h_* \gamma^{-2}$, so that, using $1 \leq \nu$ in the last step,

$$p < \nu + 1 + \frac{\nu \ln(18 h_* \gamma^{-2})}{\ln 2} \leq \nu \left(2 + \frac{\ln(18 h_* \gamma^{-2})}{\ln 2} \right).$$

Here $\ln(18 h_* \gamma^{-2}) = \ln 18 + \ln h_* + \omega = O(\omega)$ by (42), and $\nu \leq \min\{r, 2\gamma^{-2}\} \leq 2 \min\{r, \gamma^{-2}\}$, because $\nu \leq r - 2$ by (29) when $\nu \geq 1$ and $\nu < 2\gamma^{-2}$ by (31). This proves (38), and $p = O(\gamma^{-2} \omega)$ follows.

The stepsize threshold η_0. Each of the four terms of K_{h_*} in (34) is $O(A)$: $\exp(4) = O(1)$, $h_*\gamma^{-2} = O(\gamma^{-2}\omega)$ by (42), $p = O(\gamma^{-2}\omega)$ was just proved, and $\ln(2\gamma^{-2}C) = O(\omega + \ln(en))$ by (37), since $\ln\omega < \omega$ and $\ln(n+\omega^2) \le \ln n + \ln(1+\omega^2) \le \ln n + 2\omega$; here $\omega < \gamma^{-2}\omega \le A$ and $A \ge \gamma^{-2}\omega + 1 > 2$, because $\gamma^{-2}\omega > 2\ln 2 > 1$ and $\ln(en) \ge 1$. Hence $K_{h_*} \le cA$ for an absolute constant $c \ge 1$. On the other hand $K_{h_*} \ge \exp(4) > 1$, because the other three terms of (34) are nonnegative, the last one since $C \ge 80\gamma^{-4}h_*n > 1$. The function $s \mapsto s(1+\ln s)$ is increasing on $[1,\infty)$, and the ceiling in (34) adds less than one, so

$$\eta_0 < 128\,K_{h_*}(1+\ln K_{h_*}) + 1 \le 128\,cA\,(1+\ln c + \ln A) + 1 = O(A\ln A),$$

because $A > 2$ gives $\ln A > \ln 2$, so that $1 + \ln c = O(\ln A)$ and $1 = O(A\ln A)$. This proves (39).

The asymptotic exponent. All $O(\cdot)$ and $o(\cdot)$ statements below refer to $\gamma \to 0$, so that $\omega \to \infty$. Let $x := \gamma^{-2} + (\gamma^{-2}-1)/(h_*-1)$; the two expressions in (40) agree by (30) at $h = h_*$. Since $0 < \gamma^{-2} - 1 < \gamma^{-2}$ and $h_* > \omega$ by (42),

$$x = \gamma^{-2}\big(1 + O(1/\omega)\big), \tag{43}$$

so the threshold $\lfloor x \rfloor$ in (40) is $(1+o(1))\gamma^{-2}$. Assume now (40). Since $x > \gamma^{-2} > 2$, the minimum in (29) equals $\lfloor x \rfloor \ge 2$, so $\nu = \lfloor x \rfloor - 2$, and (43) gives

$$\nu = \gamma^{-2}\big(1 + O(1/\omega)\big). \tag{44}$$

In particular $\nu \to \infty$, so eventually $\max\{2,\nu\} = \nu$ and $\ln\nu = (1+o(1))\omega$, while $18h_*\gamma^{-2}/\nu = O(h_*) = O(\omega)$ by (42), so $\ln(18h_*\gamma^{-2}/\nu) = O(\ln\omega)$. Since the ceiling in (32) adds less than one,

$$1 \le p - \nu < 1 + \frac{\nu\ln(18h_*\gamma^{-2}/\nu)}{\ln\nu} = O\!\left(\gamma^{-2}\frac{\ln\omega}{\omega}\right).$$

Together with (44), and since $1/\omega = O(\ln\omega/\omega)$, this gives $p - \gamma^{-2} = (p-\nu) + (\nu - \gamma^{-2}) = O(\gamma^{-2}\ln\omega/\omega)$, which is the first equality in (41). The second follows because $\ln\omega/\omega \to 0$. □

Under the rank condition (40), as $\gamma \to 0$, the proof separates two contributions to $p - \gamma^{-2}$. The depth error $\nu - \gamma^{-2}$ is $O(\gamma^{-2}/\ln(\gamma^{-2}))$ by (44), while the overhead $p - \nu$, which absorbs the fixed coefficient of the factorial bound into the exponent, carries the additional factor $\ln\ln(\gamma^{-2})$ and determines the order displayed in (41). The orders stated in Section 2 are those of Proposition A.2. The bounds on C and p are (37) and (38), and $A \le (\gamma^{-2} + \ln n)(1 + \ln(\gamma^{-2}))$ turns (39) into $\eta_0 = \widetilde{O}(\gamma^{-2} + \ln n)$, where $\widetilde{O}$ hides factors polylogarithmic in $\gamma^{-2} + \ln n$.

A.3 THE LARGE-MARGIN CASE

This part proves Proposition A.3, the second case of Theorem 2.2(i), and shows along the way that for $\gamma \ge 1/\sqrt{2}$ the loss decreases monotonically from initialization for every stepsize. Neither the parameter h nor the horizon (11) is involved. We write $\ell(u) := \ln(1+e^{-u})$, so that $L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \ell(\langle \mathbf{z}_i, \mathbf{w}\rangle)$.

Proposition A.3 (Large margin; the second case of Theorem 2.2(i))**.** *Suppose Assumption 2.1 holds with* $\gamma \ge 1/\sqrt{2}$*. If* $\eta \ge 64n^2$*, then* $\tau \le 1$.

Proof of Proposition A.3. For each i, set $a_i := \langle \mathbf{z}_i, \mathbf{w}_*\rangle \ge \gamma$ and $\mathbf{u}_i := \mathbf{z}_i - a_i\mathbf{w}_*$, so that $\langle \mathbf{u}_i, \mathbf{w}_*\rangle = 0$ and $a_i^2 + \|\mathbf{u}_i\|^2 = \|\mathbf{z}_i\|^2 \le 1$. Hence $\|\mathbf{u}_i\|^2 \le 1 - \gamma^2$ and $\|\mathbf{z}_i\|^2 \ge \gamma^2$, and the Cauchy–Schwarz inequality gives, for all i, j,

$$\langle \mathbf{z}_i, \mathbf{z}_j\rangle = a_ia_j + \langle \mathbf{u}_i, \mathbf{u}_j\rangle \ge \gamma^2 - (1-\gamma^2) = 2\gamma^2 - 1 \ge 0, \tag{45}$$

where the last inequality is the hypothesis $\gamma^2 \ge 1/2$.

By the first identity in (10), which holds for every $t \ge 0$ and every $\eta > 0$, and by (45),

$$\langle \mathbf{z}_i, \mathbf{w}_{t+1}\rangle - \langle \mathbf{z}_i, \mathbf{w}_t\rangle = \frac{\eta}{n}\sum_{j=1}^n q_{j,t}\langle \mathbf{z}_i, \mathbf{z}_j\rangle \ge 0 \qquad (i = 1,\dots,n;\ t \ge 0;\ \eta > 0). \tag{46}$$

Thus every signed margin is nondecreasing in t, and since ℓ is decreasing, $L(\mathbf{w}_{t+1}) \le L(\mathbf{w}_t)$ for every $t \ge 0$ and every $\eta > 0$, which is the assertion made after Theorem 2.2.

It remains to examine the first step. Since $\mathbf{w}_0 = \mathbf{0}$, every weight equals $q_{j,0} = 1/2$, so $\mathbf{w}_1 = \frac{\eta}{2n}\sum_{j=1}^n \mathbf{z}_j$. Keeping only the term $j = i$ in the sum below, which is allowed by (45),

$$\langle \mathbf{z}_i, \mathbf{w}_1\rangle = \frac{\eta}{2n}\sum_{j=1}^n \langle \mathbf{z}_i, \mathbf{z}_j\rangle \geq \frac{\eta}{2n}\|\mathbf{z}_i\|^2 \geq \frac{\eta}{4n} \qquad (i = 1, \ldots, n).$$

Finally, $\ell(u) \leq e^{-u}$ and $e^s \geq s^2/2$ for $s > 0$, so for $\eta \geq 64n^2$,

$$L(\mathbf{w}_1) \leq \frac{1}{n}\sum_{i=1}^n \exp\big(-\langle \mathbf{z}_i, \mathbf{w}_1\rangle\big) \leq \exp\left(-\frac{\eta}{4n}\right) \leq \frac{32n^2}{\eta^2} \leq \frac{1}{2\eta} < \frac{1}{\eta},$$

where the next-to-last inequality is equivalent to $\eta \geq 64n^2$. By the definition (2) of the phase transition time, $\tau \leq 1$. □

A.4 The stable phase

Part (ii) of Theorem 2.2 is the monotonicity statement in the stable-phase part of Wu et al. (2024, Theorem 1), which we recall. If $L(\mathbf{w}_s) \leq 1/\eta$ at some step s, then $L(\mathbf{w}_{t+1}) \leq L(\mathbf{w}_t)$ for every $t \geq s$, and for every $t > s$,

$$L(\mathbf{w}_t) \leq \frac{2F(\mathbf{w}_s) + \ln^2\big(\gamma^2\eta(t-s)\big)}{\gamma^2\eta(t-s)}, \qquad \text{where } F(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^n \exp(-y_i\mathbf{x}_i^\top\mathbf{w}); \tag{47}$$

the rate is their Lemma 11, and the monotonicity part holds already under the weaker condition $L(\mathbf{w}_s) \leq 2/\eta$ by their Lemma 10. Since $L(\mathbf{w}_\tau) \leq 1/\eta$ by (2), taking $s = \tau$ gives part (ii). With $t = \tau + k$, the rate (47) also shows how the loss continues to decrease after the phase transition at a fixed stepsize. Here the term $F(\mathbf{w}_\tau)$ is controlled by the definition of τ. Writing $a_i := \exp(-y_i\mathbf{x}_i^\top\mathbf{w})$, we have $1 + nF(\mathbf{w}) = 1 + \sum_{i=1}^n a_i \leq \prod_{i=1}^n (1 + a_i) = \exp(nL(\mathbf{w}))$, so $F(\mathbf{w}_\tau) \leq (\exp(n/\eta) - 1)/n$, which is at most $(e-1)/\eta$ whenever $\eta \geq n$.

