TIGHT TRANSITION TIME BOUNDS FOR SEPARABLE LOGISTIC REGRESSION AT THE EDGE OF STABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study logistic regression on linearly separable data under gradient descent with a large constant stepsize η. Such dynamics may exhibit a characteristic Edge of Stability phenomenon, in which the loss initially oscillates before transitioning to a stable phase of monotone decrease. Existing work provides a tight $\Theta(1)$ bound in dimension $d = 2$ as $\eta \to \infty$ and conjectures a bound independent of η in arbitrary dimensions $d \geq 2$. In this paper, we disprove this conjecture by showing that, for every fixed sample size $n \geq 2$ and sufficiently small margin γ, the worst-case transition time is

$$\Theta\Big((\log \eta)^{\min\{n-2,d-2\}}\Big)$$

uniformly over $d \geq 2$. The key challenge in establishing a tight bound is that the sample contributing most strongly to the gradient can change repeatedly across iterations. To address this issue, we control such changes by induction on dimension and sample size, and construct matching hard instances.

1 INTRODUCTION

Classical guarantees for gradient descent (GD) typically require small stepsizes: for a β-smooth objective, a stepsize $0 < \eta < 2/\beta$ guarantees monotone loss decrease (Nesterov, 2018). However, full-batch neural network training often operates in *Edge of Stability* (EoS) regimes, where the largest Hessian eigenvalue hovers near $2/\eta$ and the loss oscillates over short time scales while decreasing over longer time scales (Cohen et al., 2021; Cohen et al., 2025). Logistic regression on linearly separable data under GD with a large stepsize offers a tractable setting for studying such EoS regimes (Wu et al., 2023; Wu et al., 2024; Crawshaw & Liu, 2026).

In this regime, GD with a large constant stepsize can exhibit two phases: an initial EoS phase with oscillatory loss, followed by a stable phase of nonincreasing loss (Wu et al., 2024). To quantify when the transition between these two phases occurs, we define the *exact EoS transition time* $\sigma_\eta(D)$ as the earliest iterate after which the loss remains nonincreasing on dataset D with stepsize η. Understanding its dependence on the stepsize η is crucial for the optimization of GD within a finite time horizon: A larger stepsize η accelerates convergence in the stable phase where loss decreases monotonically with time. However, if increasing η also extends the EoS phase, fewer iterations remain for stable descent within a fixed budget, potentially reducing the acceleration in the stable phase. Thus, sharp control of the transition time helps determine how large a constant stepsize can be used effectively within a fixed iteration budget.

The sharp dependence of the transition time on the stepsize has remained open. Wu et al. (2024) show that GD reaches a sufficient condition for permanent monotonicity within $O(\eta)$ iterations, leaving open whether a linear dependence on the stepsize is necessary in the worst case. Crawshaw & Liu (2026) sharpen this bound in data dimension $d = 2$: the first time the loss reaches a small threshold guaranteeing subsequent monotonicity is bounded independently of η for sufficiently large η. Based on numerical experiments, they conjecture that such a uniform bound independent of stepsize holds in arbitrary dimension.

In this paper, we disprove[1] this conjecture by establishing a tight worst-case dependence on η for arbitrary d. After absorbing the labels into the features, let $\mathcal{D}_{n,d}(\gamma)$ denote datasets of n feature

[1] Crawshaw & Liu (2026) study a threshold-based transition time, whereas we define the exact transition time. We show below that these two notions have the same worst-case asymptotic dependence on η.

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.

vectors $x_i \in \mathbb{R}^d$ with $\|x_i\| \le 1$ for which there exists a unit separator with margin at least γ. We consider GD from $w_0 = 0$ on the empirical logistic loss F. For every fixed $n \ge 2$, there exists $\gamma_n \in (0, 1]$ such that, for every fixed $0 < \gamma \le \gamma_n$, Theorem 1 gives

$$\sup_{D \in \mathcal{D}_{n,d}(\gamma)} \sigma_\eta(D) = \Theta\Big((\log \eta)^{\min\{n-2, d-2\}}\Big) \qquad \text{as } \eta \to \infty,$$

uniformly over $d \ge 2$. Both the constants and the stepsize threshold depend only on n, γ and are independent of dimension d, while the worst-case dataset may depend on η. Thus the worst-case transition time is bounded independently of η when $n = 2$ or $d = 2$, but grows polylogarithmically once both $n, d \ge 3$. Each increase in $\min\{n, d\}$ permits one additional logarithmic factor.

The above main result has two corollaries. **First, it disproves the conjecture of Crawshaw & Liu (2026)** that, for fixed sample size and margin, GD reaches the loss threshold $1/(8\eta)$ within a number of iterations bounded independently of η in arbitrary dimension. To state this corollary, let $k = \min\{n - 2, d - 2\}$ and define

$$\tau_\eta(D) = \inf\left\{ t \in \mathbb{Z}_{\ge 0} : F(w_t) \le \frac{1}{8\eta} \right\}.$$

Under the same assumptions, Corollary 2 gives $\sup_{D \in \mathcal{D}_{n,d}(\gamma)} \tau_\eta(D) = \Theta((\log \eta)^k)$, as $\eta \to \infty$, uniformly over $d \ge 2$. This corollary establishes the same worst-case order for the threshold transition time $\tau_\eta(D)$ as our main theorem does for the exact EoS transition time $\sigma_\eta(D)$. The corollary also indicates that transition times $\sigma_\eta(D)$ and $\tau_\eta(D)$ may differ on an individual dataset, but their worst-case orders agree. In summary, the stepsize independence established by Crawshaw & Liu (2026) is specific to the regime $d = 2$ rather than a dimension-free phenomenon.

Second, this transition time bound yields a stronger guarantee for the optimization within a finite time horizon. We recall that the previous $O(\eta)$ transition bound, together with the $\widetilde{O}(1/(\eta t))$ loss bound after t further iterations in the stable phase (Wu et al., 2024) allows a stepsize $\eta = \Theta(T)$ and gives final loss $\widetilde{O}(T^{-2})$ after T iterations. In contrast, our polylogarithmic transition bound allows a much larger constant stepsize. For $n, d \ge 3$ and every fixed $0 < \gamma \le 1$, choosing $\eta = \exp(aT^{1/k})$ for a prescribed horizon T, with $a > 0$ sufficiently small, ensures that $\tau_\eta(D) \le T/2$ for all sufficiently large T. Recall that $k = \min\{n - 2, d - 2\}$, then applying the loss bound in the stable phase over the remaining iterations gives the bound on the last-iterate loss $F(w_T)$ as

$$F(w_T) \le \frac{C}{T} \exp\Big(-aT^{1/k}\Big),$$

where $a, C > 0$ and the required lower bound on T depend only on n, γ (Corollary 3). In contrast, Wu et al. (2024) obtain an $\widetilde{O}(T^{-2})$ bound on $F(w_T)$ in general dimension by choosing $\eta = \Theta(T)$.

1.1 Technical Overview

Proving a sharp transition bound requires going beyond two existing approaches. Wu et al. (2024) obtain an $O(\eta)$ transition bound by combining progress along a separating direction with a global bound on iterate growth. This establishes eventual stability but leaves the sharp stepsize dependence unresolved. Crawshaw & Liu (2026) sharpen the analysis in $d = 2$ by tracking oscillations in the one-dimensional subspace orthogonal to the max-margin direction. Controlling these oscillations yields a stepsize-independent bound. In higher dimensions, this orthogonal subspace has dimension $d - 1 > 1$, so the same one-dimensional oscillation argument does not apply directly.

The key challenge in higher dimensions is that the samples contributing most strongly to the gradient can change repeatedly: a sample with a negligible current contribution can later become significant again, and this can happen repeatedly throughout the training trajectory. Our analysis handles these changes through recursive projections that reduce the rank of the projected sample vectors. Each recursive rank reduction contributes at most one logarithmic factor, while the lower bound constructs instances in which all such factors occur.

Upper bound. The bound follows by controlling how long the dynamics can be driven by a given subset of samples and how often samples outside this subset can begin contributing significantly to the updates. The $O(\eta)$ analysis of Wu et al. (2024) relies on a global progress argument and does

not exploit the finer structure of the oscillatory trajectory, while Crawshaw & Liu (2026) exploit this structure in $d = 2$, showing that oscillations force faster progress. Our key insight is that *after an appropriate projection, one can retain a nonempty subset of samples that remains separable while its rank decreases by at least one*, allowing us to bound the resulting dynamics recursively by induction. To bound the full trajectory before $\tau_\eta(D)$, we also control how often this reduction must be restarted. We restart the reduction whenever a large update occurs or a sample outside the retained subset becomes significant again, and show that only $O(1)$ such restarts are needed.

It therefore remains to bound the number of iterations between consecutive restarts. We call the consecutive iterations between restarts an *outer block*. Within each outer block, we recursively reduce the rank of the projected sample vectors, that is, the dimension of the subspace they span. At each recursive level, we partition the projected trajectory into intervals such that each completed interval increases the component along a separating direction by a fixed multiplicative factor. Since this component is globally bounded by $O(\eta)$, its multiplicative growth permits at most $O(\log \eta)$ intervals at each level. To bound the duration of each interval, we project again and retain a smaller subset of samples that remains separable and has lower rank. The induction then bounds the number of iterations in that interval. Summing these bounds over $O(\log \eta)$ intervals introduces at most one additional logarithmic factor.

The recursion ends when the projected sample vectors have rank one, where the duration is $O(1)$. For original data of rank $r \geq 2$, the first projection has already reduced the rank to at most $r - 1$, leaving at most $r-2$ recursive rank reductions. Each contributes at most one logarithmic factor, and summing over the $O(1)$ outer blocks does not change the order. Together with the direct bound for original rank one, this gives at most $\max\{r - 2, 0\}$ logarithmic factors.

Lower bound. The key insight for the construction of hard instances is that *a sample can contribute extremely little to the current update and still affect the transition time.* When this contribution is extremely small, the updates are driven almost entirely by the remaining samples, while those updates continue to change the sample's margin. Our construction makes the margin move only a constant fraction of the remaining distance before the sample becomes influential, allowing $\Theta(\log \eta)$ repetitions. Each additional interacting sample therefore contributes one additional factor of $\log \eta$ to the total trajectory length. This mechanism makes the transition time grow as a power of $\log \eta$, even for a fixed sample size and a fixed positive margin lower bound.

More precisely, the exponent of $\log \eta$ is determined by the sample size and dimension. For $k \geq 1$, producing k logarithmic factors uses $k + 1$ interacting samples in $k + 1$ dimensions. The recursive construction starts from a chosen initial state, which we realize after the first GD update from $w_0 = 0$ by adding one setup sample in an extra direction. Thus, for $n, d \geq 3$, we can take $k = \min\{n - 2, d - 2\}$. However, a long trajectory alone does not establish a lower bound on the exact transition time, since the loss could already be decreasing monotonically. We therefore arrange a strict loss increase at time $\Theta((\log \eta)^k)$. This forces $\sigma_\eta(D)$ to occur later, giving the matching lower bound for the exact transition time.

2 Related Work

Comparison with previous works on the transition to stable phase. The works most closely related to ours study large-stepsize GD on separable logistic regression. Wu et al. (2024) show that GD reaches a sufficiently stable regime within $O(\eta)$ iterations. After the transition with additional t steps, they obtain an $\widetilde{O}(1/(\eta t))$ loss guarantee, which leads to an $\widetilde{O}(T^{-2})$ rate when the constant stepsize is chosen on the order of the optimization horizon as $\eta = \Theta(T)$. Their result demonstrates that instability can improve optimization, but the $O(\eta)$ transition bound leaves open whether the unstable phase can grow polynomially with the stepsize.

Crawshaw & Liu (2026) substantially sharpen this picture when the data dimension $d = 2$. They study the first time the loss reaches a sufficiently small threshold that guarantees subsequent non-increasing loss, and show that this time can be bounded independently of η once the stepsize is sufficiently large. This raises the natural question of whether the same phenomenon persists in higher dimensions.

We resolve this problem by giving the tight worst-case dependence on η for arbitrary sample size and dimension. We study the exact transition time, defined as the earliest time after which the

loss remains nonincreasing, which more directly captures the dynamical transition than entry into a sufficient low-loss region. Under the assumptions of Theorem 1, the two notions have the same worst-case asymptotic order. The worst-case transition time is uniformly bounded when $n = 2$ or $d = 2$, but in higher-dimensional instances it can grow as $\Theta((\log \eta)^{\min\{n-2,d-2\}})$. Thus, the stepsize independence observed in two dimensions is a low-dimensional phenomenon rather than a dimension-free property.

Edge of stability. Cohen et al. (2021) observed that full-batch neural network training often operates at EoS, where the largest Hessian eigenvalue stays near $2/\eta$ and the loss oscillates while decreasing over longer time scales. Other work studies the mechanisms sustaining these dynamics (Lewkowycz et al., 2020; Arora et al., 2022; Damian et al., 2023; Cohen et al., 2025) and their behavior in simplified models (Chen & Bruna, 2023; Zhu et al., 2023; Ahn et al., 2023; Even et al., 2023). Related analyses account for adaptive preconditioning and stochastic gradients (Cohen et al., 2022; Mulayoff & Michaeli, 2024; Andreyev & Beneventano, 2024), with recent empirical work examining EoS in language model pretraining (Cai et al., 2026; Meterez et al., 2026). These studies focus on training dynamics and local stability. We instead study when GD on separable logistic regression enters a phase of permanent monotone loss decrease.

Separable logistic regression. On separable data, logistic loss has no finite minimizer, but GD with sufficiently small stepsizes converges in direction to the max-margin separator (Soudry et al., 2018). Related work studies convergence rates and loss tails (Ji & Telgarsky, 2019; Nacson et al., 2019a), regularization paths (Rosset et al., 2003; Ji et al., 2020), and dual characterizations (Ji & Telgarsky, 2021). Stochastic gradient descent also drives the loss to zero with a sufficiently small fixed stepsize (Nacson et al., 2019b), while finite iteration guarantees relate margins to generalization (Shamir, 2021). Under additional assumptions, Wu et al. (2023) extend convergence and implicit bias guarantees to arbitrary constant stepsizes despite possible loss oscillations. For nonseparable data, GD can instead approach periodic trajectories (Meng et al., 2024; 2025). For separable data, Tyurin (2025) relates the large stepsize limit to a batch perceptron and studies how quickly GD finds a separator, which need not mark the end of the EoS phase.

3 Main Results

This section formalizes the exact EoS transition time and presents the main results. Starting from the basic notation and necessary definitions in Section 3.1, we first establish its tight dependence on the stepsize in Section 3.2, and then derive two corollaries concerning the threshold transition time and finite-horizon optimization in Section 3.3.

3.1 Problem Setup

Notation. We consider binary logistic regression with n samples in $\mathbb{R}^d$, where $n, d \geq 2$. After absorbing the labels into the feature vectors, we write $D = (x_1, \ldots, x_n)$. For $0 < \gamma \leq 1$, let $\mathcal{D}_{n,d}(\gamma)$ denote the class of datasets for which there exists a unit vector $w^\star$ satisfying

$$\|x_i\| \leq 1, \qquad \langle w^\star, x_i \rangle \geq \gamma, \qquad i = 1, \ldots, n.$$

Repeated samples are allowed, and γ is a lower bound on the separator margin. All logarithms are natural, and all norms and inner products are Euclidean.

Gradient descent. For $D \in \mathcal{D}_{n,d}(\gamma)$, we consider the empirical logistic loss

$$F(w) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-\langle w, x_i \rangle)),$$

and gradient descent with a constant stepsize $\eta > 0$, $w_{t+1} = w_t - \eta \nabla F(w_t)$, initialized at $w_0 = 0$.

EoS transition time. For a fixed dataset D, define the exact EoS transition time as

$$\sigma_\eta(D) := \inf \{t \in \mathbb{Z}_{\geq 0} : F(w_{j+1}) \leq F(w_j) \text{ for every } j \geq t\},$$

with the convention $\inf \varnothing = \infty$. Thus, $\sigma_\eta(D)$ is the earliest time after which the loss remains nonincreasing, rather than the first iteration at which it decreases. It equals zero if the loss is nonincreasing from initialization.

3.2 TIGHT BOUNDS ON THE EOS TRANSITION TIME

This section characterizes the dependence of the worst-case EoS transition time on the stepsize.

Theorem 1 (Tight EoS transition time)**.** *For every integer $n \geq 2$, there exists a constant $\gamma_n \in (0, 1]$ such that, for every fixed $0 < \gamma \leq \gamma_n$,*

$$\sup_{D \in \mathcal{D}_{n,d}(\gamma)} \sigma_\eta(D) = \Theta\Big((\log \eta)^{\min\{n-2, d-2\}}\Big) \qquad \text{as } \eta \to \infty,$$

uniformly over all integers $d \geq 2$, where the implicit constants and the threshold on η depend only on n and γ, independent of dimension d.

The theorem shows that the worst-case EoS transition time can depend on the stepsize. While it remains bounded independently of η when $n = 2$ or $d = 2$, it grows with η once both $n, d \geq 3$. This dependence is nevertheless mild: it is only polylogarithmic in the stepsize.

The exponent $\min\{n - 2, d - 2\}$ describes how sample size and dimension control this stepsize dependence. Each unit increase in $\min\{n, d\}$ allows one additional logarithmic factor, while increasing the larger parameter alone leaves the exponent unchanged. In particular, for fixed n, the exponent saturates once $d \geq n$.

It is worth noting that the characterization is uniform in dimension d: the multiplicative constants and the threshold on η may depend only on n, γ. The upper bound holds for every fixed $0 < \gamma \leq 1$, while the matching lower bound is established for $\gamma \leq \gamma_n$. The proof further sharpens the upper-bound exponent to $\max\{r - 2, 0\}$, where $r = \operatorname{rank}(x_1, \ldots, x_n)$. Therefore, $d = 2$ suffices for a stepsize-independent bound.

3.3 COROLLARIES

We derive two corollaries of the transition bounds in Theorem 1. Corollary 2 connects the exact EoS transition time to the threshold-based notion studied in prior work, while Corollary 3 translates the transition bound into an optimization guarantee.

Threshold transition time. Prior work measures the transition through the first entry into a low-loss region that guarantees nonincreasing loss thereafter (Crawshaw & Liu, 2026). Following this convention, define

$$\tau_\eta(D) := \inf\left\{ t \in \mathbb{Z}_{\geq 0} : F(w_t) \leq \frac{1}{8\eta} \right\},$$

with the same convention $\inf \emptyset = \infty$. Corollary 2 implies that the same worst-case order holds for this threshold time.

Corollary 2 (Threshold transition time)**.** *With γ_n as in Theorem 1, for every fixed $n \geq 2$ and $0 < \gamma \leq \gamma_n$,*

$$\sup_{D \in \mathcal{D}_{n,d}(\gamma)} \tau_\eta(D) = \Theta\Big((\log \eta)^{\min\{n-2, d-2\}}\Big) \qquad \text{as } \eta \to \infty,$$

uniformly over all integers $d \geq 2$.

Crawshaw & Liu (2026) showed that $\tau_\eta(D)$ is bounded independently of η when $d = 2$, and conjectured the same bound in higher dimensions based on experiments. Corollary 2 disproves this conjecture: for fixed $n, d \geq 3$ and $0 < \gamma \leq \gamma_n$, the worst-case threshold time grows polylogarithmically with η. The exact transition time $\sigma_\eta(D)$ records when permanent monotonicity begins, while $\tau_\eta(D)$ records entry into a sufficient low-loss region. These times need not coincide for a fixed dataset, even though their worst-case asymptotic orders agree.

Proof Sketch of Corollary 2. The stability estimate in Appendix B guarantees permanent monotonicity once $F(w_t) \leq 2/\eta$, so $\sigma_\eta(D) \leq \tau_\eta(D)$. Appendix C proves $\tau_\eta(D) \leq C_{n,\gamma}(\log \eta)^{\min\{n-2,d-2\}}$ for all sufficiently large η, uniformly over $D \in \mathcal{D}_{n,d}(\gamma)$ and $d \geq 2$. Theorem 1 supplies the matching lower bound through $\sigma_\eta(D) \leq \tau_\eta(D)$, proving the corollary.

Optimization rate. The upper bound on $\tau_\eta(D)$ provided in Appendix C yields an optimization guarantee for every fixed $0 < \gamma \leq 1$. For $n, d \geq 3$, let $k = \min\{n-2, d-2\} \geq 1$. Choosing $\eta = \exp(aT^{1/k})$ with $a > 0$ sufficiently small ensures $\tau_\eta(D) \leq T/2$ for large T. The stable-phase estimate $F(w_{\tau_\eta(D)+s}) = O_\gamma(1/(\eta s))$, $s \in \mathbb{Z}_{\geq 1}$, then gives the following last-iterate guarantee.

Corollary 3 (Optimization)**.** *Fix $n \geq 3$ and $0 < \gamma \leq 1$. There exist constants $a, C > 0$ and an integer T_0, depending only on n and γ, such that the following holds for every $d \geq 3$ and every $D \in \mathcal{D}_{n,d}(\gamma)$. Let $k = \min\{n-2, d-2\}$. For every integer $T \geq T_0$, gradient descent with the constant stepsize $\eta = \exp\big(aT^{1/k}\big)$ satisfies*

$$F(w_T) \leq \frac{C}{T} \exp\Big(-aT^{1/k}\Big).$$

The stepsize is chosen before the run as a function of the prescribed horizon T and remains constant throughout optimization. For fixed n, d, γ, the corollary guarantees loss at most ϵ within $O((\log(1/\epsilon))^k)$ gradient updates as $\epsilon \to 0$. This improves the $\widetilde{O}(T^{-2})$ guarantee obtained from the $O(\eta)$ transition bound of Wu et al. (2024): the polylogarithmic transient permits an exponentially larger constant stepsize.

For completeness, when $k = 0$, we have $\tau_\eta(D) = O_{n,\gamma}(1)$ for sufficiently large η. The same analysis gives $F(w_T) = O_\gamma(1/(\eta T))$ once T exceeds twice the uniform transition bound.

4 Proof Sketch of Theorem 1

This section introduces the main ideas behind Theorem 1. The upper and lower bounds are driven by two complementary recursive mechanisms. For the upper bound, each logarithmic factor comes from partitioning a projected trajectory into $O(\log \eta)$ phases before reducing its rank. For the lower bound, a pair of coordinates generates the first logarithmic time scale. Concatenating orbit segments and adding gate coordinates produces higher powers of $\log \eta$. Throughout, we write

$$k = \min\{n-2, d-2\}, \qquad B = \frac{\eta}{n}, \qquad \lambda_{i,t} = \frac{B}{1 + \exp(\langle w_t, x_i \rangle)}.$$

For sufficiently large η, the upper bound proves $\sigma_\eta(D) \leq \tau_\eta(D) \lesssim (\log \eta)^k$ uniformly over $D \in \mathcal{D}_{n,d}(\gamma)$. The lower bound constructs a matching example when $0 < \gamma \leq \gamma_n$. The full proof is deferred to Appendix C and Appendix D.

Why the exponent is $\min\{n-2, d-2\}$. Each factor of $\log \eta$ comes from one additional recursive level, which allows the trajectory from the previous level to repeat logarithmically many times. The basic two-coordinate dynamics already produce a logarithmic trajectory starting from a fixed compact region in coefficient space; one additional sample and direction are used to reach this region from $w_0 = 0$. Thus the first logarithmic factor appears when $n, d \geq 3$. Thereafter, each additional factor requires one more interacting sample and one more independent direction. Thus, the lower-bound construction realizes $k = \min\{n-2, d-2\}$ logarithmic levels. The upper-bound rank recursion shows that no additional levels are possible, yielding the rate $\Theta\big((\log \eta)^{\min\{n-2, d-2\}}\big)$.

4.1 Upper Bound

The main obstacle is that the loss need not decrease before the transition, so it cannot serve as a global potential. Our proof instead controls the trajectory through the coefficients $\lambda_{i,t}$ in three steps. First, we reduce the problem to reaching a region of permanently monotone loss. Second, we partition the trajectory before $\tau_\eta(D)$ into outer blocks, restarting whenever a large coefficient appears or a previously negligible sample becomes significant again, and use a global coefficient budget to show that only $O(1)$ such blocks can occur. Finally, within each outer block, we retain a subset of samples that remains separable after projection and whose projected rank is at least one smaller. We then bound the length of the resulting lower-rank process recursively.

Step 1. Reduce permanent monotonicity to reaching a small loss. Lemma 4 shows that once the loss is sufficiently small, it remains nonincreasing at every subsequent iteration.

Lemma 4 (Wu et al. (2024)). *If $F(w_t) \le 2/\eta$, then $F(w_{j+1}) \le F(w_j)$ for every $j \ge t$.*

Logistic loss satisfies $\nabla^2 F(w) \preceq F(w) I$, and hence $\nabla^2 \log F(w) \preceq I$. Expanding $\log F$ to second order along one gradient step gives $\log F(w_{t+1}) - \log F(w_t) \le -\frac{\eta \|\nabla F(w_t)\|^2}{F(w_t)} \left(1 - \frac{\eta F(w_t)}{2}\right)$. Thus the loss is nonincreasing and stays at most $2/\eta$, so the argument repeats. Since τ_η uses the smaller threshold $1/(8\eta)$, this proves $\sigma_\eta(D) \le \tau_\eta(D)$. Appendix B gives the details.

Step 2. Bounding the number of outer blocks. Our goal is to show that the trajectory before τ_η can be partitioned into only $O_{n,\gamma}(1)$ outer blocks. A new block is needed only when an exceptional event occurs: either some coefficient becomes large, or a previously negligible sample becomes significant again. Let $c_t = \sum_i \lambda_{i,t}$ denote the coefficient mass at time t. Lemma 5 gives a global bound on the cumulative coefficient mass, which limits how often a block can be restarted.

Lemma 5 (Global coefficient mass). *For all sufficiently large η, there exists a universal constant $c_0 > 0$ such that $c_t \ge c_0$ for all $t < \tau_\eta$. Moreover, it holds that $\sum_{t<\tau_\eta} c_t \le \frac{2\eta}{\gamma^2} = O(\eta)$.*

Lemma 5 provides the global budget needed to control the number of outer blocks. Every restart is associated with one of two events. The first is a large coefficient, $\lambda_{i,t} \ge B/4$, which gives $c_t \ge B/4 = \Omega(\eta)$ in a single update. The second is that a previously negligible sample becomes significant again. The block construction below ensures that each such return also requires $\Omega(\eta)$ cumulative coefficient mass. Since Lemma 5 gives only $O(\eta)$ cumulative coefficient mass before τ_η, only $O(1)$ restarts can occur, and hence there are only $O(1)$ outer blocks.

To prove Lemma 5, track $\|w_t - (2\eta/\gamma)w^\star\|^2$. Before τ_η, this nonnegative quantity decreases by at least $2\eta c_t$ in each update. Its initial value is $4\eta^2/\gamma^2$, so summing these decreases over any finite prefix gives the $O(\eta)$ coefficient budget in Lemma 5. The same argument also yields $\|w_t\| \le 4\eta/\gamma$ for $t \le \tau_\eta$. Moreover, the logistic formula gives $c_t \ge c_0$ for $t < \tau_\eta$, so the finite budget ensures that $\tau_\eta < \infty$. Appendix C.2 supplies the details.

Step 3. Recursive control within each outer block. We bound the length of a single outer block by reducing it to a lower-rank process. Let $r = \mathrm{rank}(x_1, \dots, x_n)$. The case $r = 1$ is immediate, so assume $r \ge 2$. At a block start $s \ge 1$, consider $u = \frac{w_s}{\|w_s\|}$, projector $P = I - uu^\top$ and retain the samples satisfying $\langle Pw^\star, Px_i \rangle \ge \frac{\gamma}{2}$. The retained samples remain separable after projection, while their projected rank is at most $r - 1$. Thus, a single projection already reduces the rank by at least one. The remaining samples stay negligible throughout the block and contribute only a small additive error to the projected dynamics.

The projection removes the component along u, but this component still contributes to the retained margins. Since all sample margins are positive at the block start, its contribution is nonnegative and nondecreasing throughout the block. This leads to the following projected process.

Lemma 6 (Projected dynamics). *Let $y_1, \dots, y_m$ be vectors of norm at most one and rank r', separated by a unit vector with margin $\mu > 0$. Consider*

$$M_{i,t} = b_i + h_{i,t} + \langle z_t, y_i \rangle, \qquad \lambda_{i,t} = \frac{B}{1 + \exp(M_{i,t})}, \qquad z_{t+1} = z_t + \sum_i \lambda_{i,t} y_i + e_t,$$

where $z_0 = 0$, $b_i \ge \log 3$, $h_{i,0} = 0$, and each $h_{i,t}$ is nonnegative and nondecreasing. Assume that, for $0 \le t < T$, $\lambda_{i,t} < B/4, \sum_i \lambda_{i,t} \ge c > 0, \|z_t\| = O(B)$, and $\|e_t\| = O(B^{-6})$. Then, for sufficiently large B, it holds that $T \le C(1 + \log B)^{r'-1}$, where C and the threshold on B depend only on m, μ, c and the constants in the two norm bounds.

The projected dynamics within each original block satisfy the assumptions of Lemma 6: the removed direction contributes a nondecreasing offset, while the discarded samples contribute only the allowed small error. Since the projected rank is at most $r - 1$, the lemma gives a block length of $O_{n,\gamma}\big((1 + \log B)^{r-2}\big)$.

The lemma itself is proved recursively in Appendix C.1. When $r' = 1$, the coefficient mass falls below c within $O(1)$ updates. For $r' \ge 2$, after at most $O(\log B)$ initial updates, the trajectory is partitioned into $O(\log B)$ intervals. Within each interval, another projection produces the same process

class with fewer samples and rank at most $r'-1$. This leads to $T_{r'}(B) \lesssim (1+\log B)\, T_{r'-1}(B)$, and therefore $T_{r'}(B) = O\Big((1+\log B)^{r'-1}\Big)$. Together with the $O_{n,\gamma}(1)$ outer blocks and exceptional updates from Step 2, $\tau_\eta(D) \le C_{n,\gamma}(\log \eta)^{\max\{r-2,0\}}$. Since $r \le \min\{n,d\}$, the desired upper bound follows.

4.2 Lower Bound

Let $k = \min\{n-2, d-2\}$. The proof has three steps. First, we construct a two-coordinate trajectory of logarithmic length. Second, we add coordinates recursively, with each new coordinate introducing one additional logarithmic factor, yielding a trajectory of length $\Theta((\log B)^k)$. Third, we show that the loss still increases near the end of this trajectory and realize the construction by a separable dataset initialized at $w_0 = 0$. When $k = 0$, a separate example gives a strict loss increase at the first update, so $\sigma_\eta \ge 1$; see Appendix D.6.

Step 1. Constructing the base logarithmic trajectory. We begin with two sample vectors whose pairwise inner products are specified by the Gram matrix K_1, defined by $(K_1)_{ij} = \langle x_i, x_j\rangle$. We choose

$$K_1 = \begin{pmatrix} 1 & -a \\ -a & 1 \end{pmatrix}, \qquad a = \frac{4}{5}.$$

For any Gram matrix K, define $p_i = B\exp(-\langle w, x_i\rangle)$ and $h_B(x) = x/(1+x/B)$. Since the actual update coefficient is $\lambda_i = h_B(p_i)$, the induced dynamics are

$$p^+ = p \odot \exp(-K h_B(p)),$$

where the product, exponential, and h_B act coordinatewise. For K_1, when one p_i is large and the other is small, the negative off-diagonal entries make the next update decrease the large one and increase the small one. This alternating behavior is the basic mechanism used in the construction.

Lemma 7 (Base logarithmic trajectory)**.** *For all sufficiently large X satisfying $X^2 \le B$, there exists a trajectory of the K_1 dynamics with length $\Theta(\log X)$ whose initial state lies in a fixed compact subset of $(0,\infty)^2$ and whose final state has one coordinate equal to X.*

Taking $X = B^{1/4}$, Lemma 7 gives a trajectory of length $\Theta(\log B)$ starting from a fixed bounded region. This is the first logarithmic time scale in the lower-bound construction. A long trajectory alone does not yet imply a lower bound on the transition time; in Step 3, we will show that a strict loss increase occurs near the end of the constructed trajectory.

We construct the trajectory backward. Starting from a state with one large coordinate and one suitably chosen exponentially small coordinate, the exact inverse update exchanges their roles: the previously small coordinate becomes large, while the large value decreases from X to $aX + O(1)$. The resulting state has the same structure, so the same backward step can be repeated. Since $0 < a < 1$, the large coordinate decreases geometrically until it reaches a fixed constant range, which takes $\Theta(\log X)$ steps. Reversing the sequence gives the trajectory in Lemma 7. Appendix D.4, Step 1, gives the exact inverse calculation.

Step 2. Recursively adding logarithmic scales. The key idea is to make the previous-level trajectory repeat logarithmically many times. Each additional coordinate enables such a repetition, introducing one additional logarithmic factor in the trajectory length. This leads to Lemma 8.

Lemma 8 (Recursive logarithmic trajectory)**.** *For each fixed $r \ge 2$, there exists a Gram matrix K_r of size $r+1$ such that, whenever E is sufficiently large and $E \le B^{1/4}$, the K_r dynamics admit a trajectory segment of length $\Theta\big((\log E)^{r-1}\big)$, whose final state has last coordinate e^{-E} and whose initial gate coordinate is e^{-D}, where $\frac{E}{\Lambda_r} \le D \le \frac{E}{2}$ for some constant $\Lambda_r \ge 4$. The remaining coordinates at both endpoints lie in fixed compact sets independent of E.*

The construction is uniform over the bounded endpoint profiles that arise recursively: each segment starts from another admissible bounded profile, so the segments can be concatenated. We prove Lemma 8 by induction. For the base case $r = 2$, we add a third coordinate to the trajectory from Step 1. This coordinate remains exponentially small while the two-coordinate dynamics evolve, and this gives a trajectory of length $\Theta(\log E)$.

For the induction step, suppose the level-r construction has length $\Theta((\log E)^{r-1})$. We add one more coordinate and couple it weakly to the last coordinate of the existing system. The final update sends the new coordinate from a value of order E to e^{-E}. Through the negative coupling, solving this update backward places the last coordinate of the old system at $e^{-\Theta(E)}$. We then concatenate level-r segments backward, with the exponent in this coordinate decreasing by a fixed factor after each segment. This gives $\Theta(\log E)$ segments in total, whose lengths sum to $\Theta((\log E)^r)$. Along these concatenated segments, the new coordinate remains exponentially small, so its perturbation to the previous-level dynamics can be corrected backward to obtain an exact trajectory.

Finally, for $k \geq 2$, we start with $E = B^{1/4}$ and concatenate level-k segments backward until the exponent reaches a fixed constant range. This contributes one final logarithmic factor and gives

$$T = \Theta\big((\log B)^k\big).$$

The resulting trajectory starts in a fixed compact subset of the positive orthant and has total coefficient mass $O(B^{1/4})$. For $k = 1$, Step 1 gives the same conclusion with $T = \Theta(\log B)$. The full construction is provided in Appendix D.4.

Step 3. Forcing a late loss increase. We next show that the loss still increases near the end of the trajectory, since a long trajectory alone does not imply a late transition. For the auxiliary dynamics, define the scaled loss $L_B(p) = B\sum_i \log(1 + p_i/B)$. When the auxiliary coordinates correspond to samples in the original problem, L_B equals η times their contribution to F.

Proposition 9 (Late loss increase)**.** *For each fixed $k \geq 1$ and sufficiently large B, the constructed auxiliary orbit has $L_B(p_{t+1}) > L_B(p_t)$ at some $t = \Theta((\log B)^k)$.*

For $k \geq 2$, consider the last level-k segment constructed in Step 2. The full trajectory has length $\Theta((\log B)^k)$, whereas this final segment has length only $O((\log B)^{k-1})$. Hence this segment begins at time $\Theta((\log B)^k)$. At the beginning of the segment, the scaled loss is $O(1)$. During the segment, the final gate produces a coefficient of order $B^{1/4}$, and hence its coordinate reaches order $B^{1/4}$ at some later time. The scaled loss is therefore $\Omega(B^{1/4})$ at that time. Thus at least one update within the segment strictly increases the scaled loss. Since the segment itself begins at time $\Theta((\log B)^k)$, this increase also occurs at $t = \Theta((\log B)^k)$. For $k = 1$, the final update of the two-coordinate trajectory gives the strict increase directly.

It remains to realize the auxiliary trajectory from $w_0 = 0$. We add setup samples in one extra direction so that the first update reaches the required auxiliary initial state exactly. These samples then acquire margins of order B and remain negligible throughout the trajectory, since the total subsequent coefficient mass is only $O(B^{1/4})$. Hence the true gradient-descent trajectory stays close to the auxiliary one, so the late loss increase persists and $\sigma_\eta(D) \gtrsim (\log B)^k$. The construction also preserves bounded sample norms and a positive separation margin independent of B.

5 Conclusion

We established tight worst-case bounds for the Edge-of-Stability transition time of constant-stepsize gradient descent on separable logistic regression, showing that it scales as $\Theta((\log \eta)^{\min\{n-2,d-2\}})$ for sufficiently small fixed margin. Thus, stepsize independence holds when $n = 2$ or $d = 2$, while in higher-dimensional settings the EoS phase can grow polylogarithmically with the stepsize. The same order holds for the threshold transition time, yielding substantially faster finite-horizon optimization with large constant stepsizes.

Our analysis further shows that each logarithmic factor corresponds to an additional level of interaction in the dynamics. The upper bound limits these levels through recursive rank reduction, while the lower bound constructs instances that realize all of them. Together, these results give a sharp characterization of how sample size and data dimension govern the duration of large-stepsize instability. More broadly, it remains to understand whether analogous hierarchical mechanisms govern large-stepsize transitions beyond separable logistic regression.

AI USE STATEMENT

In this work, we used generative AI tools to assist with checking mathematical claims and theoretical derivations, and proofs. We did not use generative AI tools to design experiments or implement methods. Additionally, we used these tools for manuscript drafting and editing, literature searches, and LaTeX formatting. We reviewed all AI-assisted work, checking the text for accuracy and examining theoretical derivations and proofs for correctness. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19540–19569, 2023. URL https://papers.nips.cc/paper/2023/hash/3e592c571de69a43d7a870ea89c7e33a-Abstract-Conference.html.

Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging: Silver Stepsize Schedule for smooth convex optimization. *Mathematical Programming*, 213:1105–1118, 2025. doi: 10.1007/s10107-024-02164-2. URL https://link.springer.com/article/10.1007/s10107-024-02164-2.

Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for SGD. *arXiv preprint arXiv:2412.20553*, 2024. URL https://arxiv.org/abs/2412.20553.

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*, volume 162 of *Proceedings of Machine Learning Research*, pp. 948–1024. PMLR, 2022.

Kyriakos Axiotis and Maxim Sviridenko. Gradient descent converges linearly for logistic regression on separable data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1302–1319. PMLR, 2023. URL https://proceedings.mlr.press/v202/axiotis23a.html.

Han Bao, Shinsaku Sakaue, and Yuki Takezawa. Any-stepsize gradient descent for separable data under Fenchel–Young losses. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2455. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/69f98acf161316ed896047e45da3bc0c-Abstract-Conference.html.

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. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/835a0185f61867a1ea0f86155489839a-Abstract-Conference.html.

Yuhang Cai, Haofeng Huang, Haodong Wen, Deyi Liu, Yiyuan Ma, and Kaifeng Lyu. Does LLM pre-training typically occur at the edge of stability? In *Workshop on Scientific Methods for Understanding Deep Learning*, 2026. URL https://openreview.net/forum?id=QSb05IuPsy.

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*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4330–4391. PMLR, 2023. URL https://proceedings.mlr.press/v202/chen23b.html.

Jeremy M. 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 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. URL https://arxiv.org/abs/2207.14484.

Jeremy M. 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. URL https://arxiv.org/abs/2410.24206.

Michael Crawshaw and Mingrui Liu. Tight bounds for logistic regression with large stepsize gradient descent in low dimension. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 1575–1610. PMLR, 2026.

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*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11465–11492. PMLR, 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. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5da6ce80e97671b70c01a2e703b868b3-Abstract-Conference.html.

Benjamin Grimmer. Provably faster gradient descent via long steps. *SIAM Journal on Optimization*, 34(3):2588–2608, 2024. doi: 10.1137/23M1588408. URL https://epubs.siam.org/doi/10.1137/23M1588408.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1772–1798. PMLR, 2019. URL https://proceedings.mlr.press/v99/ji19a.html.

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*, volume 132 of *Proceedings of Machine Learning Research*, pp. 772–804. PMLR, 2021. URL https://proceedings.mlr.press/v132/ji21a.html.

Ziwei Ji, Miroslav Dudík, Robert E. Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2109–2136. PMLR, 2020. URL https://proceedings.mlr.press/v125/ji20a.html.

Ziwei Ji, Nathan Srebro, and Matus Telgarsky. Fast margin maximization via dual acceleration. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4860–4869. PMLR, 2021. URL https://proceedings.mlr.press/v139/ji21a.html.

Sacchit Kale, Piyushi Manupriya, Pierre Marion, Francis Bach, and Anant Raj. Stretched exponential convergence of (stochastic) gradient descent for separable logistic regression. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=R5OaFwCmS0.

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.

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, 2024. URL https://arxiv.org/abs/2406.05033.

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?, 2025. URL https://arxiv.org/abs/2507.11228.

Alexandru Meterez, Pranav Ajit Nair, Depen Morwani, Cengiz Pehlevan, Sham Kakade, and Alex Damian. A defense of the quadratic model. *arXiv preprint arXiv:2607.21716*, 2026. doi: 10.48550/arXiv.2607.21716. URL https://arxiv.org/abs/2607.21716.

Rotem Mulayoff and Tomer Michaeli. Exact mean square linear stability analysis for SGD. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 3915–3969. PMLR, 2024. URL https://proceedings.mlr.press/v247/mulayoff24a.html.

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 Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3420–3428. PMLR, 2019a. URL https://proceedings.mlr.press/v89/nacson19b.html.

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 Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3051–3059. PMLR, 2019b. URL https://proceedings.mlr.press/v89/nacson19a.html.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham, 2 edition, 2018. doi: 10.1007/978-3-319-91578-4. URL https://link.springer.com/book/10.1007/978-3-319-91578-4.

Saharon Rosset, Ji Zhu, and Trevor J. Hastie. Margin maximizing loss functions. In *Advances in Neural Information Processing Systems*, volume 16, pp. 1237–1244, 2003. URL https://papers.nips.cc/paper_files/paper/2003/hash/0fe473396242072e84af286632d3f0ff-Abstract.html.