A.5 Proof of Corollary 2.3

Proof of Corollary 2.3. (i) Let $0 < \epsilon \leq \epsilon_0$ and $\eta = 1/\epsilon$, so that $\eta \geq \eta_0$ if $\gamma < 1/\sqrt{2}$ and $\eta \geq 64n^2$ if $\gamma \geq 1/\sqrt{2}$. By Theorem 2.2(i), with $C = 1$ and $p = 0$ in the latter case,

$$\tau \leq C\,(\ln\eta)^p = C\,\ln^p(1/\epsilon).$$

By the definition (2) of the phase transition time, $L(\mathbf{w}_t) \leq 1/\eta = \epsilon$ at $t = \tau$, and by Theorem 2.2(ii) the loss decreases monotonically from step τ onwards.

(ii) Let $p \geq 1$, so that $\gamma < 1/\sqrt{2}$, and recall $T_0 = \lceil C(\ln\eta_0)^p\rceil$. For every integer $T \geq T_0$, the stepsize $\eta = \exp((T/C)^{1/p})$ satisfies $\ln\eta = (T/C)^{1/p} \geq \ln\eta_0$, hence $\eta \geq \eta_0$, and Theorem 2.2(i) gives

$$\tau \leq C\,(\ln\eta)^p = C \cdot \frac{T}{C} = T.$$

Therefore $L(\mathbf{w}_\tau) \leq 1/\eta$, and by Theorem 2.2(ii) the loss decreases monotonically from step $\tau \leq T$ onwards, so $L(\mathbf{w}_T) \leq 1/\eta = \exp(-(T/C)^{1/p})$. If instead $p = 0$, let $0 < \epsilon \leq \epsilon_0$. By part (i), $L(\mathbf{w}_t) \leq \epsilon$ for some $t \leq C$ and the loss decreases monotonically afterwards; hence $L(\mathbf{w}_T) \leq \epsilon$ for every integer $T \geq C$. □

B Proofs for Section 3

This appendix proves Theorem A.1 in full, following the structure of Section 3. Throughout, Assumption 2.1 holds and GD is run from $\mathbf{w}_0 = \mathbf{0}$; in Appendices B.1–B.4 we fix an integer $h \geq 2$ and use the general quantities θ_h, ν_h, p_h, C_h, K_h, and η_h^0 of (29)–(34). The quantities of Section 3 are the case $h = h_*$ of (36), that is, $\theta = \theta_{h_*}$, $\nu = \nu_{h_*}$, $p = p_{h_*}$, $C = C_{h_*}$, and $\eta_0 = \eta_{h_*}^0$. Appendices B.1–B.3 use only three conditions on the stepsize, namely $\eta \geq 4$, $\Lambda = \ln\eta > 10$, and the root condition $4\Lambda \leq \eta\theta_h$. Lemma B.12 in Appendix B.4 shows that every $\eta \geq \eta_h^0$ satisfies these conditions. The phase transition time τ is finite for every $\eta > 0$ by the bound (8) of Wu et al. (2024), so U in (11) is well defined. All other notation is as in Section 3.

B.1 TRAJECTORY ESTIMATES

This part proves the update decomposition (10), the bound on the cumulative coefficient on the finite horizon, Lemma 3.1, and one running-average estimate that gives (17) and is used again in Appendix B.2. Nothing in it depends on h. The standing assumption is $\eta \geq 4$. We write $\ell(u) := \ln(1 + e^{-u})$, so that $L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\langle \mathbf{z}_i, \mathbf{w}\rangle)$ by (1). Since $\ell'(u) = -1/(1 + e^u) < 0$ and $\ell''(u) = e^u/(1 + e^u)^2 > 0$, the function ℓ is decreasing and convex, L is convex, and $\ell(u) \leq e^{-u}$ because $\ln(1 + x) \leq x$ for $x \geq 0$. The weights in (9) satisfy $0 < q_{i,t} < 1$, so that $0 < Q_t < n$, $\mathbf{v}_t$ is well defined, and $0 < \mu_t < \eta$. By Assumption 2.1, $\gamma \leq 1$, and every vector $\mathbf{b} = \sum_i c_i \mathbf{z}_i$ of the convex hull of the signed examples, where $c_i \geq 0$ and $\sum_i c_i = 1$, satisfies

$$\|\mathbf{b}\| \leq \sum_{i=1}^{n} c_i \|\mathbf{z}_i\| \leq 1, \qquad \langle \mathbf{b}, \mathbf{w}_*\rangle = \sum_{i=1}^{n} c_i \langle \mathbf{z}_i, \mathbf{w}_*\rangle \geq \gamma. \tag{48}$$

Lemma B.1 (Update decomposition)**.** *For every $t \geq 0$,*

$$\nabla L(\mathbf{w}_t) = -\frac{Q_t}{n}\mathbf{v}_t = -\frac{\mu_t}{\eta}\mathbf{v}_t, \tag{49}$$

and consequently (10) holds.

Proof. Since $\ell'(\langle \mathbf{z}_i, \mathbf{w}_t\rangle) = -q_{i,t}$, the chain rule gives $\nabla L(\mathbf{w}_t) = -\frac{1}{n}\sum_i q_{i,t}\mathbf{z}_i = -(Q_t/n)\mathbf{v}_t$, and $Q_t/n = \mu_t/\eta$. Substituting (49) into (GD) with $\eta_t \equiv \eta$ gives $\mathbf{w}_{t+1} - \mathbf{w}_t = \mu_t \mathbf{v}_t$. The coefficients $q_{i,t}/Q_t$ are positive and sum to one, so $\mathbf{v}_t$ lies in the convex hull of the signed examples, and (48) gives the last two properties in (10). □

Recall $\widehat{T} = \lfloor \gamma^2 \exp(\eta/4)/\eta \rfloor$ and $U = \min\{\tau, \widehat{T}\}$ from (11). U is a well-defined nonnegative integer because τ is finite, as recalled at the start of this appendix. All counting in this appendix is restricted to the steps $0, 1, \ldots, U - 1$, and the horizon is removed only in the last step of the proof of Theorem A.1, because the following bound is proved only up to $\widehat{T}$. Its proof applies the split optimization technique introduced by Wu et al. (2024, Lemma 7).

Lemma B.2 (Norm and cumulative-coefficient bounds)**.** *Let $\eta \geq 4$. For every integer $1 \leq T \leq \widehat{T}$,*

$$\|\mathbf{w}_T\| \leq \frac{2\eta}{\gamma}, \qquad \sum_{t=0}^{T-1} \mu_t \leq 2\eta\gamma^{-2}. \tag{50}$$

Proof. Write $\mathbf{g}_t := \nabla L(\mathbf{w}_t)$. By (49) and (10),

$$\|\mathbf{g}_t\| = \frac{\mu_t}{\eta}\|\mathbf{v}_t\| \leq \frac{\mu_t}{\eta}, \qquad \langle \mathbf{g}_t, \mathbf{w}_*\rangle = -\frac{\mu_t}{\eta}\langle \mathbf{v}_t, \mathbf{w}_*\rangle \leq -\frac{\gamma\mu_t}{\eta}.$$

Define the fixed comparators

$$\mathbf{u}_1 := \frac{\eta}{4\gamma}\mathbf{w}_*, \qquad \mathbf{u}_2 := \frac{\eta}{2\gamma}\mathbf{w}_*, \qquad \mathbf{u} := \mathbf{u}_1 + \mathbf{u}_2 = \frac{3\eta}{4\gamma}\mathbf{w}_*.$$

The component $\mathbf{u}_2$ cancels the quadratic term of a step. Indeed, the two bounds on $\mathbf{g}_t$ and $0 < \mu_t < \eta$ give

$$2\langle \mathbf{g}_t, \mathbf{u}_2\rangle + \eta\|\mathbf{g}_t\|^2 \leq -\mu_t + \frac{\mu_t^2}{\eta} = -\mu_t\left(1 - \frac{\mu_t}{\eta}\right) \leq 0. \tag{51}$$

Since $\mathbf{w}_{t+1} - \mathbf{u} = (\mathbf{w}_t - \mathbf{u}) - \eta\mathbf{g}_t$, expanding the square and splitting $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$ gives

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{u}\|^2 &= \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\langle \mathbf{g}_t, \mathbf{u}_1 - \mathbf{w}_t\rangle + \eta\big(2\langle \mathbf{g}_t, \mathbf{u}_2\rangle + \eta\|\mathbf{g}_t\|^2\big) \\ &\leq \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\big(L(\mathbf{u}_1) - L(\mathbf{w}_t)\big),\end{aligned}$$

where the inequality uses (51) and the gradient inequality $L(\mathbf{u}_1) \geq L(\mathbf{w}_t) + \langle \mathbf{g}_t, \mathbf{u}_1 - \mathbf{w}_t\rangle$ of the convex function L. Every signed margin of $\mathbf{u}_1$ satisfies $\langle \mathbf{z}_i, \mathbf{u}_1\rangle = (\eta/(4\gamma))\langle \mathbf{z}_i, \mathbf{w}_*\rangle \geq \eta/4$, so $L(\mathbf{u}_1) \leq \ell(\eta/4) \leq e^{-\eta/4}$ because ℓ is decreasing and $\ell(u) \leq e^{-u}$. Summing over $t = 0, \ldots, T-1$, using $\mathbf{w}_0 = \mathbf{0}$, and discarding the nonpositive term $-2\eta\sum_{t<T} L(\mathbf{w}_t)$, we obtain

$$\|\mathbf{w}_T - \mathbf{u}\|^2 \leq \|\mathbf{u}\|^2 + 2\eta T e^{-\eta/4} \leq \|\mathbf{u}\|^2 + 2\gamma^2,$$

where the last inequality is $T \le \widehat{T} \le \gamma^2 e^{\eta/4}/\eta$. The triangle inequality and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$ now give

$$\|\mathbf{w}_T\| \le \|\mathbf{u}\| + \|\mathbf{w}_T - \mathbf{u}\| \le 2\|\mathbf{u}\| + \sqrt{2}\,\gamma = \frac{3\eta}{2\gamma} + \sqrt{2}\,\gamma \le \frac{2\eta}{\gamma}.$$

The last inequality is equivalent to $2\sqrt{2}\,\gamma^2 \le \eta$, which holds because $\gamma \le 1$ and $\eta \ge 4 > 2\sqrt{2}$. For the second bound in (50), Lemma B.1 and $\mathbf{w}_0 = \mathbf{0}$ give $\mathbf{w}_T = \sum_{t<T} \mu_t \mathbf{v}_t$, and since $\langle \mathbf{v}_t, \mathbf{w}_* \rangle \ge \gamma$ and $\mu_t > 0$,

$$\gamma \sum_{t=0}^{T-1} \mu_t \le \left\langle \sum_{t=0}^{T-1} \mu_t \mathbf{v}_t, \mathbf{w}_* \right\rangle = \langle \mathbf{w}_T, \mathbf{w}_* \rangle \le \|\mathbf{w}_T\| \le \frac{2\eta}{\gamma}.$$

Dividing by γ completes the proof. □

Recall from Section 3.1 the sets $\mathcal{T}_- = \{t < U : \langle \mathbf{z}_i, \mathbf{w}_t \rangle < 0 \text{ for some } i\}$ and $\mathcal{T}_+ = \{0, 1, \ldots, U-1\} \setminus \mathcal{T}_-$, so that every signed margin is nonnegative at each $t \in \mathcal{T}_+$.