Ohad Shamir. Gradient methods never overfit on separable data. *Journal of Machine Learning Research*, 22(85):1–20, 2021. URL https://jmlr.org/papers/v22/20-997.html.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20): 20938–20946, 2025. doi: 10.1609/aaai.v39i20.35389. URL https://ojs.aaai.org/index.php/AAAI/article/view/35389.

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, 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 Thirty-Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 5019–5073. PMLR, 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, 2025.

Ruiqi Zhang, Jingfeng Wu, and Peter Bartlett. Gradient descent converges arbitrarily fast for logistic regression via large and adaptive stepsizes. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76361–76384. PMLR, 2025a. URL https://proceedings.mlr.press/v267/zhang25cf.html.

Zihan Zhang, Jason Lee, Simon Du, and Yuxin Chen. Anytime acceleration of gradient descent. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5991–6013. PMLR, 2025b. URL https://proceedings.mlr.press/v291/zhang25a.html.

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. URL https://openreview.net/forum?id=p7EagBsMAEO.

CONTENTS

1 Introduction **1**

1.1 Technical Overview 2

2 Related Work **3**

3 Main Results **4**

3.1 Problem Setup 4

3.2 Tight Bounds on the EoS Transition Time 5

3.3 Corollaries 5

4 Proof Sketch of Theorem 1 **6**

4.1 Upper Bound 6

4.2 Lower Bound 8

5 Conclusion **9**

A Additional Related Works **15**

B Preliminaries and the monotone region **15**

B.1 Proof of Lemma 4 15

C Upper bound **16**

C.1 An auxiliary interval estimate 16

C.2 The original gradient trajectory 23

D Lower bound **30**

D.1 Exact coefficient dynamics 30

D.2 A uniform local inverse 31

D.3 Adding one negligible coordinate 33

D.4 An auxiliary orbit with a late loss increase 36

D.5 Realization by gradient descent from the origin 47

D.6 Completion of the lower bound 51

E Proof of Theorem 1 **52**

F Proofs of the corollaries **52**

F.1 Proof of Corollary 2 52

F.2 Proof of Corollary 3 53

Appendix

A Additional Related Works

Acceleration of gradient methods. Suitable stepsize schedules or momentum can improve margin convergence on separable data (Nacson et al., 2019a; Ji & Telgarsky, 2021; Ji et al., 2021). Adaptive or increasing stepsizes also accelerate logistic loss convergence (Axiotis & Sviridenko, 2023; Kale et al., 2026), including guarantees for averaged iterates (Zhang et al., 2025a). For smooth convex objectives, schedules with long steps improve convergence (Grimmer, 2024; Altschuler & Parrilo, 2025), with extensions providing guarantees at any stopping time (Zhang et al., 2025b).

Large stepsizes in related settings. The optimization benefits of large stepsizes extend to local GD with heterogeneous objectives (Crawshaw et al., 2025) and regularized logistic regression (Wu et al., 2025). Related work gives transition guarantees for two layer networks under activation and separability assumptions (Cai et al., 2024), and convergence with arbitrary constant stepsizes for a broader class of classification losses on separable data (Bao et al., 2025). In our setting, faster optimization follows from a tight bound on the EoS transition time with a fixed stepsize.

B Preliminaries and the monotone region

Throughout the appendix, we use the data class, initialization, and transition time $\sigma_\eta(D)$ from Section 3.1. The threshold time $\tau_\eta(D)$ is defined in Section 3.3 by $F(w_t) \le 1/(8\eta)$. We write $B = \eta/n$, as in Section 4. In particular,

$$F(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(\langle w, x_i\rangle), \qquad \ell(s) = \log(1+\exp(-s)),$$

and $\|x_i\| \le 1$ for every sample. All logarithms are natural. Vector norms are Euclidean, and matrix norms are the induced operator norms. For symmetric matrices, $A \preceq C$ means $z^\top A z \le z^\top C z$ for every vector z. We omit D from a transition time when the dataset is fixed.

B.1 Proof of Lemma 4

Proof. Let

$$q(s) = \frac{1}{1+\exp(s)}.$$

Then $\ell'(s) = -q(s)$ and $\ell''(s) = q(s)(1-q(s))$. Since

$$q(s) \le \log(1+\exp(-s)) = \ell(s),$$

where we used $\log(1+x) \ge x/(1+x)$ for $x \ge 0$, we obtain

$$\ell''(s) \le \ell(s).$$

Consequently,

$$\begin{aligned}\nabla^2 F(w) &= \frac{1}{n}\sum_{i=1}^{n} \ell''(\langle w, x_i\rangle) x_i x_i^\top \\ &\preceq \frac{1}{n}\sum_{i=1}^{n} \ell(\langle w, x_i\rangle) I = F(w) I.\end{aligned}$$

Therefore

$$\nabla^2 \log F(w) = \frac{\nabla^2 F(w)}{F(w)} - \frac{\nabla F(w)\nabla F(w)^\top}{F(w)^2} \preceq I.$$

For any displacement v, apply the fundamental theorem of calculus twice to $g(u) = \log F(w+uv)$ on $[0,1]$. The Hessian bound gives

$$g(1) - g(0) = g'(0) + \int_0^1 (1-u) g''(u)\,\mathrm{d}u \le \frac{\langle \nabla F(w), v\rangle}{F(w)} + \frac{1}{2}\|v\|^2.$$

Taking $v = -\eta \nabla F(w)$ and $w^+ = w + v$ yields

$$\log F(w^+) - \log F(w) \le -\eta \frac{\|\nabla F(w)\|^2}{F(w)} + \frac{\eta^2}{2}\|\nabla F(w)\|^2$$
$$= -\eta \frac{\|\nabla F(w)\|^2}{F(w)} \left(1 - \frac{\eta F(w)}{2}\right). \tag{1}$$

If $F(w) \le 2/\eta$, the right hand side is nonpositive. Hence $F(w^+) \le F(w) \le 2/\eta$. The sublevel set is therefore forward invariant and the loss is nonincreasing at every subsequent iterate. Since the threshold defining τ_η is $1/(8\eta) < 2/\eta$, we also have $\sigma_\eta(D) \le \tau_\eta(D)$. □

C Upper bound

We prove that, for each $n \ge 2$ and $0 < \gamma \le 1$, there are constants $C_{n,\gamma} > 0$ and $\eta_{n,\gamma}$ such that

$$\tau_\eta(D) \le C_{n,\gamma}(1 + \log \eta)^{\min\{n-2, d-2\}} \tag{2}$$

for every $d \ge 2$, $D \in \mathcal{D}_{n,d}(\gamma)$, and $\eta \ge \eta_{n,\gamma}$. Lemma 4 then gives the upper bound in Theorem 1.

C.1 An auxiliary interval estimate

Fix an integer $m \ge 1$ and constants

$$0 < \mu \le 1, \qquad c > 0, \qquad K > 0, \qquad E \ge 0.$$

Let $y_1, \dots, y_m$ be vectors in a finite dimensional Euclidean space with

$$\|y_i\| \le 1, \qquad \langle v^\star, y_i \rangle \ge \mu, \qquad \|v^\star\| = 1.$$

Within this subsection, write

$$r = \dim \operatorname{span}\{y_1, \dots, y_m\}.$$

Thus r is the rank of the samples in the auxiliary process.

Let T be a nonnegative integer and $B > 0$. For $0 \le t \le T$, define margins and coefficients by

$$M_{i,t} = b_i + h_{i,t} + \langle z_t, y_i \rangle, \qquad \lambda_{i,t} = \frac{B}{1 + \exp(M_{i,t})}. \tag{3}$$

The numbers b_i are fixed, and the numbers $h_{i,t}$ satisfy

$$b_i \ge \log 3, \qquad h_{i,0} = 0, \qquad 0 \le h_{i,t} \le h_{i,t+1} \qquad (0 \le t < T). \tag{4}$$

Assume that $z_0 = 0$ and

$$z_{t+1} = z_t + \sum_{i=1}^m \lambda_{i,t} y_i + e_t \qquad (0 \le t < T), \tag{5}$$

with the bounds

$$\lambda_{i,t} < \frac{B}{4}, \qquad \sum_{i=1}^m \lambda_{i,t} \ge c, \qquad \|z_t\| \le KB, \qquad \|e_t\| \le EB^{-6} \qquad (0 \le t < T). \tag{6}$$

The vectors e_t are additive errors. No condition is imposed on how the offsets or errors are chosen beyond the displayed assumptions. The value of B does not change when some samples are removed.

The margins and offsets are defined through time T, but the bounds in (6) are required only before T. In particular, neither a coefficient bound nor a norm bound is assumed at the terminal state z_T. This distinction allows the last update of an interval to be included even when that update ends the interval.

We claim that there are constants $C > 0$ and $B_0 < \infty$, depending only on m, μ, c, K, E, such that every process just described with $B \ge B_0$ satisfies

$$T \le C(1 + \log B)^{r-1}. \tag{7}$$

Because $r \le m$, the exponent is also $\min\{m-1, r-1\}$. The constants are uniform over the sample vectors and their ambient dimension. We now prove this claim in full.

We use strong induction on m. This means that, when proving the statement for m samples, we may apply it to every smaller sample count and every choice of its scalar parameters. The case $r = 1$ will be proved for every m. It includes $m = 1$ and therefore provides the base of the induction.

During the proof of this auxiliary estimate, a constant denoted by C may increase from one occurrence to the next. It depends only on m, μ, c, K, E. Every lower threshold imposed on B will have the same permitted dependence. We may assume $B \ge \mathrm{e}$ and $T \ge 1$, since $T = 0$ is immediate.

Step 1. Restriction to the sample span.

Let

$$V = \operatorname{span}\{y_1, \dots, y_m\},$$

and let P_V be the orthogonal projection onto V. For any $y_i \in V$,

$$\langle P_V z_t, y_i\rangle = \langle z_t, P_V y_i\rangle = \langle z_t, y_i\rangle.$$

Thus replacing z_t by $P_V z_t$ leaves every margin and coefficient unchanged. Projecting (5) replaces e_t by $P_V e_t$, without increasing its norm or the norm of z_t.

We must also put the separator in V. The vector $P_V v^\star$ is nonzero, since

$$\langle P_V v^\star, y_i\rangle = \langle v^\star, y_i\rangle \ge \mu > 0.$$

Moreover, orthogonal projection cannot increase a norm, so $0 < \|P_V v^\star\| \le 1$. Consequently,

$$\langle \frac{P_V v^\star}{\|P_V v^\star\|}, y_i\rangle = \frac{\langle v^\star, y_i\rangle}{\|P_V v^\star\|} \ge \mu.$$

Replace $v^\star$ by this normalized vector and retain the same notation. We may now assume

$$z_t, e_t, v^\star \in V. \tag{8}$$

This reduction will be repeated at the beginning of each recursive application. A separator inherited from a previous projection need not already belong to the smaller sample span.

Step 2. A preliminary length bound and accumulated error.

Write

$$c_t = \sum_{i=1}^{m} \lambda_{i,t}, \qquad a_t = \langle z_t, v^\star\rangle.$$

Taking the inner product of (5) with $v^\star$ gives

$$a_{t+1} - a_t = \sum_i \lambda_{i,t}\langle y_i, v^\star\rangle + \langle e_t, v^\star\rangle \ge \mu c_t - EB^{-6}.$$

Choose B so that $EB^{-6} \le \mu c/2$. Since $c_t \ge c$, we obtain

$$a_{t+1} - a_t \ge \frac{\mu}{2} c_t \ge \frac{\mu c}{2} \qquad (0 \le t < T). \tag{9}$$

As $a_0 = 0$, summing (9) over $0 \le t < T-1$ yields

$$(T-1)\frac{\mu c}{2} \le a_{T-1} \le \|z_{T-1}\| \le KB.$$

We have used the norm bound at $T-1$, not at the unbounded endpoint T. It follows that

$$T \le 1 + \frac{2KB}{\mu c}. \tag{10}$$

For $B \ge 1$, this implies

$$\sum_{t=0}^{T-1} \|e_t\| \le EB^{-6}\left(1 + \frac{2KB}{\mu c}\right) \le DB^{-5}, \qquad D = E\left(1 + \frac{2K}{\mu c}\right). \tag{11}$$

The same bound applies to the sum of error norms on any subinterval. We will use (10) only to control these errors, not as our final estimate for T.

Step 3. The case of rank one.

Suppose $r = 1$. By (8), we can write

$$y_i = \beta_i v^\star, \qquad \mu \le \beta_i \le 1, \qquad z_t = a_t v^\star.$$

Using the monotonicity of $h_{i,t}$ and (9),

$$\begin{aligned} M_{i,t+1} - M_{i,t} &= h_{i,t+1} - h_{i,t} + \beta_i(a_{t+1} - a_t) \\ &\ge \frac{\mu^2}{2} c_t. \end{aligned} \tag{12}$$

Let $\theta = \mu^2/2$. For $M \ge \log 3$ and $\Delta \ge 0$, a direct calculation gives

$$\begin{aligned} \frac{B/(1+\exp(M+\Delta))}{B/(1+\exp M)} &= \exp(-\Delta)\frac{1+\exp(-M)}{1+\exp(-(M+\Delta))} \\ &\le \frac{4}{3}\exp(-\Delta). \end{aligned} \tag{13}$$

Here $\exp(-M) \le 1/3$, and the denominator on the first line is at least one. For $t < T$, the coefficient condition in (6) implies $M_{i,t} > \log 3$. Applying (13) to (12) and then summing over i yields

$$c_{t+1} \le \frac{4}{3} c_t \exp(-\theta c_t).$$

The function $s\exp(-\theta s)$ attains its maximum at $s = 1/\theta$, where its value is $1/(\mathrm{e}\theta)$. Thus the first update gives the constant bound

$$c_1 \le C_0, \qquad C_0 = \frac{4}{3\mathrm{e}\theta}.$$

If $1 \le t < T$, summing (12) from time 1 to time $t-1$ gives

$$M_{i,t} - M_{i,1} \ge \theta \sum_{j=1}^{t-1} c_j \ge \theta c(t-1).$$

Apply (13) directly between times 1 and t. We obtain

$$c \le c_t \le \frac{4}{3} c_1 \exp(-\theta c(t-1)) \le \frac{4}{3} C_0 \exp(-\theta c(t-1)).$$

The factor $4/3$ appears only once because this is a direct comparison of two times, rather than a product of successive comparisons. If $T \ge 2$, taking $t = T-1$ shows

$$T \le 2 + \frac{\max\{0, \log(4C_0/(3c))\}}{\theta c}.$$

This also gives a constant upper bound when $T \le 1$. Therefore the auxiliary estimate holds for $r = 1$, uniformly in B.

Step 4. Growth in the separator direction.

Assume $m, r \ge 2$, and assume the auxiliary estimate has been proved for every smaller sample count. Set

$$A = \frac{16}{\mu}, \qquad z_t = a_t v^\star + \widetilde{z}_t, \qquad \widetilde{z}_t \perp v^\star.$$

By (9), the number of update indices before a_t first reaches $A \log B$ is at most

$$1 + \frac{2A \log B}{\mu c}.$$

If this level is not reached at any time $t < T$, the same expression bounds T. That is enough, because $r - 1 \geq 1$. We may therefore restrict attention to times $t < T$ with $a_t \geq A \log B$. Once this inequality holds, it remains true by (9).

At every such time,

$$\|\widetilde{z}_t\| > \frac{\mu}{4} a_t. \tag{14}$$

To prove this, suppose instead that $\|\widetilde{z}_t\| \leq \mu a_t/4$. Then, for each i,

$$\begin{aligned} M_{i,t} &= b_i + h_{i,t} + a_t \langle v^\star, y_i \rangle + \langle \widetilde{z}_t, y_i \rangle \\ &\geq \log 3 + \mu a_t - \|\widetilde{z}_t\| \\ &\geq \log 3 + \frac{3\mu}{4} a_t. \end{aligned}$$

We used $a_t \geq 0$, $h_{i,t} \geq 0$, and $\|y_i\| \leq 1$. Since $1 + \exp M \geq \exp M$, this implies

$$\sum_i \lambda_{i,t} \leq \frac{mB}{3} \exp\left(-\frac{3\mu}{4} a_t\right) \leq \frac{m}{3} B^{1-3\mu A/4} = \frac{m}{3} B^{-11}.$$

For sufficiently large B, the final expression is smaller than c, contradicting (6). This proves (14). Geometrically, if the coefficient mass remains at least c, the trajectory cannot point too closely in the separator direction, since that would make every margin large.

Step 5. Choosing a subset during one phase.

Fix a time $s < T$ with $a_s \geq A \log B$. Define

$$R = \|\widetilde{z}_s\|, \qquad u = \frac{\widetilde{z}_s}{R}, \qquad \kappa = \frac{\mu a_s}{4R}.$$

The vector u and the numbers R, κ remain fixed during this phase. By (8) and (14),

$$u \in V, \qquad \|u\| = 1, \qquad u \perp v^\star, \qquad 0 < \kappa < 1. \tag{15}$$

In particular, $u \in V$ uses the fact that both z_s and $v^\star$ belong to V. For each sample, set

$$\alpha_i = \langle u, y_i \rangle, \qquad S = \{i : \alpha_i \leq -\kappa\}.$$

The inequality $\|y_i\| \leq 1$ gives $|\alpha_i| \leq 1$. The set S contains the samples that have a sufficiently negative component along u.

For $t \geq s$, decompose

$$z_t = a_t v^\star + \rho_t u + \zeta_t, \qquad \rho_t = \langle z_t, u \rangle, \qquad \zeta_t \perp v^\star, \qquad \zeta_t \perp u. \tag{16}$$

At the start, $\rho_s = R$ and $\zeta_s = 0$. Let σ be the first time in $\{s+1, \ldots, T-1\}$ at which

$$\rho_\sigma \leq 0 \quad \text{or} \quad \|\zeta_\sigma\| \geq \frac{\mu}{8} a_\sigma. \tag{17}$$

If no such time exists, set $\sigma = T$. The phase contains the update indices $s \leq t < \sigma$. Thus it always contains at least one update, and before its endpoint we have

$$\rho_t > 0, \qquad \|\zeta_t\| < \frac{\mu}{8} a_t.$$

We next show that S is nonempty and proper. If S were empty, then $\alpha_i > -\kappa$ for every i, and at time s,

$$M_{i,s} = b_i + h_{i,s} + a_s \langle v^\star, y_i \rangle + R\alpha_i > \log 3 + \mu a_s - \kappa R = \log 3 + \frac{3\mu}{4} a_s.$$

As in Step 4, this would give total coefficient mass smaller than c. Therefore S is nonempty.

To show that S is proper, sum the exact recursion from the original initial time 0 to time s. Since $z_0 = 0$,

$$z_s = \sum_{j=0}^{s-1} \sum_i \lambda_{i,j} y_i + \mathcal{E}_s, \qquad \mathcal{E}_s = \sum_{j=0}^{s-1} e_j, \qquad \|\mathcal{E}_s\| \leq DB^{-5}.$$

If every sample belonged to S, then every $\alpha_i \le -\kappa < 0$. Taking the inner product with u would yield

$$R = \sum_{j=0}^{s-1} \sum_i \lambda_{i,j}\alpha_i + \langle \mathcal{E}_s, u\rangle \le DB^{-5},$$

because all coefficients are positive. But (14) gives

$$R > \frac{\mu}{4} a_s \ge \frac{\mu A}{4} \log B = 4 \log B.$$

These two bounds are incompatible for sufficiently large B. We have proved

$$1 \le |S| \le m - 1. \tag{18}$$

Step 6. Small coefficients outside the subset.

We prove, by induction over the update indices in the phase, that

$$0 < \rho_t \le R \quad (s \le t < \sigma), \qquad \rho_{t+1} < \rho_t \quad (s \le t < \sigma). \tag{19}$$

The first upper bound holds at $t = s$ because $\rho_s = R$. Suppose it has been proved at the current time $t < \sigma$. The stopping rule gives $\rho_t > 0$. For $i \notin S$, we have $\alpha_i > -\kappa$, and hence

$$\alpha_i \rho_t \ge -\kappa \rho_t \ge -\kappa R.$$

Using (16) and the stopping rule,

$$\begin{aligned} M_{i,t} &\ge \log 3 + \mu a_t + \alpha_i \rho_t - \|\zeta_t\| \\ &\ge \log 3 + \frac{7\mu}{8} a_t - \kappa R \\ &= \log 3 + \frac{7\mu}{8} a_t - \frac{\mu}{4} a_s \\ &\ge \log 3 + \frac{5\mu}{8} a_t. \end{aligned}$$

The last inequality follows from $a_t \ge a_s$, which is a consequence of (9). Therefore

$$\sum_{i \notin S} \lambda_{i,t} \le \frac{mB}{3} \exp\left(-\frac{5\mu}{8} a_t\right) \le \frac{m}{3} B^{1-5\mu A/8} = \frac{m}{3} B^{-9} \le B^{-6} \tag{20}$$

for sufficiently large B. Subtracting this from the total mass bound gives

$$\sum_{i \in S} \lambda_{i,t} \ge c - B^{-6} \ge \frac{c}{2}. \tag{21}$$

Taking the inner product of the update with u, we now obtain

$$\begin{aligned} \rho_{t+1} - \rho_t &= \sum_{i \in S} \alpha_i \lambda_{i,t} + \sum_{i \notin S} \alpha_i \lambda_{i,t} + \langle e_t, u\rangle \\ &\le -\kappa \sum_{i \in S} \lambda_{i,t} + \sum_{i \notin S} \lambda_{i,t} + \|e_t\| \\ &\le -\frac{\kappa c}{2} + (E+1) B^{-6}. \end{aligned}$$

We must compare the error with κ, since κ itself may depend on B. Orthogonality in $z_s = a_s v^\star + \widetilde{z}_s$ gives $R \le \|z_s\| \le KB$. Thus

$$\kappa = \frac{\mu a_s}{4R} \ge \frac{\mu A \log B}{4KB}. \tag{22}$$

Choose B so that

$$(E+1) B^{-6} \le \frac{\mu A c \log B}{16 KB}.$$

Such a threshold depends only on the fixed scalar parameters, because $B^5 \log B$ tends to infinity. The right side is at most $\kappa c/4$ by (22). Consequently,

$$\rho_{t+1} - \rho_t \le -\frac{\kappa c}{4} < 0.$$

If $t+1 < \sigma$, positivity of ρ_{t+1} again follows from the stopping rule, and the induction continues. If $t+1 = \sigma$, the strict decrease still holds, although ρ_σ may be nonpositive. This proves (19), (20), and (21) for every update in the phase, including its last update.

Step 7. The exact projected process.

Let

$$P = I - uu^\top.$$

This is the orthogonal projection onto $u^\perp$. For $i \in S$ and $s \le t \le \sigma$, define

$$y_i' = Py_i, \qquad q_t = P(z_t - z_s),$$
$$b_i' = M_{i,s}, \qquad h_{i,t}' = h_{i,t} - h_{i,s} + \alpha_i(\rho_t - R).$$

We verify each hypothesis of the auxiliary estimate for these new quantities.

First, $\|y_i'\| \le \|y_i\| \le 1$. Since $u \perp v^\star$, we have $Pv^\star = v^\star$, so

$$\langle v^\star, y_i' \rangle = \langle Pv^\star, y_i \rangle = \langle v^\star, y_i \rangle \ge \mu.$$

Thus the same unit separator is still valid. It may not belong to the span of the retained samples, but Step 1 will address this at the start of the recursive application.

Second, the rank decreases. On the space V, the kernel of P is exactly $\operatorname{span}\{u\}$. Indeed, $Px = 0$ is equivalent to $x = \langle u, x\rangle u$, and $u \in V$ by (15). The dimension formula for a linear map therefore gives $\dim P(V) = r - 1$. Since every y_i' belongs to $P(V)$,

$$1 \le r' := \dim \operatorname{span}\{y_i' : i \in S\} \le r - 1. \tag{23}$$

The lower bound follows from $S \neq \varnothing$ and $\langle v^\star, y_i' \rangle \ge \mu > 0$.

Third, the new process starts at zero, since $q_s = 0$. The original coefficient bound at time s gives

$$\frac{B}{1 + \exp(M_{i,s})} < \frac{B}{4} \quad \Longrightarrow \quad M_{i,s} > \log 3.$$

Hence $b_i' > \log 3$. We also have $h_{i,s}' = 0$, and, for $s \le t < \sigma$,

$$h_{i,t+1}' - h_{i,t}' = h_{i,t+1} - h_{i,t} + \alpha_i(\rho_{t+1} - \rho_t) \ge 0.$$

The first term is nonnegative by assumption. The product is nonnegative because both α_i and $\rho_{t+1} - \rho_t$ are negative. This remains true for the update ending at σ. Thus the new offsets are nonnegative and nondecreasing through the endpoint.

Fourth, the margins retain their exact logistic representation. By the definition of q_t,

$$z_t - z_s = q_t + (\rho_t - R)u.$$

Since $q_t \perp u$, we have $\langle q_t, y_i \rangle = \langle q_t, Py_i \rangle$. It follows that

$$\begin{aligned} M_{i,t} &= M_{i,s} + h_{i,t} - h_{i,s} + \langle z_t - z_s, y_i \rangle \\ &= M_{i,s} + h_{i,t} - h_{i,s} + \alpha_i(\rho_t - R) + \langle q_t, Py_i \rangle \\ &= b_i' + h_{i,t}' + \langle q_t, y_i' \rangle. \end{aligned} \tag{24}$$

In particular, the coefficient of y_i' is the original $\lambda_{i,t}$, with the same B.

Finally, projecting the update gives

$$\begin{aligned} q_{t+1} - q_t &= P(z_{t+1} - z_t) \\ &= \sum_{i \in S} \lambda_{i,t} y_i' + e_t', \qquad e_t' = Pe_t + \sum_{i \notin S} \lambda_{i,t} Py_i. \end{aligned} \tag{25}$$

By the triangle inequality and (20),

$$\|e_t'\| \le \|e_t\| + \sum_{i \notin S} \lambda_{i,t} \|P y_i\| \le (E+1) B^{-6}.$$

For $s \le t < \sigma$, we also have

$$\sum_{i \in S} \lambda_{i,t} \ge \frac{c}{2}, \qquad \|q_t\| \le \|z_t - z_s\| \le \|z_t\| + \|z_s\| \le 2KB.$$

Every retained coefficient is still less than $B/4$. After shifting the time origin from 0 to s and relabeling the retained samples, this process satisfies all assumptions of the auxiliary estimate with

$$(m, \mu, c, K, E) \quad \text{replaced by} \quad (|S|, \mu, c/2, 2K, E+1). \tag{26}$$

The number of samples is smaller by (18). By the induction hypothesis and (23),

$$\sigma - s \le C(1 + \log B)^{r'-1} \le C(1 + \log B)^{r-2}. \tag{27}$$

To make the constant and the threshold uniform over all possible S, take the maximum of the inductive constants for sample counts $1, \dots, m-1$ with the scalar parameters in (26). There are only finitely many such counts. The estimate does not require q_σ or the coefficients at time σ to satisfy the interior bounds.

Step 8. Counting the phases.

The preceding estimate controls one phase. We now show that there can be only a logarithmic number of phases. Suppose $\sigma < T$, so that at least one event in (17) has occurred. Set $\Delta z = z_\sigma - z_s$. Summing the update over the phase gives

$$\Delta z = g + \mathcal{E}, \qquad g = \sum_i A_i y_i, \qquad A_i = \sum_{t=s}^{\sigma-1} \lambda_{i,t} \ge 0, \qquad \|\mathcal{E}\| \le DB^{-5}.$$

The positive combination g satisfies

$$\|g\| \le \sum_i A_i \|y_i\| \le \sum_i A_i, \qquad \langle v^\star, g \rangle \ge \mu \sum_i A_i \ge \mu \|g\|.$$

Using $\|g\| \ge \|\Delta z\| - \|\mathcal{E}\|$, we obtain

$$\begin{aligned} a_\sigma - a_s &= \langle v^\star, \Delta z \rangle \\ &\ge \mu \|g\| - \|\mathcal{E}\| \\ &\ge \mu \|\Delta z\| - (1+\mu)\|\mathcal{E}\| \\ &\ge \mu \|\Delta z\| - 2DB^{-5}. \end{aligned} \tag{28}$$

Here $\mu \le 1$ was used in the last line. Thus a large displacement necessarily produces a substantial increase in the separator component.

Choose B large enough that

$$2DB^{-5} \le \frac{\mu^2 A \log B}{16}. \tag{29}$$

This threshold depends only on the fixed parameters. If $\|\zeta_\sigma\| \ge \mu a_\sigma / 8$, then the orthogonal decomposition

$$\Delta z = (a_\sigma - a_s) v^\star + (\rho_\sigma - R) u + \zeta_\sigma$$

implies $\|\Delta z\| \ge \|\zeta_\sigma\|$. Since $a_\sigma \ge a_s \ge A \log B$, equations (28) and (29) yield

$$a_\sigma - a_s \ge \frac{\mu^2}{8} a_\sigma - 2DB^{-5} \ge \frac{\mu^2}{16} a_\sigma.$$

Rearranging gives

$$a_\sigma \ge \frac{a_s}{1 - \mu^2/16}.$$

If instead $\rho_\sigma \le 0$, then

$$\|\Delta z\| \ge |\rho_\sigma - R| \ge R > \frac{\mu}{4} a_s.$$

The same two estimates imply

$$a_\sigma - a_s \ge \frac{\mu^2}{4} a_s - 2DB^{-5} \ge \frac{\mu^2}{8} a_s.$$

Since $0 < \mu \le 1$ and $(1-x)^{-1} \ge 1 + x$ for $0 \le x < 1$, both cases give

$$a_\sigma \ge \left(1 + \frac{\mu^2}{16}\right) a_s. \tag{30}$$

At time $\sigma < T$, start a new phase for the same original process, choosing a new u and a new S from its current state. The initial condition used in the positive combination argument remains $z_0 = 0$. The time shift is used only for the recursive process of Step 7. Since a_t is increasing, every new phase begins with $a_t \ge A \log B$.

Let N be the number of phases whose endpoints are strictly smaller than T, and let s_0 be the start of the first phase. After these N phases, repeated use of (30) gives

$$\left(1 + \frac{\mu^2}{16}\right)^N A \log B \le a_{s_N} \le \|z_{s_N}\| \le KB,$$

where $s_N < T$ is the endpoint of the last such phase. For $N = 0$, set $s_N = s_0$. Taking logarithms yields

$$N \le \frac{\log\big(KB/(A \log B)\big)}{\log(1 + \mu^2/16)} \le C(1 + \log B).$$

The argument $KB/(A \log B)$ of the logarithm is at least one whenever a phase exists, because then $a_{s_0} \ge A \log B$ and $a_{s_0} \le KB$. There is at most one additional phase ending at T.

Step 9. Completion of the induction.

The initial part before $a_t \ge A \log B$ has length at most $C(1+\log B)$. There are at most $C(1+\log B)$ subsequent phases, and (27) bounds each one. Therefore

$$\begin{aligned} T &\le C(1 + \log B) + C(1 + \log B)(1 + \log B)^{r-2} \\ &\le C(1 + \log B)^{r-1}, \end{aligned}$$

since $r \ge 2$. This proves the inductive step.

For completeness, the dependence of the constants can be followed through the recursion. Every recursive application removes at least one sample, so there are at most $m - 1$ reductions along any sequence of applications. The margin lower bound stays at least μ, including after the restriction to the new sample span in Step 1. After j reductions, the remaining scalar parameters have the form

$$\frac{c}{2^j}, \qquad 2^j K, \qquad E + j, \qquad 0 \le j \le m - 1.$$

Only finitely many sample counts and these finitely many parameter choices can occur. Taking maxima over them gives constants depending only on m, μ, c, K, E. No geometric quantity other than the norm and margin bounds enters the estimates. In particular, the constants do not depend on rank or ambient dimension.

This proves the precise form of Lemma 6 used below. The fixed offsets, initialization, coefficient threshold, and error bound in that statement are specified in (3) through (6).

C.2 The original gradient trajectory

We return to the original data $x_1, \dots, x_n$ and set

$$r = \dim \operatorname{span}\{x_1, \dots, x_n\}.$$

Here r again denotes the rank of the original data, rather than the rank of the preceding auxiliary process. We will obtain the stronger estimate

$$\tau_\eta \le C_{n,\gamma}(1+\log\eta)^{\max\{r-2,0\}}.$$

Since $1 \le r \le \min\{n,d\}$, this implies (2).

Throughout this proof, the constants and all lower thresholds on η depend only on n and γ. We take η large enough that $B = \eta/n \ge \mathrm{e}$. Every additional threshold imposed below will have the same permitted dependence.

Step 1. The exact coefficients and an entropy inequality.

The derivative of the scalar logistic loss is

$$\frac{\mathrm{d}}{\mathrm{d}s}\log(1+\exp(-s)) = -\frac{\exp(-s)}{1+\exp(-s)} = -\frac{1}{1+\exp s}.$$

Thus, with

$$B = \frac{\eta}{n}, \qquad m_{i,t} = \langle w_t, x_i\rangle, \qquad \lambda_{i,t} = \frac{B}{1+\exp(m_{i,t})}, \qquad c_t = \sum_i \lambda_{i,t},$$

the gradient descent update becomes

$$w_{t+1} = w_t + \sum_i \lambda_{i,t} x_i = w_t + c_t u_t, \qquad u_t = \frac{1}{c_t}\sum_i \lambda_{i,t} x_i. \tag{31}$$

Every coefficient is strictly positive, so $c_t > 0$ and u_t is well defined. As a convex combination of the samples, u_t satisfies

$$\|u_t\| \le 1, \qquad \langle w^\star, u_t\rangle \ge \gamma, \qquad 0 < c_t \le nB = \eta. \tag{32}$$

Set $\pi_{i,t} = \lambda_{i,t}/c_t$, so that $\pi_{i,t} > 0$ and $\sum_i \pi_{i,t} = 1$. The coefficient formula gives

$$m_{i,t} = \log\frac{B-\lambda_{i,t}}{\lambda_{i,t}} \le \log\frac{B}{\lambda_{i,t}}.$$

Consequently,

$$\begin{aligned}\langle w_t, u_t\rangle &= \sum_i \pi_{i,t} m_{i,t} \\ &\le \sum_i \pi_{i,t}\log\frac{B}{c_t\pi_{i,t}} \\ &= \log\frac{B}{c_t} + \sum_i \pi_{i,t}\log\frac{1}{\pi_{i,t}}.\end{aligned}$$

The last sum is bounded by $\log n$. Indeed, concavity of the logarithm and the fact that the $\pi_{i,t}$ sum to one give

$$\sum_i \pi_{i,t}\log\frac{1}{\pi_{i,t}} \le \log\left(\sum_i \pi_{i,t}\frac{1}{\pi_{i,t}}\right) = \log n.$$

We have proved

$$\langle w_t, u_t\rangle \le \log\frac{nB}{c_t} = \log\frac{\eta}{c_t}. \tag{33}$$

This inequality relates the current parameter to the direction of its next update without requiring any bound on the individual margins.

Step 2. Finite transition time and total coefficient mass.

We first obtain a lower bound on c_t before the transition. The identity

$$1+\exp(-m_{i,t}) = \frac{1}{1-\lambda_{i,t}/B}$$

implies

$$\eta F(w_t) = B \sum_i -\log(1 - \lambda_{i,t}/B). \tag{34}$$

For $0 \le q \le 1/4$, integration gives

$$-\log(1-q) = \int_0^q \frac{\mathrm{d}s}{1-s} \le \frac{q}{1-q} \le \frac{4}{3}q.$$

If every $\lambda_{i,t} < B/4$, applying this inequality in (34) yields

$$\eta F(w_t) \le \frac{4}{3} c_t.$$

When $t < \tau_\eta$, the definition of τ_η gives $\eta F(w_t) > 1/8$, so $c_t > 3/32$. If some coefficient is at least $B/4$, then instead $c_t \ge B/4 \ge 3/32$, where the last inequality holds because $B \ge \mathrm{e}$. Thus, in both cases,

$$c_t \ge \frac{3}{32} \qquad (t < \tau_\eta). \tag{35}$$

Consider the squared distance

$$Z_t = \left\| w_t - \frac{2\eta}{\gamma} w^\star \right\|^2.$$

Using (31), expanding the square, and then applying (32) and (33), we get

$$\begin{aligned} Z_{t+1} - Z_t &= 2c_t \left\langle w_t - \frac{2\eta}{\gamma} w^\star, u_t \right\rangle + c_t^2 \|u_t\|^2 \\ &\le 2c_t \left(\log \frac{\eta}{c_t} - 2\eta \right) + c_t^2 \\ &\le 2c_t \log \frac{\eta}{c_t} - 3\eta c_t. \end{aligned}$$

The last step uses $c_t^2 \le \eta c_t$. Choose η sufficiently large that

$$2 \log \frac{32\eta}{3} \le \eta.$$

Such a threshold exists since $\log \eta / \eta$ tends to zero. For $t < \tau_\eta$, (35) then gives

$$Z_{t+1} - Z_t \le -2\eta c_t. \tag{36}$$

We have not yet proved that τ_η is finite, so we sum only over a finite prefix. For any finite integer $N \le \tau_\eta$,

$$Z_N - Z_0 = \sum_{t=0}^{N-1} (Z_{t+1} - Z_t) \le -2\eta \sum_{t=0}^{N-1} c_t.$$

Since $Z_N \ge 0$ and $Z_0 = 4\eta^2/\gamma^2$, this gives

$$\sum_{t=0}^{N-1} c_t \le \frac{2\eta}{\gamma^2}, \qquad Z_N \le Z_0. \tag{37}$$

If $\tau_\eta = \infty$, then (35) and (37) would imply

$$\frac{3N}{32} \le \frac{2\eta}{\gamma^2}$$

for every positive integer N. This is impossible. Thus $\tau_\eta < \infty$, and we may now set $N = \tau_\eta$ in (37) to obtain

$$\sum_{t<\tau_\eta} c_t \le \frac{2\eta}{\gamma^2}. \tag{38}$$

For every $0 \le t \le \tau_\eta$, the second part of (37) and the triangle inequality also give

$$\|w_t\| \le \left\| w_t - \frac{2\eta}{\gamma} w^\star \right\| + \frac{2\eta}{\gamma} \le \frac{4\eta}{\gamma} = \frac{4nB}{\gamma}. \tag{39}$$

Equations (35), (38), and (39) prove Lemma 5, with the universal choice $c_0 = 3/32$.

Let N_{large} be the number of update indices $t < \tau_\eta$ for which some $\lambda_{i,t} \ge B/4$. At each such index $c_t \ge B/4$, so (38) yields

$$N_{\text{large}} \frac{B}{4} \le \sum_{t<\tau_\eta} c_t \le \frac{2\eta}{\gamma^2}, \qquad N_{\text{large}} \le \frac{8n}{\gamma^2}. \tag{40}$$

Thus only a bounded number of updates have a coefficient above the threshold.

Step 3. The case of rank one.

Suppose the original data have rank $r = 1$. Choose a unit vector v along their common line and orient it so that $\langle w^\star, v\rangle > 0$. Write $x_i = a_i v$. The separator condition gives

$$a_i \langle w^\star, v\rangle \ge \gamma.$$

Because $0 < \langle w^\star, v\rangle \le 1$, we have $a_i \ge \gamma$ for every i. At the origin, every logistic coefficient equals $B/2$. Therefore

$$w_1 = \frac{B}{2} \sum_j x_j, \qquad \langle w_1, x_i\rangle = \frac{B}{2} a_i \sum_j a_j \ge \frac{nB\gamma^2}{2} = \frac{\eta\gamma^2}{2}.$$

Using $\log(1+s) \le s$ for $s \ge 0$,

$$F(w_1) \le \frac{1}{n} \sum_i \exp(-\langle w_1, x_i\rangle) \le \exp\left(-\frac{\eta\gamma^2}{2}\right).$$

For sufficiently large η, the last expression is at most $1/(8\eta)$. Hence $\tau_\eta \le 1$, which proves (2) when $r = 1$. For the rest of the proof, assume $r \ge 2$.

Step 4. A projection at the start of an interval.

The remaining update indices are those satisfying

$$\lambda_{i,t} < \frac{B}{4} \qquad \text{for every } i. \tag{41}$$

Partition them into maximal consecutive integer intervals $[a, b)$ within $\{0, \dots, \tau_\eta - 1\}$. Here $[a, b)$ means the indices $a, a+1, \dots, b-1$, so that it contains $b-a$ updates and ends at the state w_b. There are at most $N_{\text{large}} + 1$ such intervals. Their endpoints may be a time with a large coefficient or the transition time itself.

Fix one of these intervals. We will partition it further into blocks. Let $s \in [a, b)$ be the start of the current block, initially $s = a$. Every new block will use a newly defined projection. Since $\lambda_{i,0} = B/2$, time 0 does not satisfy (41). Thus $s \ge 1$. Summing the updates from time 0 gives

$$w_s = \sum_i A_i x_i, \qquad A_i = \sum_{t=0}^{s-1} \lambda_{i,t} > 0. \tag{42}$$

The separator component is increasing, because

$$\langle w_{t+1} - w_t, w^\star\rangle = \sum_i \lambda_{i,t} \langle x_i, w^\star\rangle \ge \gamma c_t > 0.$$

The first update consequently gives

$$\langle w_s, w^\star\rangle \ge \langle w_1, w^\star\rangle = \frac{B}{2} \sum_i \langle x_i, w^\star\rangle \ge \frac{\eta\gamma}{2}.$$

Set

$$R = \|w_s\|, \qquad u = \frac{w_s}{R}, \qquad P = I - uu^\top.$$

These quantities are well defined and satisfy

$$R \geq \frac{\eta\gamma}{2}, \qquad 0 < \langle w^\star, u\rangle \leq 1. \tag{43}$$

Define the subset

$$G = \left\{ i : \langle Pw^\star, Px_i\rangle \geq \frac{\gamma}{2} \right\}. \tag{44}$$

It is proper. Indeed, $Pw_s = 0$, so (42) gives

$$0 = \langle Pw^\star, Pw_s\rangle = \sum_i A_i \langle Pw^\star, Px_i\rangle.$$

If every sample belonged to G, the last sum would be at least $(\gamma/2)\sum_i A_i > 0$, a contradiction. Therefore

$$|G| \leq n - 1. \tag{45}$$

Put $\alpha_i = \langle u, x_i\rangle$. The small coefficient condition at time s is equivalent to $m_{i,s} > \log 3$, and hence

$$\alpha_i = \frac{m_{i,s}}{R} > 0 \qquad \text{for every } i. \tag{46}$$

For $i \notin G$, the projection identity gives

$$\langle w^\star, u\rangle \alpha_i = \langle w^\star, x_i\rangle - \langle Pw^\star, Px_i\rangle > \gamma - \frac{\gamma}{2} = \frac{\gamma}{2}.$$

By (43), $0 < \langle w^\star, u\rangle \leq 1$. Dividing by this positive number shows $\alpha_i > \gamma/2$. Thus every sample outside G starts the block with margin

$$m_{i,s} = R\alpha_i \geq \frac{\eta\gamma^2}{4} \qquad (i \notin G). \tag{47}$$

This bound is proportional to η, rather than to $\log \eta$. It will limit how many times an outside sample can become relevant again.