Lemma B.3 (General form of Lemma 3.1)**.** *Let $\eta \ge 4$. Then*

$$\sum_{t<U} \mu_t \le 2\eta\gamma^{-2}, \qquad |\mathcal{T}_-| < 4n\gamma^{-2}, \tag{52}$$

and at every $t \in \mathcal{T}_+$,

$$\mu_t \ge \frac{\eta}{2} L(\mathbf{w}_t) > \frac{1}{2}, \qquad 0 \le \langle \mathbf{w}_t, \mathbf{v}_t \rangle < \Lambda. \tag{53}$$

Lemma 3.1 is this statement; neither it nor its proof involves h.

Proof. The proof of Lemma 3.1 in Section 3.1 is complete except for the first bound in (52), which it defers to this appendix. That bound is trivial when $U = 0$, and for $U \ge 1$ it is Lemma B.2 with $T = U \le \widehat{T}$. The remaining assertions then follow verbatim from the proof in Section 3.1, which uses only the weights (9), the convexity of ℓ, and $L(\mathbf{w}_t) > 1/\eta$ for $t < U \le \tau$, by the definition (2) of τ. □

The next lemma gives, for every starting step, the properties of the running averages of update directions used in Appendix B.2.

Lemma B.4 (Running averages of the update directions)**.** *Let $\eta > 0$, let $a \ge 0$ be an integer and, for every integer $t > a$, define*

$$S_t^{(a)} := \sum_{k=a}^{t-1} \mu_k, \qquad \bar{\mathbf{v}}_t^{(a)} := \frac{1}{S_t^{(a)}} \sum_{k=a}^{t-1} \mu_k \mathbf{v}_k = \frac{\mathbf{w}_t - \mathbf{w}_a}{S_t^{(a)}}.$$

Then, for every $t > a$,

$$\bar{\mathbf{v}}_t^{(a)} \in \operatorname{conv}\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}, \qquad \|\bar{\mathbf{v}}_t^{(a)}\| \le 1, \qquad \langle \bar{\mathbf{v}}_t^{(a)}, \mathbf{w}_* \rangle \ge \gamma,$$
$$\|\bar{\mathbf{v}}_{t+1}^{(a)} - \bar{\mathbf{v}}_t^{(a)}\| \le \frac{2\mu_t}{S_{t+1}^{(a)}} \le 2 \ln \frac{S_{t+1}^{(a)}}{S_t^{(a)}}, \tag{54}$$

and, for all integers $a < s \le t$,

$$\|\bar{\mathbf{v}}_t^{(a)} - \bar{\mathbf{v}}_s^{(a)}\| \le \sum_{k=s}^{t-1} \|\bar{\mathbf{v}}_{k+1}^{(a)} - \bar{\mathbf{v}}_k^{(a)}\| \le 2 \ln \frac{S_t^{(a)}}{S_s^{(a)}}. \tag{55}$$

Proof. Fix a and, in this proof only, write S_t and $\bar{\mathbf{v}}_t$ for $S_t^{(a)}$ and $\bar{\mathbf{v}}_t^{(a)}$. Every μ_k is positive, so $S_t > 0$, and $\mathbf{w}_t - \mathbf{w}_a = \sum_{k=a}^{t-1} \mu_k \mathbf{v}_k$ by Lemma B.1. Thus $\bar{\mathbf{v}}_t$ is a convex combination, with weights μ_k / S_t, of vectors of the convex hull of the signed examples; it therefore lies in this convex hull, and (48) gives the norm and margin bounds. Next, $S_{t+1} = S_t + \mu_t$ and $S_{t+1}\bar{\mathbf{v}}_{t+1} = S_t \bar{\mathbf{v}}_t + \mu_t \mathbf{v}_t$, so

$$\bar{\mathbf{v}}_{t+1} - \bar{\mathbf{v}}_t = \frac{\mu_t}{S_{t+1}} \big(\mathbf{v}_t - \bar{\mathbf{v}}_t\big), \qquad \|\bar{\mathbf{v}}_{t+1} - \bar{\mathbf{v}}_t\| \le \frac{\mu_t}{S_{t+1}} \big(\|\mathbf{v}_t\| + \|\bar{\mathbf{v}}_t\|\big) \le \frac{2\mu_t}{S_{t+1}}.$$

Moreover, $\mu_t/S_{t+1} = 1 - S_t/S_{t+1} \le \ln(S_{t+1}/S_t)$, because $1 - y \le -\ln y$ for every $y > 0$. This proves (54). The first inequality in (55) is the triangle inequality, and the second follows by summing the last bound in (54) over $s \le k < t$, where the logarithms telescope. □

For $a = 0$, since $\mathbf{w}_0 = \mathbf{0}$, the sum $S_t^{(0)}$ is the cumulative coefficient $S_t = \sum_{k<t} \mu_k$ of Section 3.2 and $\bar{\mathbf{v}}_t^{(0)} = \mathbf{w}_t/S_t = \boldsymbol{\xi}_t$ is the normalized iterate; for an interval $I = \{a, \dots, b\}$, the average $\bar{\mathbf{v}}_t^{(a)}$ is the local average $\boldsymbol{\zeta}_t$ in the proof of Lemma 3.2. At step zero every weight equals $q_{i,0} = 1/2$, so $S_1 = \mu_0 = \eta/2$, and if $\eta \ge 4$ and $U \ge 2$, then $S_U \le 2\eta\gamma^{-2}$ by Lemma B.2 with $T = U \le \widehat{T}$. Hence (55) with $a = 0$, $s = 1$, and $t = U$ gives

$$\sum_{t=1}^{U-1} \|\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t\| \le 2 \ln \frac{S_U}{S_1} \le 2 \ln(4\gamma^{-2}). \tag{56}$$

When $U \le 1$ the sum on the left of (56) is empty, and its bound by $2\ln(4\gamma^{-2}) > 0$ holds trivially because $\gamma \le 1$. Together with (54) for $a = 0$, this is (17).

B.2 Constructing the hierarchy of intervals

This part proves Lemma 3.2 for a general integer $h \ge 2$, with the exact cut points of both constructions, and shows that Lemmas 3.3 and 3.4 hold for every such h. Throughout, $\eta \ge 4$ and $\Lambda > 10$, and every interval is a set of consecutive steps in $\mathcal{T}_+$, so that Lemma B.3 provides the budget $\sum_{t<U} \mu_t \le 2\eta\gamma^{-2}$ and, at each step t of the interval, the bounds $\mu_t > 1/2$ and $0 \le \langle \mathbf{w}_t, \mathbf{v}_t \rangle < \Lambda$. We also use that $\mu_t > 0$, because every weight $q_{i,t}$ in (9) is positive, that $\|\mathbf{v}_t\| \le 1$ at every step (Lemma B.1), and that $0 < \theta_h < \gamma^2 \le 1$, because $h \ge 2$ and $\gamma \le 1$ (Assumption 2.1). Figure 1 gives an overview of the two constructions below and of the counting argument of Appendix B.3. In the figure, θ and ν stand for θ_h and ν_h.

Definition B.5 (θ_h-reference directions)**.** Let $I \subseteq \mathcal{T}_+$ be a set of consecutive steps. Vectors $\mathbf{a}_1, \dots, \mathbf{a}_j$ are *θ_h-reference directions for I* if

(a) $\mathbf{a}_i \in \operatorname{conv}\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ for every $1 \le i \le j$;
(b) $\langle \mathbf{a}_i, \mathbf{a}_k \rangle \le \theta_h$ whenever $1 \le i < k \le j$;
(c) $\langle \mathbf{a}_i, \mathbf{v}_t \rangle \le \theta_h$ whenever $1 \le i \le j$ and $t \in I$.

This is (14) with $\theta = \theta_h$. Condition (b) is void when $j = 1$. The conditions are local to one interval, and directions attached to different roots, or to different children of one parent, are never compared with each other.

Running averages. For integers $t \ge a \ge 0$ write $S_t^{(a)} := \sum_{k=a}^{t-1} \mu_k$, an empty sum when $t = a$, and for $t > a$ write $\boldsymbol{\zeta}_t := (\mathbf{w}_t - \mathbf{w}_a)/S_t^{(a)}$; the starting step a is fixed in each proof. In the notation of Lemma B.4, $\boldsymbol{\zeta}_t = \bar{\mathbf{v}}_t^{(a)}$, and $\boldsymbol{\xi}_t = \mathbf{w}_t/S_t = \bar{\mathbf{v}}_t^{(0)}$ with $S_t = S_t^{(0)}$. That lemma gives

$$\boldsymbol{\zeta}_t \in \operatorname{conv}\{\mathbf{z}_1, \dots, \mathbf{z}_n\} \quad (t > a), \qquad \|\boldsymbol{\zeta}_t - \boldsymbol{\zeta}_s\| \le 2 \ln \frac{S_t^{(a)}}{S_s^{(a)}} \quad (a < s \le t). \tag{57}$$

Both constructions below freeze such an average and start a new interval as soon as the average has moved by more than $\theta_h/2$.

Root partition. At $h = h_*$ the next lemma is Lemma 3.2(i). Its hypothesis (58) is the root condition stated at the beginning of this appendix and enters only in the last display of the proof.