Step 5. Defining each block and bounding the number of blocks.

Increase the threshold on η so that

$$8 \log B \leq \frac{\eta\gamma^2}{8}. \tag{48}$$

For the fixed u, P, G chosen at time s, let σ be the first time in $\{s+1, \ldots, b-1\}$ at which some $i \notin G$ satisfies $m_{i,\sigma} \leq 8 \log B$. If no such time exists, set $\sigma = b$. The block is the update interval $[s, \sigma)$. Its initial margins satisfy $m_{i,s} > 8 \log B$ for $i \notin G$ by (47) and (48). Consequently, at every update index in the block,

$$\lambda_{i,t} \leq B \exp(-m_{i,t}) \leq B \exp(-8 \log B) = B^{-7} \qquad (i \notin G). \tag{49}$$

If $\sigma < b$, a new block begins at time σ. We then repeat Step 4 using this new start, recomputing R, u, P, and G. In particular, (47) holds for the outside samples of this new subset. We do not reuse the preceding projection or subset. Because each block has at least one update and b is finite, this procedure partitions the entire interval.

We now count all blocks, across all maximal intervals satisfying (41). For every sample and every time,

$$\begin{aligned} |m_{i,t+1} - m_{i,t}| &= \left| \sum_j \lambda_{j,t} \langle x_i, x_j\rangle \right| \\ &\leq \sum_j \lambda_{j,t} |\langle x_i, x_j\rangle| \\ &\leq \sum_j \lambda_{j,t} = c_t. \end{aligned}$$

The last step uses $|\langle x_i, x_j\rangle| \le \|x_i\|\|x_j\| \le 1$. If a block ends at $\sigma < b$, choose an outside sample with $m_{i,\sigma} \le 8 \log B$. By telescoping its margin changes and using (47) and (48),

$$\begin{aligned}\sum_{t=s}^{\sigma-1} c_t &\ge \sum_{t=s}^{\sigma-1} |m_{i,t+1} - m_{i,t}| \\ &\ge |m_{i,\sigma} - m_{i,s}| \\ &\ge m_{i,s} - m_{i,\sigma} \\ &\ge \frac{\eta\gamma^2}{4} - 8\log B \ge \frac{\eta\gamma^2}{8}.\end{aligned}$$

Thus every such endpoint consumes at least $\eta\gamma^2/8$ of the global coefficient mass.

Let N_{return} be the number of block endpoints lying strictly inside their maximal intervals. All corresponding update intervals are disjoint, including those in different maximal intervals. The mass bound (38) therefore gives

$$N_{\mathrm{return}} \frac{\eta\gamma^2}{8} \le \frac{2\eta}{\gamma^2}, \qquad N_{\mathrm{return}} \le \frac{16}{\gamma^4}. \tag{50}$$

Each maximal interval contributes one initial block, and each interior endpoint contributes one additional block. Hence the total number satisfies

$$N_{\mathrm{blocks}} \le N_{\mathrm{large}} + 1 + N_{\mathrm{return}} \le \frac{8n}{\gamma^2} + 1 + \frac{16}{\gamma^4}. \tag{51}$$

This uses one mass budget for the whole trajectory, not a separate budget for each interval.

Step 6. Applying the auxiliary estimate to one block.

Fix a block $[s, \sigma)$, with the quantities R, u, P, G recomputed at its own initial time s. For $s \le t \le \sigma$, write

$$w_t = \rho_t u + z_t, \qquad \rho_t = \langle w_t, u\rangle, \qquad z_t = P w_t.$$

Since $w_s = Ru$, we have $\rho_s = R$ and $z_s = 0$. Taking the inner product of (31) with u, and using (46), gives

$$\rho_{t+1} - \rho_t = \sum_i \lambda_{i,t}\alpha_i \ge 0. \tag{52}$$

The α_i are fixed by the block's initial direction, so their positivity remains valid throughout the block.

For $i \in G$, define

$$y_i = Px_i, \qquad b_i = m_{i,s} = R\alpha_i, \qquad h_{i,t} = \alpha_i(\rho_t - R).$$

Then $\|y_i\| \le 1$, $b_i > \log 3$, and $h_{i,s} = 0$. By (52) and $\alpha_i > 0$, the offsets are nonnegative and nondecreasing, including through time σ. Since $z_t \perp u$ and $x_i = \alpha_i u + y_i$,

$$\begin{aligned}m_{i,t} &= \langle \rho_t u + z_t, \alpha_i u + y_i\rangle \\ &= \rho_t\alpha_i + \langle z_t, y_i\rangle \\ &= b_i + h_{i,t} + \langle z_t, y_i\rangle.\end{aligned} \tag{53}$$

This is exactly the auxiliary margin formula. Projecting (31) also gives

$$z_{t+1} = z_t + \sum_{i\in G} \lambda_{i,t} y_i + e_t, \qquad e_t = \sum_{i\notin G} \lambda_{i,t} P x_i. \tag{54}$$

Using (49),

$$\|e_t\| \le \sum_{i\notin G} \lambda_{i,t}\|Px_i\| \le nB^{-7} \le B^{-6}$$

for $B \ge n$. The global lower bound (35) gives

$$\sum_{i\in G} \lambda_{i,t} = c_t - \sum_{i\notin G} \lambda_{i,t} \ge \frac{3}{32} - nB^{-7} \ge \frac{3}{64} \qquad (s \le t < \sigma) \tag{55}$$

for sufficiently large B. In particular, G is nonempty. Together with (45), this proves $1 \le |G| \le n-1$.

Nonemptiness of G also implies $Pw^\star \neq 0$, by its definition in (44). Thus the unit vector

$$v^\star = \frac{Pw^\star}{\|Pw^\star\|}$$

satisfies

$$\langle v^\star, y_i \rangle = \frac{\langle Pw^\star, Px_i \rangle}{\|Pw^\star\|} \ge \frac{\gamma}{2} \qquad (i \in G),$$

where we used $\|Pw^\star\| \le 1$. The norm bound (39) yields

$$\|z_t\| \le \|w_t\| \le \frac{4nB}{\gamma} \qquad (s \le t < \sigma).$$

Every retained coefficient remains less than $B/4$ because all update indices in the block satisfy (41).

It remains to verify the rank reduction. Let $V = \operatorname{span}\{x_1, \dots, x_n\}$, so that $\dim V = r$. Equation (42) implies $u \in V$. The kernel of P on V is $\operatorname{span}\{u\}$, and hence $\dim P(V) = r-1$. Because $y_i = Px_i \in P(V)$ for $i \in G$,

$$1 \le r_G := \dim \operatorname{span}\{y_i : i \in G\} \le r-1.$$

After shifting time by s and relabeling the retained samples, all hypotheses of the auxiliary estimate (7) now hold with

$$m = |G|, \qquad \mu = \frac{\gamma}{2}, \qquad c = \frac{3}{64}, \qquad K = \frac{4n}{\gamma}, \qquad E = 1.$$

The auxiliary estimate gives

$$\sigma - s \le C_{n,\gamma}(1+\log B)^{r_G - 1} \le C_{n,\gamma}(1+\log B)^{r-2}. \tag{56}$$

To make the constants uniform, take the maximum of the auxiliary estimate's constants and thresholds over $1 \le m \le n-1$ with the displayed scalar parameters. This maximum depends only on n and γ. The recursive sequence starts at $z_s = 0$ because we recomputed the projection at this particular block start. The endpoint σ need not satisfy a coefficient or norm bound for the application of the auxiliary estimate. Thus the last update of the block is included even if it ends at a large coefficient or at the transition time.

Step 7. Summing the interval bounds.

The update indices $0, \dots, \tau_\eta - 1$ form a disjoint union of the N_{large} indices counted in (40) and the blocks constructed above. Consequently,

$$\begin{aligned}
\tau_\eta &= N_{\mathrm{large}} + \sum_{\text{blocks } [s,\sigma)} (\sigma - s) \\
&\le \frac{8n}{\gamma^2} + \left(\frac{8n}{\gamma^2} + 1 + \frac{16}{\gamma^4} \right) C_{n,\gamma}(1+\log B)^{r-2} \\
&\le C_{n,\gamma}(1+\log B)^{r-2}.
\end{aligned}$$

The constant in the last line has been enlarged. This is allowed because $r \ge 2$ and $B \ge \mathrm{e}$, so the logarithmic factor is at least one. Combining this with Step 3 gives

$$\tau_\eta \le C_{n,\gamma}(1+\log B)^{\max\{r-2,0\}}.$$

Since $B = \eta/n$ and $n \ge 2$,

$$1 + \log B \le 1 + \log \eta.$$

Because $1 \le r \le \min\{n, d\}$ and $n, d \ge 2$, we also have

$$\max\{r-2, 0\} \le \min\{n-2, d-2\}.$$

This proves the claimed upper bound for $\tau_\eta(D)$. Together with Lemma 4, it also gives the upper bound for $\sigma_\eta(D)$.

All thresholds imposed directly on the original trajectory depend only on n and γ. In the auxiliary estimate, the recursive sample count decreases at every application, and all constants depend only on its listed scalar parameters. The outer application takes a maximum over only $n-1$ possible sample counts. Thus one value of $\eta_{n,\gamma}$ and one constant $C_{n,\gamma}$ work for every ambient dimension and every dataset satisfying the given norm and margin conditions.

D LOWER BOUND

We prove the lower bound in Theorem 1. Fix

$$k = \min\{n-2, d-2\}.$$

When $k \geq 1$, our goal is to construct a linearly separable dataset for which a strict loss increase occurs after order $(\log \eta)^k$ iterations. A strict increase at such a late time is stronger than merely delaying entry into a low loss set. It directly implies that permanent monotonicity has not started earlier.

The proof has two parts. We first construct a long exact trajectory for an auxiliary logistic system with $k+1$ interacting coordinates. The construction is recursive. A new coordinate stays exponentially small while the old dynamics evolve, but its logarithmic depth changes by a fixed fraction across each old episode. This lets one concatenate order $\log B$ copies of the previous construction, where

$$B = \frac{\eta}{n}.$$

At the last level, the new coordinate produces a large pulse near the end of the trajectory. We then realize the auxiliary trajectory, up to an exponentially small error, as the true gradient descent trajectory after the first update from $w_0 = 0$.

Throughout this section, k is fixed. Constants may depend on k and on matrices or compact sets fixed at earlier stages of the construction. They do not depend on B or on the scale of the current episode. All asymptotic statements are uniform over points in the stated compact sets. The index r below counts recursive levels. It is distinct from the sample rank used in Appendix C.

D.1 EXACT COEFFICIENT DYNAMICS

Consider vectors $v_1, \ldots, v_m$ with Gram matrix

$$K_{ij} = \langle v_i, v_j \rangle.$$

For a parameter vector w, write

$$m_i = \langle w, v_i \rangle$$

for the margin of the ith vector. Define

$$p_i = B\mathrm{e}^{-m_i}, \qquad h_B(x) = \frac{x}{1 + x/B}.$$

The quantity p_i is not the coefficient in the gradient update. The actual coefficient is

$$\frac{B}{1+\mathrm{e}^{m_i}} = \frac{B\mathrm{e}^{-m_i}}{1+\mathrm{e}^{-m_i}} = h_B(p_i).$$

Therefore one gradient update gives

$$w^+ = w + \sum_{i=1}^{m} h_B(p_i) v_i.$$

Taking inner products with the vectors v_i gives

$$m^+ = m + K h_B(p),$$

where h_B acts coordinatewise. Since $p_i = B\mathrm{e}^{-m_i}$, we obtain the exact coefficient map

$$p^+ = p \odot \exp\bigl(-K h_B(p)\bigr). \tag{57}$$

Here $\odot$ and the exponential are coordinatewise. We call a finite sequence satisfying (57) an exact orbit. For such a coefficient state, define the scaled loss by

$$L_B(p) = B \sum_i \log(1 + p_i/B). \tag{58}$$

When these vectors are samples in the original dataset, $L_B(p)$ equals η times their contribution to $F(w)$. Since $\eta > 0$, comparing scaled losses preserves the direction of a loss change.

It is convenient to use logarithmic coordinates

$$s = \log p.$$

Then the map becomes

$$\Phi_{B,K}(s) = s - Kh_B(\mathrm{e}^s). \tag{59}$$

Every auxiliary orbit constructed below satisfies (57) exactly. We never replace logistic loss by exponential loss.

We will repeatedly use

$$0 \le x - h_B(x) = \frac{x^2}{B+x} \le \frac{x^2}{B}, \qquad h_B'(x) = \frac{1}{(1+x/B)^2}. \tag{60}$$

If

$$t_B(x) = xh_B'(x) = \frac{x}{(1+x/B)^2},$$

then a direct derivative calculation gives

$$\left|\frac{d}{du}\log t_B(\mathrm{e}^u)\right| = \left|\frac{B-\mathrm{e}^u}{B+\mathrm{e}^u}\right| \le 1. \tag{61}$$

This simple bound is what makes the local inverse estimate below uniform even when some coordinates are exponentially small.

D.2 A UNIFORM LOCAL INVERSE

The recursive construction is most naturally built backward. We therefore need a local inverse whose size does not shrink when one coordinate becomes extremely small.

Lemma 10. *Fix $0 < \kappa < H < \infty$. There are constants $\delta > 0$, $R > 1$, $\alpha > 0$, and $L \ge 2$, depending only on κ and H, with the following property. Suppose*

$$\kappa I \preceq K \preceq HI$$

and every coordinate of p belongs to

$$(0, \delta] \cup [R, B].$$

Then $\Phi_{B,K}$ has a local inverse on the ball of radius α around $\Phi_{B,K}(\log p)$. This inverse sends the center to $\log p$ and is L Lipschitz.

Proof. Set

$$s_0 = \log p, \qquad D = \mathrm{diag}\big(t_B(p_1), \dots, t_B(p_m)\big).$$

Differentiating (59) gives

$$J := D\Phi_{B,K}(s_0) = I - KD.$$

We first prove that J^{-1} has a norm bound depending only on κ and H.

Let

$$A = \{i : p_i \ge R\}, \qquad S = \{1, \dots, m\} \setminus A.$$

For $i \in A$, the assumptions $R \le p_i \le B$ imply

$$t_B(p_i) = \frac{p_i}{(1+p_i/B)^2} \ge \frac{p_i}{4} \ge \frac{R}{4}.$$

Hence

$$D_A \succeq \frac{R}{4} I.$$

For $i \in S$, we have $p_i \le \delta$, so

$$0 \preceq D_S \preceq \delta I.$$

Assume first that both A and S are nonempty. In block form,

$$J = \begin{pmatrix} E & -K_{AS}D_S \\ -K_{SA}D_A & I - K_{SS}D_S \end{pmatrix}, \qquad E = I - K_{AA}D_A.$$

Because every principal submatrix of K satisfies the same lower spectral bound,

$$K_{AA} \succeq \kappa I.$$

Choose $R \geq 8/\kappa$. Since $D_A^{-1} \preceq 4I/R$, we obtain

$$K_{AA} - D_A^{-1} \succeq \frac{\kappa}{2} I.$$

Using

$$E = (D_A^{-1} - K_{AA}) D_A,$$

we get

$$E^{-1} = D_A^{-1}(D_A^{-1} - K_{AA})^{-1}.$$

Therefore

$$\|E^{-1}\| \leq \frac{8}{\kappa R}, \qquad \|D_A E^{-1}\| \leq \frac{2}{\kappa}. \tag{62}$$

Now solve $Jz = y$. The first block equation gives

$$z_A = E^{-1} y_A + E^{-1} K_{AS} D_S z_S.$$

Substituting this expression into the second block equation gives

$$S_0 z_S = y_S + K_{SA} D_A E^{-1} y_A,$$

where

$$S_0 = I - \big(K_{SS} + K_{SA} D_A E^{-1} K_{AS}\big) D_S.$$

Every submatrix of K has operator norm at most H. Using (62),

$$\|I - S_0\| \leq \left(H + \frac{2H^2}{\kappa}\right) \delta.$$

Choose δ small enough that the right hand side is at most $1/2$. Then the geometric series for S_0^{-1} converges and

$$\|S_0^{-1}\| \leq 2.$$

For clarity, the two block equations give the explicit bounds

$$\|z_S\| \leq 2\left(1 + \frac{2H}{\kappa}\right) \|y\|, \qquad \|z_A\| \leq \frac{8}{\kappa R} \left(\|y\| + H\delta \|z_S\|\right).$$

Since $R \geq 8/\kappa$ and we may take $\delta \leq 1$, summing these bounds shows that

$$\|J^{-1}\| \leq M$$

for a constant M that depends only on κ and H.

If A is empty, then $\|KD\| \leq H\delta \leq 1/2$ after decreasing δ, so $I - KD$ is inverted by a geometric series. If S is empty, the estimate follows directly from (62). Thus the same bound on J^{-1} holds in all cases.

We next show that the inverse exists on a ball whose radius is independent of the sizes of the small coordinates. If $\|v\| \leq \rho$, then $|v_i| \leq \rho$ for every coordinate. By (61),

$$\mathrm{e}^{-\rho} \leq \frac{t_B(\mathrm{e}^{s_{0,i}+v_i})}{t_B(\mathrm{e}^{s_{0,i}})} \leq \mathrm{e}^{\rho}.$$

Thus the derivative at the displaced point is

$$D\Phi_{B,K}(s_0 + v) = I - KD(I + Q_v), \qquad \|Q_v\| \leq \mathrm{e}^{\rho} - 1,$$

where Q_v is diagonal. Since $J = I - KD$, we have

$$J^{-1} KD = J^{-1} - I.$$

Therefore

$$\begin{aligned} J^{-1}\big(D\Phi_{B,K}(s_0 + v) - J\big) &= -J^{-1} KD Q_v \\ &= -(J^{-1} - I) Q_v. \end{aligned}$$

Choose $0 < \rho \le 1$ so that

$$(M+1)(\mathrm{e}^{\rho}-1) \le \frac{1}{4}.$$

Then for every s in the closed ball $\|s - s_0\| \le \rho$,

$$\|I - J^{-1} D\Phi_{B,K}(s)\| \le \frac{1}{4}. \tag{63}$$

Take a target y with

$$\|y - \Phi_{B,K}(s_0)\| \le \frac{\rho}{4M}$$

and define

$$\Psi_y(s) = s + J^{-1}\big(y - \Phi_{B,K}(s)\big).$$

At the center,

$$\|\Psi_y(s_0) - s_0\| \le M\|y - \Phi_{B,K}(s_0)\| \le \frac{\rho}{4}.$$

By (63), Ψ_y has Lipschitz constant at most $1/4$ on the ball. Hence for $\|s - s_0\| \le \rho$,

$$\|\Psi_y(s) - s_0\| \le \frac{\rho}{4} + \frac{1}{4}\|s - s_0\| \le \frac{\rho}{2}.$$

Thus Ψ_y maps the closed ball into itself and is a contraction. Starting at s_0 and repeatedly applying Ψ_y, the distances between successive iterates decrease by a factor at most $1/4$. Their sum is finite, so the iterates converge in the closed ball. Continuity gives a fixed point, and the same contraction inequality makes it unique. Its defining identity is exactly

$$\Phi_{B,K}(s) = y.$$

If y' is another target and s' is its fixed point, then

$$\|s - s'\| \le \frac{1}{4}\|s - s'\| + M\|y - y'\|.$$

Therefore

$$\|s - s'\| \le 2M\|y - y'\|.$$

Taking

$$\alpha = \frac{\rho}{4M}, \qquad L = \max\{2, 2M\}$$

proves the claim. The inverse is continuously differentiable on the interior of its target ball. Indeed, (63) implies that $D\Phi_{B,K}$ is invertible throughout the construction, so the inverse function theorem applies at each selected solution. Differentiating $\Phi_{B,K}(\mathcal{R}(y)) = y$ gives $D\mathcal{R}(y) = D\Phi_{B,K}(\mathcal{R}(y))^{-1}$. □

D.3 Adding one negligible coordinate

The next lemma says that a coordinate can be added without changing a previously constructed orbit by more than an exponentially small amount. This is the main device that allows us to build the hierarchy one coordinate at a time.

Lemma 11. *Fix $C \ge 1$, $q \ge 0$, and $a_0 > 0$. Let $B \ge 1$ and $T \ge 1$. Let*

$$\widetilde{K} = \begin{pmatrix} K & -b \\ -b^{\top} & 1 \end{pmatrix}, \qquad b \ge 0,$$

and suppose that both K and $\widetilde{K}$ satisfy common fixed positive upper and lower spectral bounds. Let $p_0, \dots, p_T$ be an exact orbit for K. Assume that every coordinate of every p_t lies in the ranges of Lemma 10 for these common bounds and that, for a scale $D \ge 1$,

$$T \le C(1 + \log D)^q, \qquad \max_{t,i} p_{i,t} \le CD, \qquad \sum_{t<T}\sum_{i} h_B(p_{i,t}) \le CD.$$

Let $Y > 0$ be the prescribed terminal value of the new coordinate. Assume either

$$\log Y \le -a_0 D$$

or

$$|\log Y| \le C \log D, \qquad b^\top h_B(p_{T-1}) \ge a_0 D,$$

Then, for all sufficiently large D, there is an exact orbit $(\widehat{p}_t, g_t)_{t=0}^{T}$ for $\widetilde{K}$ ending at (p_T, Y) such that

$$\max_{0 \le t < T} (\|\log \widehat{p}_t - \log p_t\| + g_t) \le \exp(-cD)$$

for a constant $c > 0$ depending only on the fixed parameters and spectral bounds.