Lemma B.6 (General form of Lemma 3.2(i))**.** *Assume*

$$4\Lambda \le \eta\theta_h. \tag{58}$$

There are an integer $M \ge 0$ and nonempty sets $J_1, \dots, J_M$ of consecutive steps such that

$$\begin{aligned} &\mathcal{T}_+ \setminus \{0\} = J_1 \,\dot\cup \cdots \dot\cup\, J_M, \qquad \max J_i < \min J_{i+1} \ \ (1 \le i < M), \\ &M \le |\mathcal{T}_-| + 1 + \left\lceil \frac{4\ln(4\gamma^{-2})}{\theta_h} \right\rceil, \end{aligned} \tag{59}$$

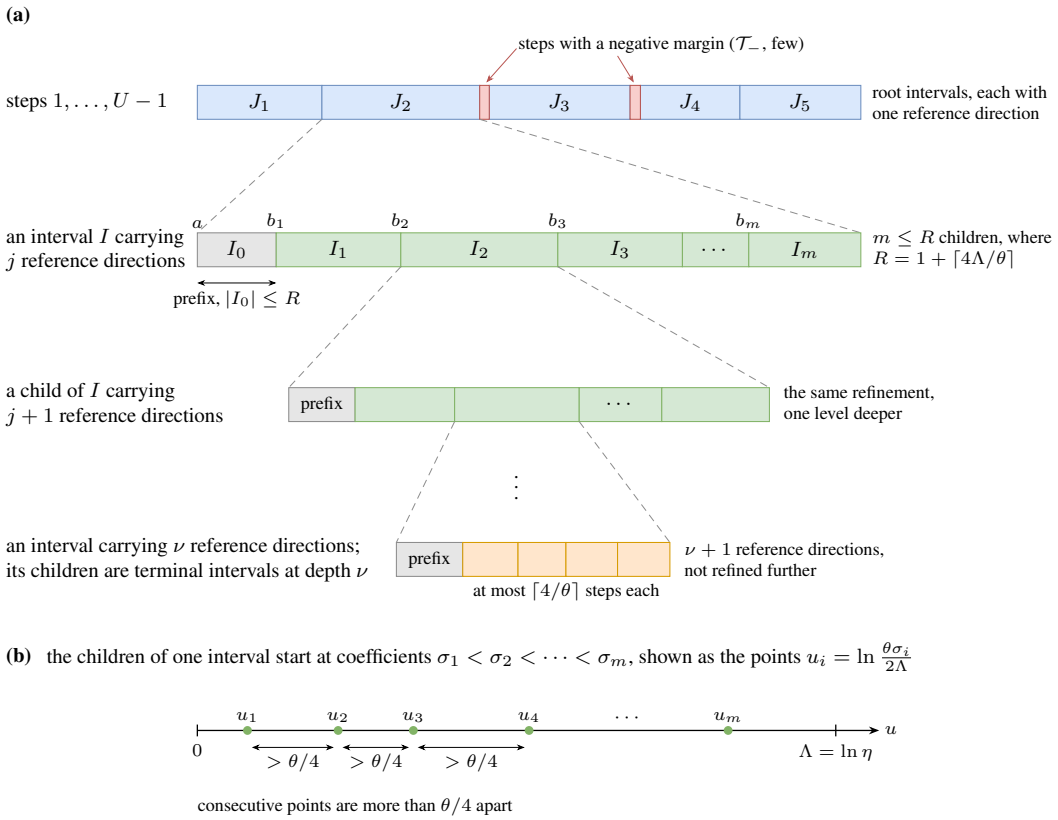

Figure 1: The hierarchy of intervals. (a) The steps in $\mathcal{T}_+ \setminus \{0\}$ are partitioned into root intervals, each carrying one reference direction. An interval carrying j reference directions is split into a prefix of at most R steps and at most R children, and each child inherits the j directions and gains one more (Lemma 3.2). An interval reached after ν refinements carries $\nu+1$ reference directions; it is terminal, is not refined further, and has at most $\lceil 4/\theta \rceil$ steps (Lemma 3.4). (b) The children of one interval start at coefficients σ_i, measured from the start of the parent, and the points $u_i = \ln(\theta\sigma_i/(2\Lambda))$ are more than $\theta/4$ apart, so a sum over the children can be compared with an integral along this axis. Lemma 3.5 applies this at every level of the hierarchy, which yields the factorial saving.

where $M = 0$ when $\mathcal{T}_+ \setminus \{0\}$ is empty. For every $1 \le i \le M$, the vector $\mathbf{a}_1^{(i)} := \boldsymbol{\xi}_{\min J_i}$ is a θ_h-reference direction for J_i.

Proof. Put $\mathcal{A} := \mathcal{T}_+ \setminus \{0\}$. If $\mathcal{A} = \varnothing$, take $M = 0$. Otherwise $U \ge 2$, because $\mathcal{A} \subseteq \{1, \dots, U-1\}$.

Runs. Let $\mathcal{K}_g = \{s_g, s_g + 1, \dots, e_g\}$, $1 \le g \le G$, be the maximal runs of consecutive integers in $\mathcal{A}$, in chronological order. Between two consecutive runs lies a step of $\{1, \dots, U-1\} \setminus \mathcal{T}_+ \subseteq \mathcal{T}_-$, and different gaps contain different steps, so $G - 1 \le |\mathcal{T}_-|$.

Cut points. For each $1 \le g \le G$ set $c_{g,1} := s_g$ and define recursively

$$c_{g,j+1} := \min\left(\left\{ t \in \{c_{g,j}+1, \dots, e_g\} : \|\boldsymbol{\xi}_t - \boldsymbol{\xi}_{c_{g,j}}\| > \frac{\theta_h}{2} \right\} \cup \{e_g + 1\} \right). \tag{60}$$

As long as $c_{g,j} \le e_g$, every element of the set in (60) is at least $c_{g,j} + 1$, so $c_{g,j} < c_{g,j+1} \le e_g + 1$. Hence the value $e_g + 1$ is reached after finitely many applications of (60), $m_g := \min\{j \ge 1 : c_{g,j+1} = e_g + 1\}$ is finite, and $c_{g,j} \le e_g$ for $1 \le j \le m_g$ by the minimality of m_g. The pieces

$$J_{g,j} := \{c_{g,j}, c_{g,j} + 1, \dots, c_{g,j+1} - 1\} \qquad (1 \le j \le m_g)$$

are therefore nonempty and partition $\mathcal{K}_g$ in chronological order. A crossing step $c_{g,j+1} \le e_g$ is the first step of the next piece, so no step belongs to two pieces. Moreover,

$$\|\boldsymbol{\xi}_t - \boldsymbol{\xi}_{c_{g,j}}\| \le \frac{\theta_h}{2} \qquad (t \in J_{g,j}), \tag{61}$$

because the left side vanishes at $t = c_{g,j}$, while a step t with $c_{g,j} < t < c_{g,j+1}$ lies in the index range of (60) and precedes the minimum, so it fails the strict inequality there.

Counting. For $1 \le j < m_g$ we have $c_{g,j+1} \le e_g$, so (60) and the triangle inequality give

$$\frac{\theta_h}{2} < \|\boldsymbol{\xi}_{c_{g,j+1}} - \boldsymbol{\xi}_{c_{g,j}}\| \le \sum_{t=c_{g,j}}^{c_{g,j+1}-1} \|\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t\|.$$

The summation range is the piece $J_{g,j}$ itself, and the pieces are pairwise disjoint subsets of $\{1, \dots, U-1\}$. There are $M - G$ pairs (g, j) with $1 \le j < m_g$, where $M := \sum_{g=1}^{G} m_g$. Summing over these pairs and applying (56),

$$\frac{(M-G)\theta_h}{2} \le \sum_{t=1}^{U-1} \|\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t\| \le 2\ln(4\gamma^{-2}).$$

Hence $M - G \le 4\ln(4\gamma^{-2})/\theta_h \le \lceil 4\ln(4\gamma^{-2})/\theta_h \rceil$, and together with $G \le |\mathcal{T}_-| + 1$ this is the bound on M. Relabeling the M pieces chronologically as $J_1, \dots, J_M$ gives (59).

The reference direction. Fix $1 \le i \le M$, put $c := \min J_i$, and let $t \in J_i$. By (57) with $a = 0$, the vector $\mathbf{a}_1^{(i)} = \boldsymbol{\xi}_c$ lies in the convex hull of the signed examples, which is condition (a) of Definition B.5; condition (b) is void. Since $t \ge 1$, we have $S_t \ge S_1 = \mu_0 = \eta/2$, as computed before (56). Writing $\boldsymbol{\xi}_c = \boldsymbol{\xi}_t + (\boldsymbol{\xi}_c - \boldsymbol{\xi}_t)$ and using the Cauchy–Schwarz inequality with $\|\mathbf{v}_t\| \le 1$, then $\langle \mathbf{w}_t, \mathbf{v}_t \rangle < \Lambda$ and (61), and finally (58),

$$\langle \mathbf{a}_1^{(i)}, \mathbf{v}_t \rangle \le \frac{\langle \mathbf{w}_t, \mathbf{v}_t \rangle}{S_t} + \|\boldsymbol{\xi}_c - \boldsymbol{\xi}_t\| < \frac{2\Lambda}{\eta} + \frac{\theta_h}{2} \le \theta_h,$$

which is condition (c). □

Refinement. At $h = h_*$ the next lemma is Lemma 3.2(ii): the quantity R_h is the R of the main text, (62) is (15), and (64) is (16).