Proof. For reference, the first terminal condition is

$$\log Y \le -a_0 D. \tag{64}$$

and the second is

$$|\log Y| \le C \log D, \qquad b^\top h_B(p_{T-1}) \ge a_0 D. \tag{65}$$

We construct an approximate sequence first and then correct it backward using Lemma 10.

We first keep the old orbit exactly equal to p_t and solve only for an approximate new coordinate. Set $\bar{g}_T = Y$. For $t = T-1, \dots, 0$, choose the small solution of

$$\log \bar{g}_t - h_B(\bar{g}_t) = \log \bar{g}_{t+1} - b^\top h_B(p_t). \tag{66}$$

We need to verify that this small solution exists at every step.

For a sufficiently negative number v, consider

$$\log x - h_B(x) = v.$$

On $0 < x \le 1/4$ the derivative with respect to $\log x$ is

$$1 - t_B(x) \ge \frac{3}{4}.$$

At $x = \mathrm{e}^v$, the left hand side is at most v. At $x = \exp(v + 2\mathrm{e}^v)$, the left hand side is at least v for all sufficiently negative v. Therefore there is a unique solution in $(0, 1/4)$ and it satisfies

$$v \le \log x \le v + 2\mathrm{e}^v. \tag{67}$$

Under (65), the right hand side of (66) at $t = T-1$ is at most

$$C \log D - a_0 D \le -\frac{a_0 D}{2}$$

for large D. Under (64), it is at most $-a_0 D$ because $b \ge 0$. Thus (67) gives

$$\bar{g}_{T-1} \le \exp(-a_0 D/3)$$

when D is large enough.

At an earlier step, the nonnegativity of b gives

$$\log \bar{g}_t \le \log \bar{g}_{t+1} + 2\bar{g}_{t+1}.$$

Let

$$q_D = \exp(-a_0 D/3).$$

Suppose the coordinates already constructed backward are at most $2q_D$. Then the total possible increase in their logarithms over all remaining steps is at most

$$4T q_D.$$

Since T is bounded by a fixed power of $\log D$, we have $4T q_D < \log 2$ for all sufficiently large D. Starting from $\bar{g}_{T-1} \le q_D$, backward induction yields

$$\bar{g}_t \le 2q_D \qquad 0 \le t < T. \tag{68}$$

This also confirms that every scalar inverse was taken on the intended small branch.

The sequence $(p_t, \bar{g}_t)$ is not yet an exact orbit of the enlarged system because the new coordinate exerts a small force on the old coordinates. In logarithmic variables let

$$z_t = (\log p_t, \log \bar{g}_t).$$

Equation (66) makes the last coordinate equation exact. The old coordinates satisfy

$$\Phi_{B,\widetilde{K}}(z_t) = z_{t+1} + \big(bh_B(\bar{g}_t), 0\big).$$

By (68), the defect has norm at most Cq_D.

Start from the exact terminal point $\widehat{z}_T = z_T$. At each $t < T$, the old coordinates p_t satisfy the required ranges and $\bar{g}_t \le 2q_D \le \delta$. Thus Lemma 10 applies at z_t for the enlarged matrix. No inverse is needed at the terminal state, so Y need not be in either coordinate range. Move backward using these local inverses, and let

$$\varepsilon_t = \|\widehat{z}_t - z_t\|.$$

If L is the uniform Lipschitz constant of the local inverse, then

$$\varepsilon_t \le L\varepsilon_{t+1} + CLq_D.$$

Since $\varepsilon_T = 0$, iteration gives

$$\varepsilon_t \le CTL^T q_D.$$

Taking logarithms and using $T \le C(1 + \log D)^q$ gives

$$\begin{aligned}\log(CTL^T q_D) &= -\frac{a_0 D}{3} + \log C + \log T + T \log L \\ &\le -\frac{a_0 D}{3} + O((1 + \log D)^q) \\ &\le -\frac{a_0 D}{6}\end{aligned}$$

for sufficiently large D. The last inequality follows from $(1 + \log D)^q = o(D)$ for fixed q. Thus, with $c = a_0/6$,

$$\max_{0 \le t \le T} \varepsilon_t \le \exp(-cD). \tag{69}$$

To justify every inverse application, choose D so large that $CTL^T q_D + Cq_D < \alpha$, where α is the target radius in Lemma 10. If the correction has been constructed through time $t+1$, the next target is at distance at most $\varepsilon_{t+1} + Cq_D < \alpha$ from $\Phi_{B,\widetilde{K}}(z_t)$. The inverse therefore exists at time t and gives the stated recurrence. This backward induction constructs the whole orbit. Since the logarithm of the corrected new coordinate differs from $\log \bar{g}_t$ by at most $\exp(-cD)$, after decreasing c,

$$\max_{t<T} g_t \le \exp(-cD).$$

Write the exact new coordinate as g_t. Summing its logarithmic update gives

$$\log g_T - \log g_0 = \sum_{t<T} b^\top h_B(\widehat{p}_t) - \sum_{t<T} h_B(g_t).$$

Using $g_T = Y$,

$$-\log g_0 = \sum_{t<T} b^\top h_B(\widehat{p}_t) - \log Y - \sum_{t<T} h_B(g_t).$$

For $|v| \le \epsilon \le 1$, the derivative bound $t_B(x) \le h_B(x)$ gives

$$|h_B(x\mathrm{e}^v) - h_B(x)| \le (\mathrm{e}^\epsilon - 1)h_B(x) \le 2\epsilon h_B(x).$$

Apply this with the logarithmic error in (69). The old total coefficient mass is at most CD, so its change is at most $CD\exp(-cD)$. The new coordinate contributes at most $T\exp(-cD)$ before time T. Decreasing c absorbs both factors. Therefore

$$-\log g_0 = b^\top \sum_{t<T} h_B(p_t) - \log Y + O(\exp(-cD)). \tag{70}$$

This identity will determine the scale of the new gate in the recursive construction. □

D.4 AN AUXILIARY ORBIT WITH A LATE LOSS INCREASE

We now construct the orbit underlying Lemmas 7 and 8 and Proposition 9 in Section 4. An orbit segment at one recursive level will be called an episode. Its endpoints and length will be specified in the proof. The loss comparisons for this auxiliary trajectory use L_B from (58). The corresponding contribution to the original empirical loss is L_B/η.

Proposition 12. *For each integer $k \geq 1$, there are a positive definite matrix K of size $k+1$ with unit diagonal and a compact set $\mathcal{P} \subset (0,\infty)^{k+1}$ such that, for every sufficiently large B, an exact orbit starting in $\mathcal{P}$ satisfies*

$$L_B(p_{t+1}) > L_B(p_t) \quad \textit{at some } t = \Theta((\log B)^k).$$

Proof. The proof is recursive. We begin with two coordinates and then add one coordinate at a time. Following Section 4, we call each added coordinate a gate. When its value is e^{-E}, its depth is E. We will also prove that the constructed orbit has length $T = \Theta((\log B)^k)$ and total coefficient mass $O(B^{1/4})$. These estimates will be used to realize the orbit from $w_0 = 0$ in Appendix D.5.

Step 1. The two coordinate orbit. Fix

$$a = \frac{4}{5}, \qquad K_1 = \begin{pmatrix} 1 & -a \\ -a & 1 \end{pmatrix}.$$

The negative off diagonal entries mean that a large coefficient in one coordinate suppresses itself and increases the other coordinate. We now make this precise by constructing predecessors.

Let (X, Y) be a target with X large, $X^2 \leq B$, and

$$|\log Y - \log(X/a) + a h_B(X)| \leq H_0, \tag{71}$$

where H_0 is fixed. We seek a predecessor (u, v). The inverse equations from (57) are

$$\log u - h_B(u) = \log X - a h_B(v), \tag{72}$$

$$\log v - h_B(v) + a h_B(u) = \log Y. \tag{73}$$

We first solve for u as a function of v on $[aX/2, 3aX/2]$. On this interval, v is of order X, and $v \leq B$ for large X because $X^2 \leq B$. Therefore

$$h_B(v) \geq \frac{v}{2}.$$

The right hand side of (72) is then at most $-cX$. By the scalar inverse estimate (67), there is a unique small solution and

$$u(v) \leq \exp(-cX). \tag{74}$$

Implicit differentiation of (72) gives

$$u'(v) = -\frac{a h_B'(v)}{1/u(v) - h_B'(u(v))} < 0.$$

Substitute $u(v)$ into (73) and define

$$\Psi(v) = \log v - h_B(v) + a h_B(u(v)) - \log Y.$$

For v of order X,

$$h_B'(v) = 1 + O(1/X)$$

because $v/B = O(1/X)$. Hence

$$\Psi'(v) = \frac{1}{v} - h_B'(v) + a h_B'(u(v)) u'(v) \leq -\frac{1}{2}$$

for all sufficiently large X.

At $v = aX$, use (71). The residual is

$$\Psi(aX) = 2\log a + a h_B(X) - h_B(aX) - \Big(\log Y - \log(X/a) + a h_B(X)\Big) + a h_B(u(aX)).$$

The quantity in parentheses is bounded by H_0. Also, by (60),

$$|ah_B(X) - h_B(aX)| \le \frac{aX^2}{B} + \frac{a^2X^2}{B} = O(1).$$

Finally $h_B(u(aX))$ is exponentially small. Thus $|\Psi(aX)| \le C$ for a fixed $C > 0$. Set

$$v_- = aX - (2C + 2), \qquad v_+ = aX + (2C + 2).$$

For sufficiently large X, both endpoints lie in $[aX/2, 3aX/2]$, where the derivative bound above holds. Integrating that bound gives

$$\Psi(v_-) \ge \Psi(aX) + \frac{aX - v_-}{2} \ge -C + (C + 1) = 1,$$
$$\Psi(v_+) \le \Psi(aX) - \frac{v_+ - aX}{2} \le C - (C + 1) = -1.$$

The intermediate value theorem therefore gives a root in (v_-, v_+). Since $\Psi'(v) < 0$, it is the unique root in this interval and satisfies

$$v = aX + O(1). \tag{75}$$

Together with (74), this gives the desired predecessor.

There is an exact expression for the error in the corresponding relation for the new pair. Rearranging (72),

$$\log u = \log X - ah_B(v) + h_B(u).$$

Therefore

$$\log u - \log(v/a) + ah_B(v) = \log \frac{aX}{v} + h_B(u).$$

By (75), the first term on the right is $O(1/X)$, and the second is exponentially small. Hence the new pair satisfies the same type of condition as (71), with an error that is in fact smaller.

We may now repeat the selected inverse. At every step the dominant coordinate switches from one component to the other. If X_j denotes the dominant value after the jth backward step, then

$$X_{j+1} = aX_j + O(1). \tag{76}$$

Choose a large fixed stopping level L_0. Stop when the dominant value first becomes at most $2L_0$. Iterating (76) gives

$$X_j = a^jX + O\left(\sum_{i=0}^{j-1} a^i\right) = a^jX + O(1).$$

The first crossing has $3L_0/2 \le X_J \le 2L_0$ for large enough L_0, because $X_{J-1} > 2L_0$ and $X_J = aX_{J-1} + O(1)$ with $a = 4/5$. Taking L_0 larger than the fixed error in the last display shows that a^JX is between two positive fixed constants. Therefore

$$J = \frac{\log X}{|\log a|} + O(1). \tag{77}$$

Reversing the backward sequence produces an exact forward orbit of length $\Theta(\log X)$, proving Lemma 7.

We also record two estimates that will be needed later. Summing the dominant values in backward time gives

$$\sum_{j=1}^{J} X_j = \frac{a}{1-a}X + O(J).$$

The small coordinates are uniformly bounded, so they contribute $O(J)$. Also $X_j \le Ca^jX$ before the stopping index, after increasing C. Replacing X_j by $h_B(X_j)$ therefore changes the sum by at most

$$\frac{1}{B}\sum_{j=1}^{J} X_j^2 = O(1)$$

because $X^2 \le B$. Consequently the total coefficient mass of the exact pair orbit is

$$\sum_{t<T}\big(h_B(p_{1,t}) + h_B(p_{2,t})\big) = \frac{a}{1-a}X + O(\log X). \tag{78}$$

The initial point obtained at the fixed stopping scale lies in a fixed compact subset of $(0,\infty)^2$.

When $k \ge 2$, we also prepare the finite collection of pair targets that will be needed by the later recursion. For

$$0 \le j \le k-2,$$

let Q_2^j be the union of the following set and the set obtained by swapping its two coordinates:

$$\left\{(X,Y) : a^{2j}L_0 \le X \le 3a^{-2j}L_0,\ |\log Y - \log(X/a) + aX| \le 2 + j\right\}. \tag{79}$$

Each Q_2^j is a compact subset of the positive quadrant with nonempty interior. The difference between $h_B(X)$ and X tends to zero uniformly on these fixed compact sets as B tends to infinity. Hence the predecessor estimate (75) and the residual estimate following it also hold uniformly over every Q_2^j. If L_0 is chosen large enough, the selected predecessor of every point in Q_2^j lies a fixed positive distance inside Q_2^{j+1} whenever $j < k-2$. The reason is simple. Its dominant coordinate is $aX + O(1)$, which lies strictly inside the enlarged interval at the next index, and its new residual has magnitude less than one. Since there are only finitely many compact sets, Lemma 10 gives one positive target radius and one derivative bound for every inverse branch that will be used later.

Taking $X = B^{1/4}$ already proves the proposition when $k = 1$, except for the strict loss increase. We verify that at the end of the proof.

Step 2. The first gate. Assume now that $k \ge 2$. We add a third coordinate. Fix

$$c = \frac{1}{5}, \qquad K_2 = \begin{pmatrix} 1 & -a & -c \\ -a & 1 & -c \\ -c & -c & 1 \end{pmatrix}.$$

The eigenvalues are

$$1 + a, \qquad \frac{3+\sqrt{6}}{5}, \qquad \frac{3-\sqrt{6}}{5}.$$

In particular, K_2 is positive definite with a fixed spectral gap from zero.

We call the third coordinate the gate. Write the terminal gate as $\exp(-E)$. The quantity E is its logarithmic depth. We now show that one exact three coordinate episode changes this depth by a fixed fraction while taking $\Theta(\log E)$ updates.

Fix

$$(P,Q) \in Q_2^j$$

for some $0 \le j \le k-2$. All estimates below are uniform over these finitely many compact target sets. Assume

$$E_0 \le E \le B^{1/4}.$$

We first solve the last update backward. Seek a predecessor of the form

$$(M, \mathrm{e}^{-\ell}, R).$$

The exact equations are

$$\log P = \log M - h_B(M) + a h_B(\mathrm{e}^{-\ell}) + c h_B(R), \tag{80}$$

$$\log Q = -\ell + a h_B(M) - h_B(\mathrm{e}^{-\ell}) + c h_B(R), \tag{81}$$

$$-E = \log R + c h_B(M) + c h_B(\mathrm{e}^{-\ell}) - h_B(R). \tag{82}$$

Ignoring logarithmic terms and the exponentially small coordinate suggests

$$M \approx cR, \qquad R - cM \approx E.$$

Solving these leading equations gives

$$\bar{M} = \frac{cE}{1-c^2}, \qquad \bar{R} = \frac{E}{1-c^2}.$$

Equation (81) then suggests

$$\bar{\ell} = (1+a)\bar{M}.$$

Let $f(M, R, \ell)$ be the vector obtained by moving the left hand sides of (80) through (82) to the right. At the center $(\bar{M}, \bar{R}, \bar{\ell})$, the terms of order E cancel by construction. For example,

$$\bar{M} - c\bar{R} = 0, \qquad \bar{R} - c\bar{M} = E.$$

The remaining terms contain $\log \bar{M}$, $\log \bar{R}$, the bounded numbers $\log P$ and $\log Q$, and the logistic corrections from (60). Since $\bar{M}, \bar{R} = \Theta(E)$ and $E \le B^{1/4}$,

$$\frac{\bar{M}^2}{B} + \frac{\bar{R}^2}{B} = O(1).$$

The terms containing $\mathrm{e}^{-\bar{\ell}}$ are exponentially small. Hence

$$\|f(\bar{M}, \bar{R}, \bar{\ell})\| \le C \log E.$$

We next compute the derivative. The three rows are

$$\left(\frac{1}{M} - h'_B(M), ch'_B(R), -at_B(\mathrm{e}^{-\ell})\right),$$

$$\left(ah'_B(M), ch'_B(R), -1 + t_B(\mathrm{e}^{-\ell})\right),$$

and

$$\left(ch'_B(M), \frac{1}{R} - h'_B(R), -ct_B(\mathrm{e}^{-\ell})\right).$$

On a ball of radius $C_1 \log E$ around the center, M, R, and ℓ remain positive and of order E. By (60),

$$h'_B(M) = 1 + O(E/B), \qquad h'_B(R) = 1 + O(E/B).$$

Also $1/M + 1/R = O(1/E)$ and $t_B(\mathrm{e}^{-\ell})$ is exponentially small. Therefore the derivative differs by

$$O(E^{-1} + E/B + \exp(-c_0 E))$$

from

$$J_0 = \begin{pmatrix} -1 & c & 0 \\ a & c & -1 \\ c & -1 & 0 \end{pmatrix}.$$

Its determinant is $1 - c^2 > 0$.

Choose C_1 large enough that

$$\|J_0^{-1} f(\bar{M}, \bar{R}, \bar{\ell})\| \le \frac{C_1 \log E}{2}.$$

Then increase the lower threshold on E so that

$$\|J_0^{-1}(Df - J_0)\| \le \frac{1}{2}$$

throughout the ball of radius $C_1 \log E$. The map

$$x \mapsto x - J_0^{-1} f(x)$$

then sends this ball into itself and has Lipschitz constant at most $1/2$. The contraction theorem gives an exact solution with

$$M = \frac{cE}{1-c^2} + O(\log E), \qquad R = \frac{E}{1-c^2} + O(\log E), \qquad \ell = (1+a)M + O(\log E). \quad (83)$$

We solve one more update backward. Seek a predecessor

$$(\mathrm{e}^{-A}, N, \mathrm{e}^{-H}).$$

The exact equations are

$$\log M = -A - h_B(\mathrm{e}^{-A}) + a h_B(N) + c h_B(\mathrm{e}^{-H}), \tag{84}$$

$$-\ell = \log N + a h_B(\mathrm{e}^{-A}) - h_B(N) + c h_B(\mathrm{e}^{-H}), \tag{85}$$

$$\log R = -H + c h_B(\mathrm{e}^{-A}) + c h_B(N) - h_B(\mathrm{e}^{-H}). \tag{86}$$

The leading solution has

$$N \approx \ell, \qquad A \approx a\ell - \log M, \qquad H \approx c\ell - \log R.$$

At this center the terms of order E again cancel. The remaining logarithmic terms give a residual of size $O(\log E)$. If the variables are ordered as (N, A, H), the derivative rows converge to

$$(a, -1, 0), \qquad (-1, 0, 0), \qquad (c, 0, -1).$$

Indeed, the terms containing e^{-A} and e^{-H} are exponentially small, while $h_B'(N) = 1 + O(E/B)$ and $1/N = O(1/E)$. Thus the limiting derivative is

$$J_1 = \begin{pmatrix} a & -1 & 0 \\ -1 & 0 & 0 \\ c & 0 & -1 \end{pmatrix},$$

which has determinant 1. On a ball of radius $C_2 \log E$, the derivative error is

$$O(E^{-1} + E/B + \exp(-c_0 E)).$$

Choose C_2 to dominate the central residual after multiplication by J_1^{-1}, and then increase the lower threshold on E so that the derivative error gives contraction factor at most $1/2$. This gives an exact solution with

$$N = (1+a)M + O(\log E), \qquad A = aN - \log M + O(1), \qquad H = cN - \log R + O(1). \tag{87}$$

We verify the target condition before applying the pair construction. Equation (84) gives

$$-A = \log M - a h_B(N) + h_B(\mathrm{e}^{-A}) - c h_B(\mathrm{e}^{-H}).$$

Thus the pair with dominant coordinate N satisfies

$$-A - \log(N/a) + a h_B(N) = \log(aM/N) + h_B(\mathrm{e}^{-A}) - c h_B(\mathrm{e}^{-H}).$$

The right hand side tends to $\log(a/(1+a))$, whose absolute value is less than one. Hence (71) holds with a fixed residual bound. Moreover, $N = \Theta(E)$ and $E \le B^{1/4}$ imply $N^2 \le B$ for sufficiently large B. We may therefore apply the selected pair inverse until its dominant coordinate returns to the fixed stopping scale. By (77), this prefix has length $\Theta(\log E)$. By (78), its coefficient mass is

$$\frac{a}{1-a} N + O(\log E).$$

The third coordinate at the end of this prefix is e^{-H}. We check all assumptions of Lemma 11 at scale $D = N$. The eigenvalues of K_1 and K_2 lie in $[1/16, 2]$. Choose L_0 large enough that every dominant pair coordinate is at least the corresponding R in Lemma 10, and every small coordinate is at most its δ. The largest coordinate is $O(N)$, which is less than B for large B because $N = \Theta(E)$ and $E \le B^{1/4}$. The preceding pair estimates give length $O(\log N)$ and total mass $O(N)$. Finally, (87) gives $H \ge cN/2$, so the prescribed terminal third coordinate satisfies (64). We may therefore apply the lemma with

$$b = c(1,1)^\top.$$

This produces an exact three coordinate prefix in which the first two coordinates differ from the pair orbit by only an exponentially small amount, and the third coordinate remains exponentially small throughout the prefix. The initial pair state of the unperturbed prefix lies a fixed positive distance inside Q_2^0. The correction tends to zero uniformly, so the corrected initial pair state remains in a fixed compact subset of the interior of Q_2^0.