Lemma B.7 (General form of Lemma 3.2(ii))**.** *Let $I = \{a, a+1, \dots, b\} \subseteq \mathcal{T}_+$ carry the θ_h-reference directions $\mathbf{a}_1, \dots, \mathbf{a}_j$, let $R_h := 1 + \lceil 4\Lambda/\theta_h \rceil$, and let I_0 be the set of steps $t \in I$ with $S_t^{(a)} < 2\Lambda/\theta_h$. There are an integer $m \ge 0$ and nonempty sets $I_1, \dots, I_m$ of consecutive steps such that*

$$\begin{aligned} &I = I_0 \,\dot\cup\, I_1 \,\dot\cup\, \cdots \,\dot\cup\, I_m, \qquad \max I_i < \min I_{i+1} \;\; (0 \le i < m), \\ &|I_0| \le \left\lceil \frac{4\Lambda}{\theta_h} \right\rceil \le R_h, \qquad m \le R_h; \end{aligned} \tag{62}$$

the case $m = 0$ is allowed, with $I_0 = I$. For $1 \le i \le m$ put $b_i := \min I_i$ and

$$\sigma_i := \sum_{t=a}^{b_i-1} \mu_t = S_{b_i}^{(a)}, \qquad \mathbf{a}_{j+1}^{(i)} := \frac{1}{\sigma_i} \sum_{k=a}^{b_i-1} \mu_k \mathbf{v}_k = \boldsymbol{\zeta}_{b_i}. \tag{63}$$

Then $\mathbf{a}_1, \dots, \mathbf{a}_j, \mathbf{a}_{j+1}^{(i)}$ are θ_h-reference directions for I_i, and

$$\sigma_i \ge \frac{2\Lambda}{\theta_h} \;\; (1 \le i \le m), \qquad \ln \frac{\sigma_{i+1}}{\sigma_i} > \frac{\theta_h}{4} \;\; (1 \le i < m). \tag{64}$$

Moreover,

$$R_h \le \frac{6\Lambda}{\theta_h}. \tag{65}$$

Proof. Since $\lceil x \rceil < x + 1$ for every real x and $\theta_h < 1 < \Lambda$, we have $R_h < 2 + 4\Lambda/\theta_h < 6\Lambda/\theta_h$, which is (65). We also use twice that an integer smaller than $1 + x$ is at most $\lceil x \rceil$.

The prefix. Since every μ_k is positive, $t \mapsto S_t^{(a)}$ is increasing on $\{a, \dots, b+1\}$, and $S_a^{(a)} = 0$. Hence $I_0 = \{a, \dots, b_1 - 1\}$, where

$$b_1 := \min\left(\left\{ t \in \{a+1, \dots, b\} : S_t^{(a)} \ge \frac{2\Lambda}{\theta_h} \right\} \cup \{b+1\} \right). \tag{66}$$

The step $b_1 - 1$ belongs to I_0, and $\mu_k > 1/2$ for $a \le k \le b_1 - 2$, so

$$\frac{|I_0| - 1}{2} = \frac{b_1 - 1 - a}{2} \le S^{(a)}_{b_1 - 1} < \frac{2\Lambda}{\theta_h}.$$

The coefficient in this display is measured before the last step of the prefix. That step may carry the sum past the threshold, so $\sum_{t \in I_0} \mu_t$ itself need not be smaller than $2\Lambda/\theta_h$, and the argument does not use such a bound. The display gives $|I_0| < 1 + 4\Lambda/\theta_h$, hence $|I_0| \le \lceil 4\Lambda/\theta_h \rceil \le R_h$. If $b_1 = b + 1$, take $m = 0$; then $I_0 = I$, (62) holds, and the remaining assertions are void. Assume from now on that $b_1 \le b$.

The local radial bound. Let $b_1 \le t \le b$. Every signed margin at the step $a \in \mathcal{T}_+$ is nonnegative and $\mathbf{v}_t$ is a convex combination of the signed examples, so $\langle \mathbf{w}_a, \mathbf{v}_t \rangle \ge 0$. Together with $\langle \mathbf{w}_t, \mathbf{v}_t \rangle < \Lambda$ and $S^{(a)}_t \ge S^{(a)}_{b_1} \ge 2\Lambda/\theta_h$, this gives

$$\langle \boldsymbol{\zeta}_t, \mathbf{v}_t \rangle = \frac{\langle \mathbf{w}_t, \mathbf{v}_t \rangle - \langle \mathbf{w}_a, \mathbf{v}_t \rangle}{S^{(a)}_t} \le \frac{\langle \mathbf{w}_t, \mathbf{v}_t \rangle}{S^{(a)}_t} < \frac{\Lambda}{S^{(a)}_t} \le \frac{\theta_h}{2} \qquad (b_1 \le t \le b), \tag{67}$$

which is (18).

The children. Starting from b_1, define recursively

$$b_{i+1} := \min\left(\left\{ t \in \{b_i + 1, \dots, b\} : \|\boldsymbol{\zeta}_t - \boldsymbol{\zeta}_{b_i}\| > \frac{\theta_h}{2} \right\} \cup \{b + 1\} \right), \tag{68}$$

let $m := \min\{i \ge 1 : \ b_{i+1} = b + 1\}$, which is finite because $b_i < b_{i+1} \le b + 1$ as long as $b_i \le b$, and put

$$I_i := \{b_i, b_i + 1, \dots, b_{i+1} - 1\} \qquad (1 \le i \le m).$$

Since $b_i \le b$ for $1 \le i \le m$ by the minimality of m, these sets are nonempty, satisfy $\min I_i = b_i$, and partition $\{b_1, \dots, b\} = I \setminus I_0$ in chronological order, which gives the first line of (62). A crossing step $b_{i+1} \le b$ is the first step of I_{i+1} and does not belong to I_i, so no step is counted twice. As for (61), the minimality in (68) gives

$$\|\boldsymbol{\zeta}_t - \boldsymbol{\zeta}_{b_i}\| \le \frac{\theta_h}{2} \qquad (t \in I_i, \ 1 \le i \le m). \tag{69}$$

Separation and the number of children. By monotonicity and (66), $\sigma_i = S^{(a)}_{b_i} \ge S^{(a)}_{b_1} \ge 2\Lambda/\theta_h$ for every $1 \le i \le m$. For $1 \le i < m$ we have $b_{i+1} \le b$, so (68) and (57) give

$$\frac{\theta_h}{2} < \|\boldsymbol{\zeta}_{b_{i+1}} - \boldsymbol{\zeta}_{b_i}\| \le 2 \ln \frac{S^{(a)}_{b_{i+1}}}{S^{(a)}_{b_i}} = 2 \ln \frac{\sigma_{i+1}}{\sigma_i},$$

which completes (64). Summing over $1 \le i < m$, the logarithms telescope. Since $b_m \le b < U$, the budget gives $\sigma_m \le \sum_{k<U} \mu_k \le 2\eta\gamma^{-2}$, and $\sigma_1 \ge 2\Lambda/\theta_h$, so

$$\frac{(m-1)\theta_h}{4} \le \ln \frac{\sigma_m}{\sigma_1} \le \ln\left(2\eta\gamma^{-2} \cdot \frac{\theta_h}{2\Lambda} \right) = \ln \frac{\eta}{h\Lambda} < \Lambda,$$

where the equality is $\gamma^{-2}\theta_h = 1/h$ and the last inequality uses $h\Lambda > 1$. Hence $m < 1 + 4\Lambda/\theta_h$, so $m \le \lceil 4\Lambda/\theta_h \rceil \le R_h$, which completes (62).

The reference directions of a child. Fix $1 \le i \le m$. The new direction $\mathbf{a}^{(i)}_{j+1} = \boldsymbol{\zeta}_{b_i}$ of (63) lies in the convex hull of the signed examples by (57), which is condition (a) of Definition B.5. The inherited directions satisfy conditions (a) and (b) among themselves by hypothesis, and condition (c) on I_i because $I_i \subseteq I$. For $1 \le s \le j$, condition (c) on I gives $\langle \mathbf{a}_s, \mathbf{v}_k \rangle \le \theta_h$ for $a \le k < b_i$, and the weights μ_k/σ_i are positive and sum to one, so

$$\langle \mathbf{a}_s, \mathbf{a}^{(i)}_{j+1} \rangle = \frac{1}{\sigma_i} \sum_{k=a}^{b_i - 1} \mu_k \langle \mathbf{a}_s, \mathbf{v}_k \rangle \le \theta_h.$$

Thus the inherited inner-product bounds pass through the convex average exactly, and no error accumulates with the depth. For $t \in I_i$ we have $b_1 \le b_i \le t \le b$, and the Cauchy–Schwarz inequality with $\|\mathbf{v}_t\| \le 1$, (67), and (69) give

$$\langle \mathbf{a}_{j+1}^{(i)}, \mathbf{v}_t \rangle \le \langle \boldsymbol{\zeta}_t, \mathbf{v}_t \rangle + \|\boldsymbol{\zeta}_{b_i} - \boldsymbol{\zeta}_t\| < \frac{\theta_h}{2} + \frac{\theta_h}{2} = \theta_h.$$

The last two displays are (19), and they verify conditions (b) and (c) for the new direction. Hence $\mathbf{a}_1, \dots, \mathbf{a}_j, \mathbf{a}_{j+1}^{(i)}$ are θ_h-reference directions for I_i. □

All the σ_i in (64) are measured from the first step a of the parent, and Appendix B.3 uses the separation in this form.

Capacity and terminal intervals. At $h = h_*$, parts (i) and (ii) of the next lemma are Lemmas 3.3 and 3.4, which are proved in full in Section 3.2.

Lemma B.8 (General form of Lemmas 3.3 and 3.4)**.** (i) *If* $\mathbf{b}_1, \dots, \mathbf{b}_m \in \operatorname{conv}\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ *satisfy* $\langle \mathbf{b}_i, \mathbf{b}_k \rangle \le \theta_h$ *for all* $i \ne k$*, then*

$$m \le \min\left\{ r,\ \left\lfloor \frac{1-\theta_h}{\gamma^2 - \theta_h} \right\rfloor \right\} \le \nu_h + 2. \tag{70}$$

(ii) *If an interval* $I \subseteq \mathcal{T}_+$ *carries* $\nu_h + 1$ *reference directions in the sense of Definition B.5, then*

$$|I| \le \left\lceil \frac{4}{\theta_h} \right\rceil \le \frac{6}{\theta_h} = 6h\gamma^{-2}.$$

Proof. The proofs of Lemmas 3.3 and 3.4 use θ only through $0 < \theta < \gamma^2$ and, in the last step of Lemma 3.3, through the identity (30), which identifies the floor in (70) with the floor in the definition (29) of ν_h, so that, by (29), the minimum in (70) is at most $\nu_h + 2$. The facts about the trajectory used in the proof of Lemma 3.4 are the gradient identity of Lemma B.1 and the bounds $\mu_t \ge \eta L(\mathbf{w}_t)/2$ and $L(\mathbf{w}_t) > 1/\eta$ on $\mathcal{T}_+$ of Lemma B.3, none of which involves h. Both proofs therefore apply with θ_h and ν_h in place of θ and ν, the second one using part (i), and give (70) and $|I| \le \lceil 4/\theta_h \rceil$; finally, $\theta_h < 1$ gives $\lceil 4/\theta_h \rceil < 4/\theta_h + 1 < 6/\theta_h = 6h\gamma^{-2}$. □

A root interval carries one reference direction, and each application of Lemma B.7 adds one. An interval carrying $\nu_h + 1$ reference directions is called *terminal*; it is not refined further, and Lemma B.8(ii), whose proof rests on the capacity bound of Lemma B.8(i), bounds its length. Appendix B.3 organizes the resulting intervals into trees and counts them.