Let D be the gate depth at the beginning of the resulting three coordinate episode. Formula (70) gives

$$D = H + c \sum_{t<T} \big(h_B(p_{1,t}) + h_B(p_{2,t})\big) + O(1).$$

Substituting the pair mass estimate and (87),

$$D = H + c\frac{a}{1-a}N + O(\log E) = \frac{c}{1-a}N + O(\log E).$$

Using (83) and (87),

$$\frac{D}{E} = \frac{c^2(1+a)}{(1-a)(1-c^2)} + O\left(\frac{\log E}{E}\right).$$

For $a = 4/5$ and $c = 1/5$, the constant is $3/8$. Hence, for all sufficiently large E,

$$\frac{E}{4} \le D \le \frac{E}{2}. \tag{88}$$

We have proved more than the depth ratio. The episode has length $\Theta(\log E)$, its total coefficient mass is $O(E)$, and the last update has a gate coefficient of order E because $R = \Theta(E)$. Before this last part, the gate is exponentially small. These estimates are exactly what will be needed in the recursive step.

Step 3. Adding further gates. For $3 \le r \le k$, define

$$K_r = \begin{pmatrix} K_{r-1} & -\varepsilon_r e_r \\ -\varepsilon_r e_r^\top & 1 \end{pmatrix}, \tag{89}$$

where e_r is the last standard basis vector in $\mathbb{R}^r$. The new coordinate is coupled only to the previous gate. We choose the couplings recursively. Once the construction through level r and its constant C_r have been fixed, choose $\varepsilon_{r+1} > 0$ to satisfy both bounds in (95) below. In particular, $\varepsilon_{r+1} \le 2^{-r-7}$, so for the fixed final level k,

$$\sum_{j=3}^{k} \varepsilon_j \le \sum_{r=2}^{k-1} 2^{-r-7} < \frac{1}{256}.$$

When $k = 2$, both sums are zero. Each coupling is fixed before constructing the next level and does not depend on B or the episode scale. The off diagonal perturbation that turns $\mathrm{diag}(K_{r-1}, 1)$ into K_r has operator norm ε_r. Therefore, for every unit vector z, its quadratic form changes by at most ε_r. Taking the minimum and maximum over unit vectors shows that the smallest and largest eigenvalues can move by at most ε_r at this step. Since the smallest eigenvalue of K_2 is larger than $1/10$, the largest equals $9/5$, and the total later perturbation is less than $1/256$, we have

$$\frac{1}{16} I \preceq K_r \preceq 2I \tag{90}$$

for every level that we use. This allows us to use the same type of local inverse estimate at every recursive step.

We now explain the induction. At level r, the matrix K_r has $r+1$ coordinates. We write it as

$$K_r = \begin{pmatrix} K_{r-1} & -b_r \\ -b_r^\top & 1 \end{pmatrix}, \qquad b_2 = c(1,1)^\top, \qquad b_r = \varepsilon_r e_r \quad (r \ge 3). \tag{91}$$

The last coordinate is the current gate. An endpoint has the form

$$(q, \mathrm{e}^{-D}),$$

where q contains the first r coordinates and stays in a fixed compact set.

There are two technical facts that must be preserved from one level to the next. First, the bounded profile must admit the few backward steps needed by later levels. Second, every coordinate must remain either uniformly small or uniformly large so that Lemma 10 applies. Since the final depth k is fixed, only finitely many such backward steps are ever needed.

We implement the first fact with a short chain of compact sets

$$Q_r^0 \subset Q_r^1 \subset \cdots \subset Q_r^{k-r} \subset (0,\infty)^r.$$

For every $j < k - r$, each target in Q_r^j has a selected predecessor for the bounded r coordinate dynamics that lies in a fixed compact subset of the interior of Q_r^{j+1}. The selected inverse is defined

on a logarithmic ball of a fixed positive radius, uniformly over the target. At level $r = 2$, these compact sets are obtained directly from the pair construction. We enlarge the allowed interval for the dominant coordinate and the bounded residual in (71) a fixed amount at each step. The estimates $v = aX + O(1)$ and

$$\log u - \log(v/a) + ah_B(v) = O(1/X)$$

show that the selected predecessor lies strictly inside the next enlarged set once the fixed stopping scale is chosen large enough. Compactness then gives one inverse radius that works uniformly on each set.

For the second fact, take δ and R from Lemma 10 using (90). Choose fixed numbers $\delta_* > 0$ and $R_* > 0$ such that

$$\mathrm{e}^{k+2}\delta_* < \delta, \qquad \mathrm{e}^{-k-2}R_* > R.$$

Choose the pair stopping scale large enough that its dominant coordinate is at least R_* and its other coordinate is at most δ_*, including on all the compact pair sets. Choose every subsequent fixed stopping depth large enough to give the same bounds for each newly constructed coordinate. Each correction below has logarithmic size less than one. Along any chain of constructions, a coordinate is corrected at most once per added coordinate, so these strict inequalities allow all corrections through level k. Concatenating episodes introduces no additional correction at their common endpoints.

The upper endpoint B requires a separate check. We maintain the bound $p_{i,t} \le C_r E$ at level r, where $E \le B^{1/4}$. For the finitely many levels, choose B sufficiently large that $\mathrm{e}^{k+2}\max_r C_r B^{1/4} < B$. Thus the corrected large coordinates also remain below B. This proves that the inverse lemma's two ranges are preserved whenever it is used.

The induction claim is as follows. For each $2 \le r \le k$, there are constants $\Lambda_r \ge 4$, $c_r > 0$, $C_r > 0$, and $E_r^* > 0$ that work for every $0 \le j \le k - r$ and every $q^+ \in Q_r^j$. For all $E_r^* \le E \le B^{1/4}$, there is an exact K_r orbit from

$$(q^-, \mathrm{e}^{-D})$$

to

$$(q^+, \mathrm{e}^{-E}).$$

The initial bounded profile q^- belongs to a fixed compact subset of the interior of Q_r^0. Throughout each episode, including both endpoints, every coordinate remains in the two ranges required by Lemma 10. The initial depth and the length satisfy

$$\frac{E}{\Lambda_r} \le D \le \frac{E}{2}, \tag{92}$$

$$c_r(\log E)^{r-1} \le T_r(E) \le C_r(\log E)^{r-1}. \tag{93}$$

We also maintain estimates needed for the next level. After decreasing c_r and increasing C_r if necessary, they are

$$\sum_{t<T_r(E)} \sum_{i=1}^{r+1} h_B(p_{i,t}) \le C_r E, \qquad \max_{0\le t\le T_r(E),\, i} p_{i,t} \le C_r E,$$
$$h_B(p_{r+1,T_r(E)-1}) \ge c_r E. \tag{94}$$

The last inequality bounds the gate coefficient at the last update. It also gives a lower bound for the gate's total mass, while the first inequality gives its upper bound. Step 2 establishes these inequalities for $r = 2$: the two final states have coordinates $O(E)$, the earlier pair mass is $O(E)$, and $R = \Theta(E)$ at the last update. Throughout the proof, a coordinate is $p_{i,t}$ and its update coefficient is $h_B(p_{i,t})$.

Assume the claim holds at level $r < k$. All constants and compact sets at level r are now fixed. We construct level $r + 1$.

Preparing the bounded target sets at the next level. With C_r now fixed, choose the new coupling

$$\varepsilon = \varepsilon_{r+1} > 0$$

small enough that

$$\varepsilon \le 2^{-r-7}, \qquad \varepsilon \le \frac{1}{8\sqrt{C_r + 1}}. \tag{95}$$

The second inequality will later give the upper depth bound $D \le E/2$.

Fix $j \le k - r - 1$ and a target $q^+ \in Q_r^j$. For a large number d, consider the exact equation

$$\log q^+ = \log q - K_{r-1} h_B(q) + b_r h_B(\mathrm{e}^{-d}). \tag{96}$$

Without the last term, this is precisely the bounded predecessor equation already available at level r. The perturbation has norm at most

$$\|b_r\| \mathrm{e}^{-d}.$$

For large d, it lies inside the fixed inverse neighborhood. Hence (96) has a selected solution, which we denote by $q(d)$. It lies in a fixed compact subset of the interior of Q_r^{j+1}.

We also need to know that $q(d)$ changes very little when d changes. Let $\mathcal{R}$ denote the selected inverse in logarithmic coordinates. Then

$$\log q(d) = \mathcal{R}\left(\log q^+ - b_r h_B(\mathrm{e}^{-d})\right).$$

Differentiate with respect to d. The derivative of $\mathcal{R}$ is uniformly bounded on the compact target set, and

$$\left| \frac{d}{dd} h_B(\mathrm{e}^{-d}) \right| = t_B(\mathrm{e}^{-d}) \le \mathrm{e}^{-d}.$$

Therefore

$$\|q'(d)\| \le C\mathrm{e}^{-d}. \tag{97}$$

Choose a constant $M \ge 1$ that bounds

$$b_r^\top h_B(q(d))$$

on every branch that can occur. Such an M exists because all $q(d)$ lie in finitely many fixed compact sets. Let

$$H_0 = M + 3.$$

Choose a fixed stopping depth $A > E_r^*$. Increase it so that $A/2 - kH_0 - 1$ is positive and is already large enough for (97), for e^{-d} to be in the small coordinate range, and for $C\mathrm{e}^{-d} \le 1/2$. This includes every depth in the intervals used below.

For $0 \le j \le k - r - 1$, define the depth interval

$$I_j = \left[\frac{A}{2} - jH_0, 2\Lambda_r A + jH_0 \right] \tag{98}$$

and define

$$Q_{r+1}^j = \left\{ (q, \mathrm{e}^{-d}) : q \in Q_r^j,\ d \in I_j \right\}.$$

These are compact sets. We now verify the required bounded predecessor property for them before constructing any level $r+1$ episode.

Take a target $(q^+, \mathrm{e}^{-d^+}) \in Q_{r+1}^j$ with $j < k - r - 1$. A predecessor under the K_r dynamics must have the form $(q(d), \mathrm{e}^{-d})$, where $q(d)$ solves (96) and the last coordinate equation is

$$d = d^+ + b_r^\top h_B(q(d)) - h_B(\mathrm{e}^{-d}). \tag{99}$$

On the interval

$$[d^+ - 1, d^+ + M + 1],$$

the right hand side of (99) lies between $d^+ - 1$ and $d^+ + M$. Its derivative has magnitude at most $C\mathrm{e}^{-d}$ by (97). By the choice of A, this is at most $1/2$. Hence the right hand side is a contraction of the interval into itself and has a unique fixed point.

The interval I_{j+1} is obtained from I_j by enlarging both endpoints by $H_0 = M + 3$. Thus the new depth lies a fixed positive distance inside I_{j+1}. The point $q(d)$ lies a fixed positive distance inside Q_r^{j+1}. Therefore the full predecessor lies in a fixed compact subset of the interior of Q_{r+1}^{j+1}. Lemma 10 then gives a uniform inverse neighborhood. This proves the bounded predecessor property at level $r+1$.

Solving the last update of a new episode. Fix a terminal point

$$(q^+, \mathrm{e}^{-d^+}, \mathrm{e}^{-E})$$

with $(q^+, \mathrm{e}^{-d^+}) \in Q_{r+1}^j$ and $j \le k - r - 1$. We seek the state immediately before the last update in the form

$$(q, \mathrm{e}^{-\Delta}, Y).$$

The exact equations are

$$\log q^+ = \log q - K_{r-1} h_B(q) + b_r h_B(\mathrm{e}^{-\Delta}), \tag{100}$$

$$E = h_B(Y) - \log Y - \varepsilon h_B(\mathrm{e}^{-\Delta}), \tag{101}$$

$$\Delta = d^+ + b_r^\top h_B(q) - h_B(\mathrm{e}^{-\Delta}) + \varepsilon h_B(Y). \tag{102}$$

For each sufficiently large Δ, the first equation has the selected solution $q(\Delta)$ constructed above. It lies in a fixed compact subset of the interior of Q_r^{j+1}.

Now fix Δ and solve (101). Define

$$H_B(Y) = h_B(Y) - \log Y.$$

For $Y \in [E/2, 2E]$ and $E \le B^{1/4}$,

$$H_B'(Y) = h_B'(Y) - \frac{1}{Y} \ge \frac{1}{2}$$

when E is large. At $Y = E/2$, we have $H_B(Y) < E$. At $Y = 2E$, (60) gives

$$h_B(2E) \ge 2E - \frac{4E^2}{B},$$

so $H_B(2E) > E + 1$ for large E. Since the perturbation $\varepsilon h_B(\mathrm{e}^{-\Delta})$ is smaller than one, (101) has a unique root in this interval. Denote it by $Y(\Delta)$.

Substituting $Y = E$ into (101) leaves a residual of size $O(\log E)$. Since the derivative is at least $1/2$,

$$Y(\Delta) = E + O(\log E). \tag{103}$$

Differentiating (101) with respect to Δ gives

$$H_B'(Y) Y'(\Delta) = -\varepsilon t_B(\mathrm{e}^{-\Delta}).$$

Therefore

$$|Y'(\Delta)| \le C\mathrm{e}^{-\Delta}. \tag{104}$$

Substitute $q(\Delta)$ and $Y(\Delta)$ into the right hand side of (102). Since d^+ and the bounded profile stay in fixed compact sets, and since (103) holds, this right hand side equals

$$\varepsilon E + O(\log E).$$

Its derivative with respect to Δ is $O(\mathrm{e}^{-\Delta})$ by (97) and (104). Hence, for large E, it maps

$$[\varepsilon E/2, 2\varepsilon E]$$

into itself and has derivative at most $1/2$ in magnitude. The contraction theorem gives a fixed point. Thus the exact last update is solved and

$$Y = E + O(\log E), \qquad \Delta = \varepsilon E + O(\log E). \tag{105}$$

After increasing the lower threshold on E, we may also assume

$$\Delta \le \frac{E}{2} \qquad \text{and} \qquad \Delta \le B^{1/4}.$$

Constructing the old orbit ending at depth Δ. Starting from the exact endpoint $(q, \mathrm{e}^{-\Delta})$, repeatedly apply level r episodes backward. The first terminal profile is in Q_r^{j+1}. Every earlier episode begins from a bounded profile in the fixed compact interior of Q_r^0. Hence every invocation satisfies the induction hypothesis.

Write $E_0 = \Delta$. If $E_{\ell+1}$ is the initial depth of the episode whose terminal depth is E_ℓ, then (92) gives

$$\frac{E_\ell}{\Lambda_r} \le E_{\ell+1} \le \frac{E_\ell}{2}. \tag{106}$$

Choose E large enough that $\Delta > \Lambda_r A$. Apply an episode whenever the current depth exceeds $\Lambda_r A$, and stop at the first depth at most $\Lambda_r A$. Every application then has terminal depth greater than $A > E_r^*$. At the last application, the lower bound in (106) keeps the resulting depth greater than A. The upper bound halves the depth at each application, so the process stops after $O(\log \Delta)$ episodes with depth in $(A, \Lambda_r A]$.

Reading the concatenation in forward time gives an exact old orbit ending at $(q, \mathrm{e}^{-\Delta})$. Let S_g be the total coefficient mass of its current gate. Since $E_\ell \le 2^{-\ell}\Delta$, the mass estimates in the induction hypothesis give

$$c_r \Delta \le S_g \le 2C_r \Delta. \tag{107}$$

The same geometric sum gives total coefficient mass at most $2C_r\Delta$ and maximum coordinate at most $C\Delta$. The last level r episode in forward time also gives

$$h_B(p_{\text{old gate},T-1}) \ge c_r \Delta. \tag{108}$$

We now compute the length. There are at most $C \log \Delta$ episodes, and each has length at most $C_r(\log \Delta)^{r-1}$. Hence

$$T \le C(\log \Delta)^r.$$

For the lower bound, (106) gives

$$E_\ell \ge \Delta \Lambda_r^{-\ell}.$$

For every integer

$$0 \le \ell \le \frac{\log \Delta}{2 \log \Lambda_r},$$

we have $E_\ell \ge \sqrt{\Delta}$. These episodes all occur when Δ is large. Each one has length at least

$$c_r \left(\frac{1}{2} \log \Delta \right)^{r-1}.$$

There are at least a fixed positive multiple of $\log \Delta$ such episodes. Therefore

$$T = \Theta((\log \Delta)^r). \tag{109}$$

Adding the new gate along the old orbit. Apply Lemma 11 to the old orbit with coupling vector

$$b = \varepsilon e_{r+1}$$

and terminal new coordinate Y from (105). By (108),

$$b^\top h_B(p_{T-1}) \ge \varepsilon c_r \Delta.$$

Also $|\log Y| = O(\log E) = O(\log \Delta)$ because $\Delta = \Theta(E)$. The length, mass, maximum coordinate, and coordinate range assumptions of Lemma 11 have all been verified. Thus the enlarged orbit exists exactly. The new gate is exponentially small before its final activation, and every old logarithmic coordinate changes by at most $\exp(-c\Delta)$.

Let D_{new} be the initial depth of the new gate. Formula (70) and (107) give

$$D_{\text{new}} = \varepsilon S_g - \log Y + O(\exp(-c\Delta)).$$

By (105), for large E,

$$\frac{\varepsilon E}{2} \le \Delta \le 2\varepsilon E.$$

Therefore

$$D_{\mathrm{new}} \geq \frac{c_r \varepsilon^2}{2} E - O(\log E)$$

and

$$D_{\mathrm{new}} \leq 4 C_r \varepsilon^2 E + O(\log E).$$

The coupling ε was chosen after C_r was fixed. By (95), the upper coefficient is smaller than $1/2$. After increasing the lower threshold on E, the logarithmic term is absorbed into the linear term. Hence there is a constant $\Lambda_{r+1} \geq 4$ such that

$$\frac{E}{\Lambda_{r+1}} \leq D_{\mathrm{new}} \leq \frac{E}{2}. \tag{110}$$

Since $\log \Delta = \log E + O(1)$, (109) gives length $\Theta((\log E)^r)$ after adding the last update. Its gate coefficient is $h_B(Y) = \Theta(E)$. All earlier new gate coefficients are exponentially small. The old total mass is $O(\Delta) = O(E)$, so the total mass at the new level is $O(E)$. The maximum coordinate is also $O(E)$ by the old coordinate bound and $Y = E + O(\log E)$.

The old initial bounded profile lies a fixed positive distance inside Q_r^0. Its initial gate depth lies in $(A, \Lambda_r A]$, whose closure is strictly inside I_0. The logarithmic correction therefore keeps the new initial bounded profile in a fixed compact subset of the interior of Q_{r+1}^0. The coordinate ranges are preserved by the strict bounds chosen above. Finally, append the exact update from (100) through (102). Its input is the uncorrected terminal state of the extended orbit, so the two pieces meet exactly. This adds one update and proves (94) at level $r+1$. All constants and compact sets are fixed before its new lower threshold on E is chosen. This completes the induction and proves Lemma 8 in its precise form (92) through (94).

Step 4. Concatenating the final level. For $k \geq 2$, take the matrix K_k. Choose a fixed bounded terminal profile and set the terminal gate depth to

$$E_0 = B^{1/4}.$$

Repeatedly apply the level k episode backward until the depth reaches a fixed constant scale. By (92), the depths decrease by fixed factors. The same calculation used in (109), now with episode length $\Theta((\log E)^{k-1})$, gives a total length

$$T = \Theta((\log B)^k). \tag{111}$$

The geometric sum of the terminal depths and (94) give

$$\sum_{t<T} \sum_i h_B(p_{i,t}) = O(B^{1/4}), \qquad \max_{0 \leq t \leq T,\, i} p_{i,t} = O(B^{1/4}). \tag{112}$$

The starting depth stays between fixed positive bounds, and the bounded profile belongs to a fixed compact subset of the positive orthant. Their product gives the fixed compact set $\mathcal{P}$ in the proposition, independently of B.

For $k = 1$, take the pair construction from Step 1 with $X = B^{1/4}$. Equations (77) and (78) give the same conclusions with length $\Theta(\log B)$.

Step 5. A late strict increase. We use the scaled loss L_B from (58).