B.3 Counting intervals

This part proves, for a general integer $h \ge 2$, the inequality (24) for the children of an interval, the descendant count of Lemma 3.5, and the per-root count (25), which Appendix B.4 adds over the root intervals of Lemma B.6. Throughout this part, $\eta \ge 4$ and $\Lambda = \ln \eta > 10$, and we use the bound $0 \le \nu_h < 2\gamma^{-2}$ of (31).

The tree of a root interval. Let J be a nonempty set of consecutive steps in $\mathcal{T}_+$ carrying one θ_h-reference direction (Definition B.5), for instance a root interval of Lemma B.6. Recursive refinement of J produces a rooted tree. A node is an interval together with its reference directions, and the root J has depth zero. A terminal node, that is, a node carrying $\nu_h + 1$ reference directions (Appendix B.2), is not refined. A node I carrying at most ν_h reference directions is partitioned by Lemma B.7 into a prefix I_0 and children $I_1, \dots, I_m$ as in (62), and each child, together with the reference directions of I and the new one supplied by the lemma, is a node at the next depth. We write $\operatorname{succ}(I) := \{I_1, \dots, I_m\}$ for such a node and $\operatorname{succ}(I) := \varnothing$ for a terminal one. The case $m = 0$ is allowed; then $I_0 = I$, $\operatorname{succ}(I) = \varnothing$, and that branch ends. The root carries one reference direction and each child gains exactly one, so a node at depth j carries exactly $j + 1$ of them. Hence the nodes at depth ν_h are exactly the terminal ones, every node at smaller depth is refined, and when $\nu_h = 0$ the tree is the root alone.

For a node I and an integer $j \geq 0$ let $D_j(I)$ be the number of descendants of I at depth j below I. Thus $D_0(I) = 1$ and

$$D_{j+1}(I) = \sum_{I' \in \mathrm{succ}(I)} D_j(I') \qquad (j \geq 0), \tag{71}$$

where an empty sum is zero.

For a nonempty interval I of the construction define its coefficient and its logarithmic size by

$$\mu(I) := \sum_{t \in I} \mu_t, \qquad \chi_h(I) := \max\left\{0,\ \ln \frac{\theta_h \mu(I)}{2\Lambda}\right\}, \tag{72}$$

which is (21) with $\theta = \theta_h$; the normalization is the coefficient threshold $2\Lambda/\theta_h$ of Lemma B.7. Since $I \subseteq \{0, \dots, U-1\}$ and every μ_t is positive, the first bound in (52) (Lemma B.3, which requires $\eta \geq 4$) gives $\mu(I) \leq 2\eta\gamma^{-2}$, so $\theta_h \mu(I)/(2\Lambda) \leq (\gamma^2/h)(2\eta\gamma^{-2})/(2\Lambda) = \eta/(h\Lambda)$, and $h\Lambda \geq 1$ gives

$$0 \leq \chi_h(I) \leq \max\left\{0,\ \ln \frac{\eta}{h\Lambda}\right\} \leq \ln \eta = \Lambda. \tag{73}$$

For integers $k \geq 0$ let $E_k(s) := \sum_{i=0}^{k} s^i/i!$ be the truncated exponential. On $[0, \infty)$ every E_k is positive and nondecreasing, and $E_k \leq E_{k+1}$. When $\nu_h \geq 1$ put $\alpha := 2\nu_h/\Lambda$.

Counting siblings. By (64), the children of a node start at coefficients whose logarithms, measured on the coefficient axis of the parent, are more than $\theta_h/4$ apart (Figure 1(b)). Counting the siblings jointly along this axis gives the next lemma, which is (24) with $\theta = \theta_h$ and $\nu = \nu_h$.

Lemma B.9 (Bound on the children of an interval)**.** *Let $\nu_h \geq 1$, and let $I = \{a, \dots, b\}$ be a nonterminal node with children $I_1, \dots, I_m$ given by Lemma B.7. Then, for every integer $k \geq 0$,*

$$\sum_{i=1}^{m} E_k\big(\alpha \chi_h(I_i)\big) \leq \frac{6}{\alpha\theta_h}\, E_{k+1}\big(\alpha \chi_h(I)\big). \tag{74}$$

Proof. If $m = 0$, the left side is zero and the right side is positive. Assume $m \geq 1$, and let $b_i := \min I_i$ and $\sigma_i := \sum_{t=a}^{b_i - 1} \mu_t$ for $1 \leq i \leq m$, as in Lemma B.7.

By (62), the sets $I_0, I_1, \dots, I_m$ are consecutive pieces of the consecutive set I with $\max I_i < \min I_{i+1}$, so $I_i = \{b_i, \dots, b_{i+1} - 1\}$ for $1 \leq i < m$. Hence a completed child satisfies

$$\mu(I_i) = \sigma_{i+1} - \sigma_i \leq \sigma_{i+1} \qquad (1 \leq i < m),$$

while the last child satisfies only $\mu(I_m) \leq \mu(I)$, because $I_m \subseteq I$ and every μ_t is positive. Define

$$u_j := \ln \frac{\theta_h \sigma_j}{2\Lambda} \qquad (2 \leq j \leq m).$$

The first part of (64) gives $u_j \geq 0$, and the second part gives $u_{j+1} - u_j = \ln(\sigma_{j+1}/\sigma_j) > \theta_h/4$ for $2 \leq j < m$. Since $\{a, \dots, b_j - 1\} \subseteq I$, we have $\sigma_j \leq \mu(I)$, and therefore $u_j \leq \chi_h(I)$. By (72), χ_h is a nondecreasing function of the coefficient, so the two bounds on $\mu(I_i)$ give

$$\chi_h(I_i) \leq \max\{0, u_{i+1}\} = u_{i+1} \quad (1 \leq i < m), \qquad \chi_h(I_m) \leq \chi_h(I).$$

Let $f\colon [0, \infty) \to [0, \infty)$ be nondecreasing, hence integrable on bounded intervals. The intervals $[u_j, u_j + \theta_h/4]$, $2 \leq j \leq m$, are pairwise disjoint, because consecutive left endpoints differ by more than $\theta_h/4$, and they are contained in $[0, \chi_h(I) + \theta_h/4]$. Hence the bounds on $\chi_h(I_i)$, the monotonicity of f, and $(\theta_h/4) f(u_j) \leq \int_{u_j}^{u_j + \theta_h/4} f(u)\, du$ give

$$\sum_{i=1}^{m} f\big(\chi_h(I_i)\big) \leq f\big(\chi_h(I)\big) + \sum_{j=2}^{m} f(u_j) \leq f\big(\chi_h(I)\big) + \frac{4}{\theta_h} \int_0^{\chi_h(I) + \theta_h/4} f(u)\, du, \tag{75}$$

which is (23) with $\theta = \theta_h$.

Take $f(u) := E_k(\alpha u)$, which is nonnegative and nondecreasing on $[0,\infty)$. Since $E'_{k+1} = E_k$ and $E_{k+1}(0) = 1$,

$$\int_0^v E_k(\alpha u)\,du = \frac{E_{k+1}(\alpha v) - 1}{\alpha} \qquad (v \ge 0).$$

Moreover $E'_{k+1} = E_k \le E_{k+1}$ on $[0,\infty)$, so $\frac{d}{ds}\big(e^{-s}E_{k+1}(s)\big) = e^{-s}\big(E_k(s) - E_{k+1}(s)\big) \le 0$ and $s \mapsto e^{-s}E_{k+1}(s)$ is nonincreasing there, that is,

$$E_{k+1}(u+v) \le e^v E_{k+1}(u) \qquad (u, v \ge 0). \tag{76}$$

We evaluate the integral at the upper limit $\chi_h(I) + \theta_h/4$, substitute it into (75), drop the term $-4/(\alpha\theta_h) < 0$, and apply $E_k \le E_{k+1}$ and (76) with $u = \alpha\chi_h(I)$ and $v = \alpha\theta_h/4$. This gives

$$\begin{aligned}\sum_{i=1}^m E_k\big(\alpha\chi_h(I_i)\big) &\le E_k\big(\alpha\chi_h(I)\big) + \frac{4}{\alpha\theta_h}E_{k+1}\left(\alpha\chi_h(I) + \frac{\alpha\theta_h}{4}\right)\\ &\le \frac{\alpha\theta_h + 4e^{\alpha\theta_h/4}}{\alpha\theta_h}\,E_{k+1}\big(\alpha\chi_h(I)\big).\end{aligned}$$

By $\nu_h < 2\gamma^{-2}$, $h \ge 2$, and $\Lambda > 10$, we have $0 < \alpha\theta_h = 2\nu_h\gamma^2/(h\Lambda) < 4/(h\Lambda) < 1/5$. Since $e^s \le (1-s)^{-1}$ for $0 \le s < 1$, we get $e^{\alpha\theta_h/4} < (1 - 1/20)^{-1} < 5/4$, hence $\alpha\theta_h + 4e^{\alpha\theta_h/4} < 1/5 + 5 < 6$, and (74) follows. □

Counting descendants. Iterating Lemma B.9 down the tree gives the general form of Lemma 3.5, which is the case $h = h_*$.

Lemma B.10 (General form of Lemma 3.5)**.** *Let $\nu_h \ge 1$. For every node I of the construction and every $0 \le j \le \nu_h$,*

$$D_j(I) \le 2\,\frac{(6\Lambda/\theta_h)^j}{j!} = 2\,\frac{(6h\gamma^{-2}\Lambda)^j}{j!}. \tag{77}$$

Proof. We first prove the weighted estimate

$$D_j(I) \le \left(\frac{6}{\alpha\theta_h}\right)^j E_j\big(\alpha\chi_h(I)\big) \tag{78}$$

by induction on $j \ge 0$, simultaneously for all nodes I of the construction. For $j = 0$ both sides equal one. Assume that (78) holds at j for every node, and fix a node I. If $\mathrm{succ}(I) = \varnothing$, then $D_{j+1}(I) = 0$ by (71), and the right side of (78) at $j+1$ is positive. Otherwise I is nonterminal with children $I_1, \dots, I_m$, and (71), the induction hypothesis applied to every child, and Lemma B.9 with $k = j$ give

$$D_{j+1}(I) = \sum_{i=1}^m D_j(I_i) \le \left(\frac{6}{\alpha\theta_h}\right)^j \sum_{i=1}^m E_j\big(\alpha\chi_h(I_i)\big) \le \left(\frac{6}{\alpha\theta_h}\right)^{j+1} E_{j+1}\big(\alpha\chi_h(I)\big).$$

This proves (78) for every $j \ge 0$.