Assume first that $k \geq 2$. Consider the final level k episode in the concatenation above. Its terminal depth is $E = B^{1/4}$. Let t_0 be its starting time. The episode itself has length $O((\log B)^{k-1})$, while the entire orbit has length $\Theta((\log B)^k)$. Therefore

$$t_0 = T - O((\log B)^{k-1}) = \Theta((\log B)^k). \tag{113}$$

At time t_0, the bounded profile stays in a fixed compact set and the gate is exponentially small in E. Hence

$$L_B(p_{t_0}) \leq C. \tag{114}$$

During the same episode, the final gate produces a coefficient at least cE. Since $h_B(x) \leq x$, at some later time t_1 the gate coordinate itself satisfies

$$p_{\mathrm{gate},t_1} \geq cE.$$

Thus

$$L_B(p_{t_1}) \geq B \log\left(1 + \frac{cB^{1/4}}{B}\right).$$

For small $z \geq 0$, $\log(1+z) \geq z/2$. Here $z = cB^{-3/4}$, so for large B,

$$L_B(p_{t_1}) \geq c' B^{1/4}. \tag{115}$$

Equations (114) and (115) show that the scaled loss is much larger at t_1 than at t_0. Telescoping the difference

$$L_B(p_{t_1}) - L_B(p_{t_0}) = \sum_{j=t_0}^{t_1-1} \big(L_B(p_{j+1}) - L_B(p_j)\big)$$

shows that at least one update with $j \geq t_0$ has a strict loss increase. By (113), this update occurs at time $\Theta((\log B)^k)$.

For $k = 1$, use the final update of the pair orbit. The target is

$$X = B^{1/4}, \qquad Y = \frac{X}{a}\exp(-ah_B(X)).$$

Its predecessor satisfies

$$v = aX + O(1), \qquad u \leq \exp(-cX).$$

Since $X = o(B)$,

$$B\log(1 + x/B) = x + O(x^2/B)$$

uniformly for $0 \leq x \leq X$. Therefore

$$L_B(X,Y) - L_B(u,v) = (1-a)X + O(1) + O(X^2/B) > 0$$

for large B. By (77), this final update has index $\Theta(\log B)$. The proposition follows. □

D.5 Realization by gradient descent from the origin

The auxiliary orbit starts from a positive compact set, while the original problem requires gradient descent to start from $w_0 = 0$. We now add one extra direction and use the first gradient update to place the auxiliary coordinates at the required initial state.

Proposition 13. *Let $k = \min\{n-2, d-2\} \geq 1$. There are constants $\bar{\gamma}_k, c_k > 0$ such that, for every sufficiently large η, some $D \in \mathcal{D}_{n,d}(\bar{\gamma}_k)$ satisfies*

$$\sigma_\eta(D) \geq c_k(1 + \log\eta)^k.$$

The stepsize threshold may depend on n and k.

Proof. Let

$$m = k + 1, \qquad s = n - m.$$

Since $k \leq n-2$, we have $s \geq 1$. Let K be the matrix from Proposition 12. It is positive definite and has unit diagonal. Therefore there are linearly independent vectors $\widetilde{v}_1, \ldots, \widetilde{v}_m \in \mathbb{R}^m$ with Gram matrix K. For example, if $K^{1/2}$ is its positive definite square root, take

$$\widetilde{v}_i = K^{1/2} e_i.$$

Then

$$\langle \widetilde{v}_i, \widetilde{v}_j \rangle = K_{ij}, \qquad \|\widetilde{v}_i\| = 1.$$

Choose a fixed $\rho > 0$ sufficiently small and set

$$v_i = \rho \widetilde{v}_i.$$

Let

$$S = \sum_{i=1}^{m} v_i.$$

We choose ρ small enough that

$$\|S\| \le \frac{1}{8}. \tag{116}$$

The Gram matrix of the v_i is $\rho^2 K$.

Recall that the original coefficient scale is $B = \eta/n$. Define

$$\widehat{B} = \rho^2 B.$$

Construct the orbit in the proof of Proposition 12, with its length and mass estimates (111) and (112), using parameter $\widehat{B}$. For $k = 1$, use (77) and (78) instead. Denote the orbit by $\widehat{p}_t$ and rescale its coordinates by

$$p_t^* = \frac{\widehat{p}_t}{\rho^2}.$$

Because

$$h_{\widehat{B}}(\widehat{p}) = \frac{\widehat{p}}{1 + \widehat{p}/(\rho^2 B)} = \rho^2 h_B(\widehat{p}/\rho^2),$$

we have

$$p_{t+1}^* = p_t^* \odot \exp\bigl(-\rho^2 K h_B(p_t^*)\bigr).$$

Thus p_t^* is an exact isolated coefficient orbit for the vectors v_i at the original scale B.

The scaling also preserves the loss structure up to a fixed constant. Indeed,

$$L_{\widehat{B}}(\widehat{p}) = \widehat{B} \sum_i \log\left(1 + \frac{\widehat{p}_i}{\widehat{B}}\right) = \rho^2 B \sum_i \log\left(1 + \frac{p_i^*}{B}\right) = \rho^2 L_B(p^*).$$

Since ρ is fixed, a bounded loss state for the $\widehat{B}$ orbit remains bounded for the rescaled orbit, and a pulse of order $\widehat{B}^{1/4}$ becomes a pulse of order $B^{1/4}$ up to a fixed multiplicative constant.

Its length is $\Theta((\log B)^k)$. Its total coefficient mass and its largest coordinate are $O(B^{1/4})$. Its initial point remains in a fixed compact subset of the positive orthant. In particular,

$$\mu_i := \log \frac{B}{p_{i,0}^*} = \log B + O(1). \tag{117}$$

Let V be the $m \times m$ matrix whose columns are the v_i. Since these vectors are linearly independent, V is invertible. Choose $u_B \in \mathbb{R}^m$ so that

$$V^\top u_B = \frac{2}{B}(\mu_1, \dots, \mu_m)^\top. \tag{118}$$

By (117),

$$\|u_B\| \le \frac{C \log B}{B}. \tag{119}$$

We now define the actual data in $\mathbb{R}^{m+1}$. For $1 \le i \le m$, let

$$x_i = (v_i, 0).$$

Use the same additional vector for all remaining samples:

$$y_B = \left(\frac{u_B - S}{s}, \frac{1}{2}\right), \qquad x_{m+1} = \cdots = x_n = y_B. \tag{120}$$

Since $m + 1 = k + 2 \le d$, we append zero coordinates if necessary.

We first check the norm condition. By (116) and (119), for large B,

$$\|u_B - S\| \le \frac{1}{4}.$$

Therefore

$$\|y_B\|^2 = \frac{\|u_B - S\|^2}{s^2} + \frac{1}{4} \le \frac{1}{16} + \frac{1}{4} < 1.$$

The vectors x_i with $i \le m$ also have norm at most one because $\rho \le 1$.

Next we construct a separator with a margin independent of B. Let

$$v = (V^\top)^{-1}\mathbf{1}.$$

Then

$$\langle v, v_i \rangle = 1$$

for every i. Set

$$w_{\mathrm{aux}} = \frac{v}{\|v\|}, \qquad \gamma_{\mathrm{aux}} = \frac{1}{\|v\|}.$$

Thus

$$\langle w_{\mathrm{aux}}, v_i \rangle = \gamma_{\mathrm{aux}}.$$

Define

$$w^\star = \left(\frac{1}{2} w_{\mathrm{aux}}, \frac{\sqrt{3}}{2} \right).$$

This is a unit vector. For $i \le m$,

$$\langle w^\star, x_i \rangle = \frac{\gamma_{\mathrm{aux}}}{2}.$$

For the copied vector y_B,

$$\langle w^\star, y_B \rangle = \frac{1}{2s} \langle w_{\mathrm{aux}}, u_B - S \rangle + \frac{\sqrt{3}}{4}.$$

By taking ρ small enough first, and then B large enough, the absolute value of the first term is at most $1/8$. Hence

$$\langle w^\star, y_B \rangle \ge -\frac{1}{8} + \frac{\sqrt{3}}{4} > \frac{1}{4}.$$

Therefore every sample has margin at least

$$\bar{\gamma}_k = \min \left\{ \frac{\gamma_{\mathrm{aux}}}{2}, \frac{1}{4} \right\} > 0. \tag{121}$$

This constant depends only on the fixed auxiliary construction.

We now check the first gradient update. At $w_0 = 0$, every logistic coefficient equals $B/2$. Therefore

$$\begin{aligned} w_1 &= \frac{B}{2} \left(\sum_{i=1}^{m} (v_i, 0) + s y_B \right) \\ &= \frac{B}{2} \left(S + u_B - S, \frac{s}{2} \right) \\ &= \left(\frac{B}{2} u_B, \frac{Bs}{4} \right). \end{aligned}$$

For $i \le m$, (118) gives

$$\langle w_1, x_i \rangle = \frac{B}{2} \langle u_B, v_i \rangle = \mu_i.$$

Hence

$$B \exp\bigl(-\langle w_1, x_i \rangle\bigr) = p^*_{i,0}.$$

Thus the auxiliary coefficient state is hit exactly after the first update.

The copied setup samples must remain negligible afterwards. Their margin at w_1 is

$$\langle w_1, y_B \rangle = \frac{B}{2s} \langle u_B, u_B - S \rangle + \frac{Bs}{8}.$$

By (119), the first term is $O(\log B)$ uniformly over $s \ge 1$. Therefore

$$\langle w_1, y_B \rangle \ge \frac{B}{16} \tag{122}$$

for all sufficiently large B.

Let

$$z_t = \log p_t^*$$

be the logarithmic state of the isolated auxiliary orbit. Let $\widetilde{z}_t$ be the true logarithmic auxiliary coefficient state at the iterate w_{1+t}. Define

$$M_t = \langle w_{1+t}, y_B \rangle, \qquad g_t = B\mathrm{e}^{-M_t}, \qquad \epsilon_t = \|\widetilde{z}_t - z_t\|.$$

We have $\epsilon_0 = 0$.

Assume temporarily that $\epsilon_t \le 1$ on an initial time interval. Then each true auxiliary coefficient differs from its reference value by at most a factor e. Since $h_B(\mathrm{e}x) \le \mathrm{e}h_B(x)$,

$$\sum_{t<U} \sum_{i=1}^{m} h_B(\mathrm{e}^{\widetilde{z}_{i,t}}) \le C_k B^{1/4} \tag{123}$$

for every prefix of length U inside this interval.

The setup margin evolves according to

$$M_{t+1} - M_t = \sum_{i=1}^{m} h_B(\mathrm{e}^{\widetilde{z}_{i,t}}) \langle x_i, y_B \rangle + s h_B(g_t) \|y_B\|^2.$$

The last term is nonnegative. Also $|\langle x_i, y_B \rangle| \le 1$. Using (122) and (123),

$$M_t \ge \frac{B}{16} - C_k B^{1/4} \ge \frac{B}{32}$$

for all sufficiently large B. Consequently

$$h_B(g_t) \le g_t \le B\mathrm{e}^{-B/32}. \tag{124}$$

The isolated logarithmic update is

$$\Phi_{B,\rho^2 K}(z) = z - \rho^2 K h_B(\mathrm{e}^z).$$

The true auxiliary update contains one additional term from the s copies of y_B. Its ith component has magnitude

$$s|\langle x_i, y_B \rangle| h_B(g_t).$$

Because

$$s\langle x_i, y_B \rangle = \langle v_i, u_B - S \rangle,$$

this factor is bounded independently of s. Hence the additional update vector has norm at most

$$C_k B\mathrm{e}^{-B/32}. \tag{125}$$

On the unit ball around the reference orbit, all auxiliary coordinates are $O(B^{1/4})$. Since $t_B(x) \le x$, the derivative of the isolated logarithmic map satisfies

$$\|D\Phi_{B,\rho^2 K}\| \le 1 + C_k B^{1/4}.$$

Combining this with (125),

$$\epsilon_{t+1} \le (1 + C_k B^{1/4})\epsilon_t + C_k B\mathrm{e}^{-B/32}. \tag{126}$$

The auxiliary orbit has length at most $C_k(\log B)^k$. Iterating (126) from $\epsilon_0 = 0$ gives

$$\max_{t \le T} \epsilon_t \le C_k B T (1 + C_k B^{1/4})^T \mathrm{e}^{-B/32}.$$

Taking logarithms of the prefactor shows that it contributes only

$$O((1 + \log B)^{k+1}),$$

which is negligible compared with $B/32$. Hence

$$\max_{t \le T} \epsilon_t \le \mathrm{e}^{-B/64} \tag{127}$$

for large B.

This also justifies the temporary assumption $\epsilon_t \le 1$. If there were a first time at which the error exceeded one, the argument above would apply to the entire preceding prefix and would imply that the error at that first exit time is at most $\mathrm{e}^{-B/64} < 1$, a contradiction.

It remains to transfer the late loss increase. For $k \ge 2$, let $t_0 < t_1$ be the bounded state and the later pulse state from the auxiliary construction. We have

$$t_0 \ge c_k(\log B)^k, \qquad L_B(p^*_{t_0}) \le C_k, \qquad L_B(p^*_{t_1}) \ge c_k B^{1/4}.$$

The logarithmic error (127) changes the auxiliary scaled loss by $o(1)$ at the bounded state and by $o(B^{1/4})$ at the pulse state. To see this, note that

$$\frac{\partial}{\partial \log p}\left[B \log\left(1 + \frac{p}{B}\right)\right] = h_B(p) \le p,$$

and all reference coordinates are $O(B^{1/4})$.

The copied setup samples contribute at most

$$sB\mathrm{e}^{-B/32} = o(1)$$

to the scaled loss at every time under consideration. Therefore

$$\eta F(w_{1+t_0}) \le C_k + o(1), \qquad \eta F(w_{1+t_1}) \ge c_k B^{1/4} + o(B^{1/4}).$$

For large B, the second quantity is strictly larger. Telescoping shows that some update with index at least $1 + t_0$ strictly increases the true loss. Hence

$$\sigma_\eta(D) \ge c_k(\log B)^k.$$

When $k = 1$, the auxiliary pair orbit has a final scaled loss increase of order $B^{1/4}$. The error estimate (127) changes it by only $o(B^{1/4})$, and the setup samples contribute $o(1)$. Thus the same strict increase occurs for the true gradient descent trajectory at an index of order $\log B$.

Finally, n is fixed and $B = \eta/n$. Therefore

$$\log B = \log \eta - \log n$$

and for all sufficiently large η,

$$\log B \ge \frac{1}{4}(1 + \log \eta).$$

This proves the proposition. □

D.6 Completion of the lower bound

It remains to handle $k = 0$, which means $n = 2$ or $d = 2$. We only need a constant lower bound on the transition time.

Choose

$$u = \frac{1}{4}, \qquad v = \frac{3}{4}.$$

Let $g_n > 0$ be small enough that

$$ng_n^2 \le \frac{1}{32}.$$

In two dimensions, define

$$x_1 = (u, g_n), \qquad x_2 = \cdots = x_n = (-v, g_n), \qquad w^\star = (0, 1).$$

After decreasing g_n if necessary, all sample norms are at most one. The vector $w^\star$ separates every sample with margin g_n.

At the origin every logistic coefficient equals $B/2$, so

$$w_1 = \frac{B}{2}\sum_{j=1}^{n} x_j.$$

For the first sample,

$$\begin{aligned}\langle w_1, x_1\rangle &= \frac{B}{2}\left(u^2 - (n-1)uv + ng_n^2\right)\\ &\le \frac{B}{2}\left(-\frac{1}{8} + \frac{1}{32}\right)\\ &= -\frac{3B}{64}.\end{aligned}$$

Hence

$$F(w_1) \ge \frac{1}{n}\log\left(1 + \exp(3B/64)\right).$$

For large B, this is larger than

$$F(w_0) = \log 2.$$

Thus the first update strictly increases the loss and

$$\sigma_\eta(D) \ge 1.$$

If $d > 2$, append zero coordinates.

For $k \ge 1$, Proposition 13 gives a margin $\bar{\gamma}_k > 0$ that does not depend on B, on the number of copied setup samples, or on the ambient dimension. For a fixed sample size n, only the finitely many values

$$k = 0, 1, \ldots, n-2$$

can occur. For $n = 2$, set $\gamma_2 = g_2$. For $n \ge 3$, define

$$\gamma_n = \min\{g_n, \bar{\gamma}_1, \ldots, \bar{\gamma}_{n-2}\} > 0.$$

After taking the largest of the finitely many stepsize thresholds, the construction gives, uniformly over every $d \ge 2$,

$$\sup_{D \in \mathcal{D}_{n,d}(\gamma_n)} \sigma_\eta(D) \ge c_n(1 + \log\eta)^{\min\{n-2, d-2\}}.$$

If $0 < \gamma \le \gamma_n$, then

$$\mathcal{D}_{n,d}(\gamma_n) \subseteq \mathcal{D}_{n,d}(\gamma),$$

so the same lower bound holds for $\mathcal{D}_{n,d}(\gamma)$. This is the lower bound required in Theorem 1.

E Proof of Theorem 1

Proof. Fix $n \ge 2$ and $0 < \gamma \le \gamma_n$, where γ_n is chosen in Appendix D.6. The lower construction gives, uniformly for all $d \ge 2$ and all sufficiently large η,

$$\sup_{D \in \mathcal{D}_{n,d}(\gamma)} \sigma_\eta(D) \ge c_n(1 + \log\eta)^{\min\{n-2, d-2\}}.$$

On the other hand, Appendix C proves for every fixed $0 < \gamma \le 1$ that

$$\tau_\eta(D) \le C_{n,\gamma}(1 + \log\eta)^{\min\{n-2, d-2\}}$$

uniformly over $D \in \mathcal{D}_{n,d}(\gamma)$ and $d \ge 2$. Lemma 4 gives $\sigma_\eta(D) \le \tau_\eta(D)$, so the same expression is an upper bound for $\sigma_\eta(D)$. Taking the supremum over D proves Theorem 1. All constants and the stepsize threshold depend only on n and γ, not on d. □

F Proofs of the corollaries

F.1 Proof of Corollary 2

Proof. Fix $n \ge 2$ and $0 < \gamma \le \gamma_n$. Lemma 4 gives $\sigma_\eta(D) \le \tau_\eta(D)$ for each dataset. Together with (2), this yields

$$\sup_{D \in \mathcal{D}_{n,d}(\gamma)} \sigma_\eta(D) \le \sup_{D \in \mathcal{D}_{n,d}(\gamma)} \tau_\eta(D) \le C_{n,\gamma}(1 + \log\eta)^{\min\{n-2, d-2\}}.$$

Theorem 1 gives the matching lower bound for the first supremum. Both constants and the common stepsize threshold are independent of d, as required. □

F.2 Proof of Corollary 3

Proof. We first record a simple decay estimate after the threshold time. Suppose $F(w_t) \le 1/(8\eta)$ and η is sufficiently large depending only on n. Write

$$\ell_{i,t} = \log(1 + \exp(-\langle w_t, x_i\rangle)), \qquad q_{i,t} = \frac{1}{1 + \exp(\langle w_t, x_i\rangle)} = 1 - \exp(-\ell_{i,t}).$$

Since $\ell_{i,t} \le nF(w_t) \le n/(8\eta) \le 1$, the elementary bound $1 - \exp(-x) \ge x/2$ for $0 \le x \le 1$ gives

$$-\langle \nabla F(w_t), w^\star\rangle = \frac{1}{n}\sum_i q_{i,t}\langle w^\star, x_i\rangle \ge \frac{\gamma}{n}\sum_i q_{i,t} \ge \frac{\gamma}{2}F(w_t). \tag{128}$$

Hence $\|\nabla F(w_t)\| \ge \gamma F(w_t)/2$.

Since $\eta F(w_t) \le 1/8$, we have $1 - \eta F(w_t)/2 \ge 15/16$. Substituting the gradient lower bound into (1) gives

$$\log \frac{F(w_{t+1})}{F(w_t)} \le -\frac{15\gamma^2}{64}\eta F(w_t).$$

All losses are positive. Exponentiating, taking reciprocals, and using $\exp(x) \ge 1 + x$ gives

$$\frac{1}{F(w_{t+1})} \ge \frac{\exp(15\gamma^2\eta F(w_t)/64)}{F(w_t)} \ge \frac{1}{F(w_t)} + \frac{15\gamma^2}{64}\eta. \tag{129}$$

Lemma 4 ensures that the threshold condition continues to hold. Sum (129) over the s updates starting at τ_η and discard the positive initial reciprocal. For every integer $s \ge 1$, we obtain

$$F(w_{\tau_\eta + s}) \le \frac{64}{15\gamma^2\eta s}. \tag{130}$$

Now fix $n \ge 3$ and $0 < \gamma \le 1$. The upper bound in Appendix C supplies constants C_0 and η_0, uniform in d, such that

$$\tau_\eta(D) \le C_0(1 + \log\eta)^k, \qquad k = \min\{n-2, d-2\} \ge 1.$$

Increase η_0 to at least $\max\{\mathrm{e}, n/8\}$ so that the preceding decay estimate applies. Choose

$$a = \min\left\{1, \frac{1}{2^{n-1}C_0}\right\}.$$

Choose an integer $T_0 \ge 1$ with $aT_0^{1/(n-2)} \ge \log\eta_0$. For $T \ge T_0$ and $1 \le k \le n-2$, set $\eta = \exp(aT^{1/k})$. Then $aT^{1/k} \ge aT^{1/(n-2)} \ge \log\eta_0 \ge 1$, so $\eta \ge \eta_0$ and

$$\tau_\eta(D) \le C_0(2aT^{1/k})^k \le C_0 2^{n-2}a^k T \le \frac{T}{2}.$$

Applying (130) with $s = T - \tau_\eta(D) \ge T/2$ yields

$$F(w_T) \le \frac{128}{15\gamma^2\eta T} \le \frac{C}{T}\exp(-aT^{1/k}),$$

which proves Corollary 3 with $C = 128/(15\gamma^2)$. The choices of a, C, and T_0 depend only on n and γ. □