Now let $0 \le j \le \nu_h$. By (73) and the monotonicity of E_j, we have $E_j(\alpha\chi_h(I)) \le E_j(\alpha\Lambda)$, and $\alpha\Lambda = 2\nu_h \ge 2j$. The terms $a_k := (\alpha\Lambda)^k/k!$ of $E_j(\alpha\Lambda)$ satisfy $a_{k-1}/a_k = k/(\alpha\Lambda) \le j/(2\nu_h) \le 1/2$ for $1 \le k \le j$, so $a_{j-i} \le 2^{-i}a_j$ for $0 \le i \le j$ and

$$E_j(\alpha\Lambda) = \sum_{i=0}^j a_{j-i} \le a_j \sum_{i=0}^j 2^{-i} < 2\,\frac{(\alpha\Lambda)^j}{j!}.$$

Substituting into (78), the powers of α cancel, which gives $D_j(I) \le 2(6\Lambda/\theta_h)^j/j!$, and $1/\theta_h = h\gamma^{-2}$ gives the second form in (77). □

Counting the steps of a root interval. Combining Lemma B.10 with the lengths of prefixes and terminal nodes gives the general form of (25), which is the case $h = h_*$ of the following bound.

Proposition B.11 (Number of steps of a root interval)**.** *Let J be a nonempty set of consecutive steps in $\mathcal{T}_+$ carrying one θ_h-reference direction. Then*

$$|J| \le \left(3\nu_h + 12h\gamma^{-2}\right) \frac{(6h\gamma^{-2}\Lambda)^{\nu_h}}{\nu_h!}. \tag{79}$$

Proof. If $\nu_h = 0$, then J is terminal, and Lemma B.8(ii) gives $|J| \le \lceil 4/\theta_h \rceil \le 6/\theta_h = 6h\gamma^{-2}$, which is smaller than the right side $12h\gamma^{-2}$ of (79).

Let $\nu_h \ge 1$, consider the tree of J constructed above, and put $z := 6\Lambda/\theta_h = 6h\gamma^{-2}\Lambda$. For every $0 \le j \le \nu_h$, the set J is the disjoint union of the prefixes of the nodes at depths below j and of the nodes at depth j. This is trivial for $j = 0$, and the step from $j < \nu_h$ to $j+1$ replaces every node I at depth j, which is nonterminal, by its partition (62) into I_0 and the nodes of $\mathrm{succ}(I)$; a node without children contributes only its prefix $I_0 = I$. At $j = \nu_h$ the nodes at depth j are the terminal ones. There are $D_j(J)$ nodes at depth j, every prefix has at most R_h steps by (62), and every terminal node has at most $\lceil 4/\theta_h \rceil$ steps by Lemma B.8(ii). Therefore

$$|J| \le R_h \sum_{j=0}^{\nu_h - 1} D_j(J) + \left\lceil \frac{4}{\theta_h} \right\rceil D_{\nu_h}(J). \tag{80}$$

By $\nu_h < 2\gamma^{-2}$, $h \ge 2$, and $\Lambda > 10$, we have $z/\nu_h > 3h\Lambda > 18$, so each term of $\sum_{j<\nu_h} z^j/j!$ is less than $1/18$ of the next, and the sum is at most $(18/17)\, z^{\nu_h - 1}/(\nu_h - 1)!$. Lemma B.10 therefore gives

$$\sum_{j=0}^{\nu_h - 1} D_j(J) \le 2 \sum_{j=0}^{\nu_h - 1} \frac{z^j}{j!} < 3\, \frac{z^{\nu_h - 1}}{(\nu_h - 1)!}, \qquad D_{\nu_h}(J) \le 2\, \frac{z^{\nu_h}}{\nu_h!}.$$

Finally, $R_h \le 6\Lambda/\theta_h = z$ by (65) and $\lceil 4/\theta_h \rceil \le 6/\theta_h$ by Lemma B.8(ii). Substituting these four bounds into (80) and using $z \cdot z^{\nu_h - 1}/(\nu_h - 1)! = \nu_h z^{\nu_h}/\nu_h!$ gives

$$|J| \le 3\nu_h \frac{z^{\nu_h}}{\nu_h!} + \frac{12}{\theta_h} \frac{z^{\nu_h}}{\nu_h!} = \left(3\nu_h + 12h\gamma^{-2}\right) \frac{(6h\gamma^{-2}\Lambda)^{\nu_h}}{\nu_h!},$$

which is (79). □

B.4 Stepsize conditions and completion of the proof

This part proves that every $\eta \ge \eta_h^0$ satisfies the stepsize conditions (26) for general h (Lemma B.12), completes the proof of Theorem A.1 and with it that of the low-margin case of Theorem 2.2(i), and derives the factorial bound that is available before the absorption (Proposition B.13). Throughout, Assumption 2.1 holds, the integer $h \ge 2$ is fixed, $\Lambda = \ln \eta$, and the only condition on the stepsize is $\eta \ge \eta_h^0$. The restriction $0 < \gamma < 1/\sqrt{2}$ of Theorem A.1 is used only in the proofs of the theorem and of Proposition B.13, where it gives $\gamma^{-2} > 2$; Lemma B.12, on which Appendices B.1–B.3 rely for their standing assumptions, uses only $\gamma \le 1$ from Assumption 2.1.

Two facts about the constants (29)–(34) are used several times. First, $\gamma \le 1$, $h \ge 2$, and $n \ge 1$ give

$$C_h \ge 80\gamma^{-4}hn \ge 160n\gamma^{-4} > 4n\gamma^{-2} \ge 4\gamma^{-2} > 1. \tag{81}$$

Second, let $\nu_h \ge 1$ and put $b_h := \max\{2, \nu_h\}$. The quotient inside the ceiling of (32) is positive and $p_h \ge \nu_h + 1$, as observed after that display, and a ceiling is at least its argument, so

$$p_h - \nu_h \ge \frac{\nu_h \ln(18h\gamma^{-2}/\nu_h)}{\ln b_h} > 0 \qquad (\nu_h \ge 1). \tag{82}$$

Since $p_h = 0$ when $\nu_h = 0$, we have $p_h \ge 0$ in every case.

Lemma B.12 (Stepsize conditions)**.** *Every $\eta \ge \eta_h^0$ satisfies*

$$\eta \ge 4, \qquad \Lambda > 10, \qquad 4h\gamma^{-2}\Lambda \le \eta, \qquad C_h \Lambda^{p_h} < \widehat{T}. \tag{83}$$

The third condition is the root condition $4\Lambda \le \eta\theta_h$, and at $h = h_*$ the four conditions are (26).

Proof. Write $K := K_h$, $\eta_* := \eta_h^0$, and $s := 1 + \ln K$. The terms $h\gamma^{-2}$, p_h, and $\ln(2\gamma^{-2}C_h)$ of K in (34) are nonnegative, the last one because $2\gamma^{-2}C_h > 1$ by (81). Hence

$$K \geq \exp(4), \qquad s \geq 5, \qquad h\gamma^{-2} \leq K, \qquad p_h \leq K, \qquad \ln(2\gamma^{-2}C_h) \leq K.$$

The ceiling in (34) adds less than one and $Ks \geq 1$, so

$$128Ks \leq \eta_* \leq 128Ks + 1 \leq 129Ks.$$

In particular $\eta_* \geq 640\exp(4) > \exp(10)$, because $\exp(6) < 404 < 640$, and every $\eta \geq \eta_*$ satisfies $\eta \geq 4$ and $\Lambda > 10$. The upper bound gives

$$\ln \eta_* \leq \ln 129 + (s-1) + \ln s \leq 3s,$$

where the last step uses $\ln 129 < 5 \leq s$ and $\ln s \leq s$. The function $s \mapsto s - \ln(3s)$ has derivative $1 - 1/s > 0$ for $s \geq 5$ and the positive value $5 - \ln 15$ at $s = 5$, so $\ln(3s) \leq s$. Therefore

$$10 < \ln \eta_* \leq 3s, \qquad 0 < \ln\ln \eta_* \leq s. \tag{84}$$

The root condition. By $h\gamma^{-2} \leq K$ and (84),

$$4h\gamma^{-2} \ln \eta_* \leq 12Ks < 128Ks \leq \eta_*.$$

The function $\eta \mapsto \eta/\ln\eta$ has derivative $(\ln\eta - 1)/(\ln\eta)^2 > 0$ for $\eta > e$, so $\eta/\Lambda \geq \eta_*/\ln\eta_* \geq 4h\gamma^{-2}$ for every $\eta \geq \eta_*$. This is the third condition, and dividing it by $h\gamma^{-2} = 1/\theta_h$ gives $4\Lambda \leq \eta\theta_h$.

The horizon condition. For $\eta \geq \eta_*$ define

$$\Phi(\eta) := \frac{\eta}{4} - \ln\big(2\gamma^{-2}\eta\, C_h(\ln\eta)^{p_h}\big).$$

At $\eta = \eta_*$, (84) together with $0 \leq p_h \leq K$ and $\ln(2\gamma^{-2}C_h) \leq K$ gives

$$\begin{aligned}\ln\big(2\gamma^{-2}\eta_* C_h(\ln\eta_*)^{p_h}\big) &= \ln(2\gamma^{-2}C_h) + \ln\eta_* + p_h \ln\ln\eta_* \\ &\leq K + 3s + Ks \leq 3Ks < 32Ks \leq \frac{\eta_*}{4},\end{aligned}$$

where $K + 3s + Ks \leq 3Ks$ because $K \leq Ks$ and $3s \leq Ks$, by $s \geq 1$ and $K > 3$, and the last step is $\eta_* \geq 128Ks$. Thus $\Phi(\eta_*) > 0$. Moreover,

$$\Phi'(\eta) = \frac{1}{4} - \frac{1}{\eta} - \frac{p_h}{\eta\ln\eta} > \frac{1}{4} - \frac{1}{128} - \frac{1}{640} > 0 \qquad (\eta \geq \eta_*),$$

because $\eta > \exp(10) > 128$, $\eta \geq 128Ks \geq 640K \geq 640p_h$, and $\ln\eta > 1$. Hence $\Phi(\eta) > 0$ for every $\eta \geq \eta_*$, and exponentiating gives

$$\frac{\gamma^2\exp(\eta/4)}{\eta} > 2C_h\Lambda^{p_h}.$$

Finally, $C_h\Lambda^{p_h} \geq 1$ because $C_h > 1$, $\Lambda > 10$, and $p_h \geq 0$, so the definition (11) and the elementary bound $\lfloor u \rfloor > u - 1$ give the strict inequality

$$\widehat{T} > \frac{\gamma^2\exp(\eta/4)}{\eta} - 1 > 2C_h\Lambda^{p_h} - 1 \geq C_h\Lambda^{p_h}. \qquad \square$$

We now combine the counts of the preceding parts.

Proof of Theorem A.1. Fix $\eta \geq \eta_h^0$, and let $0 < \gamma < 1/\sqrt{2}$ as in the theorem, so that $\gamma^{-2} > 2$. By Lemma B.12, the hypotheses $\eta \geq 4$, $\Lambda > 10$, and $4\Lambda \leq \eta\theta_h$ of the results cited below hold; the fourth condition $C_h\Lambda^{p_h} < \widehat{T}$ is used only in the last step, where we pass from $U = \min\{\tau, \widehat{T}\}$ to τ. Each of the steps $0, 1, \ldots, U-1$ is step zero, a step of $\mathcal{T}_-$, or a step of $\mathcal{T}_+ \setminus \{0\}$, so $U \leq 1 + |\mathcal{T}_-| + |\mathcal{T}_+ \setminus \{0\}|$, and we bound $1 + |\mathcal{T}_-|$ and $|\mathcal{T}_+ \setminus \{0\}|$ separately.

Step zero and $\mathcal{T}_-$. Lemma B.3 gives $|\mathcal{T}_-| < 4n\gamma^{-2}$. Since $n\gamma^{-2} > 1$ and $h\gamma^{-2} > 1$,

$$1 + |\mathcal{T}_-| < 1 + 4n\gamma^{-2} < 5n\gamma^{-2} \leq 5\gamma^{-4}h\big(n + h\ln(4\gamma^{-2})\big). \tag{85}$$

Root intervals. By the root condition, Lemma B.6 partitions $\mathcal{T}_+ \setminus \{0\}$ into M consecutive root intervals, each carrying one reference direction. By $|\mathcal{T}_-| < 4n\gamma^{-2}$, $1/\theta_h = h\gamma^{-2}$, and $\lceil u \rceil < u+1$, their number satisfies

$$M \le |\mathcal{T}_-| + 1 + \lceil 4h\gamma^{-2}\ln(4\gamma^{-2})\rceil < 4\gamma^{-2}\big(n + h\ln(4\gamma^{-2})\big) + 2 \le 5\gamma^{-2}\big(n + h\ln(4\gamma^{-2})\big).$$

The last step holds because $\ln(4\gamma^{-2}) > \ln 8 > 2$ gives $n + h\ln(4\gamma^{-2}) > 5$ and therefore $\gamma^{-2}(n + h\ln(4\gamma^{-2})) > 10 > 2$. Each root interval J is refined recursively by Lemma B.7 into the tree of Appendix B.3, and Proposition B.11 gives

$$|J| \le \big(3\nu_h + 12h\gamma^{-2}\big)F_h, \qquad F_h := \frac{(6h\gamma^{-2}\Lambda)^{\nu_h}}{\nu_h!}.$$

Since $3\nu_h < 6\gamma^{-2} \le 3h\gamma^{-2}$ by (31) and $h \ge 2$, the first factor is at most $15h\gamma^{-2}$. Multiplying by the bound on M,

$$|\mathcal{T}_+ \setminus \{0\}| \le 75\gamma^{-4}h\big(n + h\ln(4\gamma^{-2})\big)F_h. \tag{86}$$

The structural count. The factor F_h is at least one. This is an equality when $\nu_h = 0$. When $\nu_h \ge 1$, we have $6h\gamma^{-2}\Lambda > 2\gamma^{-2} > \nu_h$ by $6h\Lambda > 2$ and (31), and $\nu_h! \le \nu_h^{\nu_h}$, so $F_h \ge (6h\gamma^{-2}\Lambda/\nu_h)^{\nu_h} > 1$. The right side of (85) is therefore at most $5\gamma^{-4}h(n + h\ln(4\gamma^{-2}))F_h$, and adding (86) gives, by $U \le 1 + |\mathcal{T}_-| + |\mathcal{T}_+ \setminus \{0\}|$ and (33),

$$U \le (5+75)\gamma^{-4}h\big(n + h\ln(4\gamma^{-2})\big)F_h = C_h\Lambda^{\nu_h}\,\frac{(6h\gamma^{-2})^{\nu_h}}{\nu_h!}. \tag{87}$$

At $h = h_*$ this is (27).

Absorbing the fixed coefficient. If $\nu_h = 0$, then $p_h = 0$ and (87) reads $U \le C_h = C_h\Lambda^{p_h}$. Assume from now on that $\nu_h \ge 1$, and recall $b_h = \max\{2, \nu_h\}$. One term of the exponential series gives $\exp(\nu_h) \ge \nu_h^{\nu_h}/\nu_h!$, that is, $\nu_h! \ge (\nu_h/e)^{\nu_h}$, and $e < 3$ gives

$$\frac{(6h\gamma^{-2})^{\nu_h}}{\nu_h!} \le \left(\frac{6eh\gamma^{-2}}{\nu_h}\right)^{\nu_h} \le \left(\frac{18h\gamma^{-2}}{\nu_h}\right)^{\nu_h}. \tag{88}$$

Multiplying the first inequality in (82) by $\ln b_h \ge \ln 2 > 0$ and exponentiating,

$$b_h^{p_h - \nu_h} \ge \left(\frac{18h\gamma^{-2}}{\nu_h}\right)^{\nu_h}. \tag{89}$$

At $h = h_*$, (88) and (89) are (28). Converting the power of b_h into a power of Λ requires two ranges of Λ. To keep the stepsize threshold polynomial, the structural count is used when Λ is at least the depth, and the coefficient budget of Lemma B.3 otherwise.

Case 1: $\Lambda \ge b_h$. By (87), (88), (89), and $p_h - \nu_h \ge 0$,

$$U \le C_h\Lambda^{\nu_h}\left(\frac{18h\gamma^{-2}}{\nu_h}\right)^{\nu_h} \le C_h\Lambda^{\nu_h}b_h^{p_h-\nu_h} \le C_h\Lambda^{\nu_h}\Lambda^{p_h-\nu_h} = C_h\Lambda^{p_h}.$$

Case 2: $10 < \Lambda < b_h$. Then $b_h > 2$, so $b_h = \nu_h$ and $\Lambda < \nu_h$. Here the structural count is replaced by a crude count. Lemma B.3 gives $\mu_t > 1/2$ at every $t \in \mathcal{T}_+$ and, in (52), $\sum_{t<U}\mu_t \le 2\eta\gamma^{-2}$. Since every μ_t is positive,

$$\frac{1}{2}|\mathcal{T}_+| \le \sum_{t\in\mathcal{T}_+}\mu_t \le \sum_{t<U}\mu_t \le 2\eta\gamma^{-2},$$

so $|\mathcal{T}_+| \le 4\gamma^{-2}\exp(\Lambda) \le C_h\exp(\Lambda)$ by (81), and $|\mathcal{T}_-| < 4n\gamma^{-2} \le C_h$ by (52) and (81). Since $\{0, 1, \dots, U-1\} = \mathcal{T}_+ \dot\cup \mathcal{T}_-$,

$$U = |\mathcal{T}_+| + |\mathcal{T}_-| \le C_h\big(1 + \exp(\Lambda)\big). \tag{90}$$

Since $\Lambda > e$, we have $\ln\Lambda > 1$ and hence $\exp(\Lambda) \le \exp(\Lambda\ln\Lambda) = \Lambda^\Lambda \le \Lambda^{\nu_h}$, the last step by $1 < \Lambda < \nu_h$. Using $\Lambda^{\nu_h} \ge 1$, $\Lambda > 2$, and $p_h \ge \nu_h + 1$, which was observed after (32), (90) gives

$$U \le C_h\big(1 + \Lambda^{\nu_h}\big) \le 2C_h\Lambda^{\nu_h} < C_h\Lambda^{\nu_h+1} \le C_h\Lambda^{p_h}.$$

Thus $U \le C_h \Lambda^{p_h}$ in every case, including $\nu_h = 0$. Because of the second case, the threshold η_h^0 never needs $\Lambda \ge \nu_h$, which would be an exponential requirement on the stepsize.

Removing the horizon. The fourth condition in (83) now gives the strict comparison

$$U \le C_h \Lambda^{p_h} < \widehat{T}.$$

Hence $U \ne \widehat{T}$, and $U = \min\{\tau, \widehat{T}\}$ forces $U = \tau$. Therefore $\tau = U \le C_h \Lambda^{p_h} = C_h (\ln \eta)^{p_h}$, which is the bound of Theorem A.1. □

By (36), Theorem A.1 at $h = h_* \ge 4$ is the bound (7) of Theorem 2.2(i) for $0 < \gamma < 1/\sqrt{2}$.

The structural count (87) is itself a bound on the phase transition time, in which the factorial saving of Appendix B.3 is still visible. It coincides with the bound of Theorem A.1 when $\nu_h = 0$, and by (88) and (89) it is at least as sharp when $\nu_h \ge 1$ and $\Lambda \ge \max\{2, \nu_h\}$.

Proposition B.13 (Factorial bound on the phase transition time)**.** *Suppose Assumption 2.1 holds with $0 < \gamma < 1/\sqrt{2}$. For every integer $h \ge 2$ and every $\eta \ge \eta_h^0$,*

$$\tau \le C_h \, \frac{(6h\gamma^{-2} \ln \eta)^{\nu_h}}{\nu_h!}. \tag{91}$$

Proof. The proof of Theorem A.1 established $U = \tau$, and substituting this identity into (87) gives (91). □