ON THE TWO FACES OF ADAM IN SEPARABLE LINEAR CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the behavior of deterministic, full-batch, bias-corrected Adam in separable linear classification with softmax parametrization under log-loss. In this setting, under a wide range of conditions Adam is known to approach max-norm-margin optimality when its stability constant ϵ is zero, while with a positive ϵ, it is known to approach Euclidean-margin optimality. Our main contribution is the quantitative description of Adam's behavior for small fixed positive ϵ. We give sufficient conditions under which an Adam-trained classifier nearly maximizes the max-norm margin before the updates become gradient-like. We also show that the classifier reaches a fixed target Euclidean margin only much later. Specifically, we show that for polynomially decreasing stepsizes with exponent a, where $1/3 < a < 1$, the updates become approximately proportional to the negative gradient after $\Theta(\log(1/\epsilon)^{1/(1-a)})$ iterations. At that time, the classifier still nearly maximizes the max-norm margin. Reaching a fixed target Euclidean margin above that of every max-norm-optimal classifier, but below the optimum, is shown to require $\epsilon^{-\Theta(1)/(1-a)}$ iterations. Under inverse-linear stepsize decay ($a = 1$), the update transition takes polynomially many iterations, whereas reaching the target margin takes exponentially many. Experiments support these predictions. The later change in the classifier can improve or worsen generalization after training error reaches zero, connecting the analysis to grokking and its reverse.

1 INTRODUCTION

Adam's implicit bias in separable linear classification depends on its stability parameter ϵ. Existing analyses show that removing this parameter leads to max-norm-margin maximization (Zhang et al., 2024), whereas keeping it positive ultimately leads to Euclidean-margin maximization (Wang et al., 2022). These results describe different limiting behaviors of the same algorithm. We study how a trajectory with a small positive ϵ evolves between these behaviors, and how the learning-rate schedule affects this evolution.

This qualitative behavior, which has already been observed empirically (Fan et al., 2025; Tsilivis et al., 2025), is not unexpected: If ϵ is sufficiently small, the gradients initially dominate it, and Adam's updates are approximately *sign-like*: They are nearly aligned with the sign function applied componentwise to the gradient. When gradients are small relative to ϵ, the updates are *gradient-like*: They are aligned, up to scaling, with the actual gradient. Now, it is also true that the updates cannot be gradient-like unless the gradients are small. The latter can only happen when the parameter vector has a large magnitude. But because initially the updates are sign-like, at this time, the parameter vector needs to be nearly ℓ_∞-margin optimal. After the updates become gradient-like, it then takes some extra time for the parameter vector to get aligned with the ℓ_2-max-margin solution direction.

Our contribution is to quantify this intuitive behavior for deterministic, full-batch, bias-corrected Adam in separable linear classification, using logistic loss for binary classification, with an extension to multiclass softmax cross-entropy. Our analysis applies to polynomially decreasing stepsizes

$$\eta_t = \eta_0(t+1)^{-a}, \qquad a \in (1/3, 1],$$

a family studied in previous implicit-bias analyses (Nacson et al., 2019; Zhang et al., 2024; Fan et al., 2025). Each trajectory uses a fixed positive ϵ; our results characterize the family of trajectories as ϵ decreases. We bound the max-norm margin gap at the time marking permanent entry into a

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.

gradient-like regime, which we call the *update transition time*. We also characterize when this update transition occurs and how much additional training is needed for the classifier to reach a prescribed Euclidean-margin accuracy.

At a high level, our main results are as follows. For $1/3 < a < 1$, the update transition time satisfies

$$T_{\text{update}} = \Theta\Big((\log(1/\epsilon))^{1/(1-a)}\Big).$$

At this time, the classifier's max-norm margin gap is small, with a bound that vanishes as ϵ decreases. We also bound its subsequent deterioration: the classifier remains nearly max-norm-margin optimal while the accumulated movement after T_{update} is small relative to the weight norm at that time. Let $\Delta > 0$ be the smallest Euclidean-margin gap among max-norm-optimal classifiers. For any fixed target gap $\zeta \in (0, \Delta)$, the first subsequent time at which the Euclidean-margin gap falls below ζ satisfies

$$T_{\text{Euclidean}} = \epsilon^{-\Theta(1)/(1-a)}.$$

Thus, the update transition occurs on a polylogarithmic scale in $1/\epsilon$, while reaching this Euclidean-margin accuracy requires a polynomial scale. For $a = 1$, these scales become polynomial and exponential, respectively.

There is an intuitive explanation for these bounds. The initial phase produces a weight vector with norm of order $\log(1/\epsilon)$. Changing the normalized parameter vector by a fixed amount requires displacement comparable to this norm: a much smaller displacement barely changes its direction. After the transition, however, the accumulated distance traveled grows only logarithmically in the cumulative stepsize, as in gradient descent on separable logistic regression (Soudry et al., 2018). Consequently, substantial classifier change requires the cumulative stepsize to be at least polynomial in $1/\epsilon$. Cumulative stepsizes grow polynomially in the iteration count when $a < 1$, but only logarithmically when $a = 1$. This explains the particularly long transition times under harmonic stepsize decay.

These results clarify the relationship between the two existing descriptions of Adam's implicit bias. They predict, for example, that inverse-square-root ($a = 1/2$) and inverse-linear ($a = 1$) stepsize decay lead to the same eventual margin geometry, but reaching a fixed Euclidean-margin accuracy takes polynomially many iterations in $1/\epsilon$ under the former and exponentially many under the latter. Depending on the data distribution, the change in the classifier can improve or worsen generalization after the training examples are already classified correctly. This connects the trajectory description to grokking and its reverse: delayed improvement or deterioration in predictive performance associated with changing implicit bias (Lyu et al., 2024).

The guarantees hold for finite positive ϵ under data- and algorithm-dependent sufficient conditions, while the scaling statements describe their behavior as ϵ decreases. We leave for future work determining the polynomial exponent in the bound for $T_{\text{Euclidean}}$ for the case when $a < 1$. Similarly, the sufficient conditions are conservative; sharpening them to characterize, for a given dataset and optimizer parameters, when the described phase behavior occurs is an important direction for future work. The theory and accompanying experiments provide a more detailed understanding of how Adam's stability parameter and learning-rate schedule determine its implicit-bias trajectory.

2 Problem Setup and Preliminaries

Let $\{(x_i, y_i)\}_{i=1}^n$ be a dataset with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$. Denote $R_1 := \max_{i\in[n]} \|x_i\|_1$ and $R_2 := \max_{i\in[n]} \|x_i\|_2$. Define $z_i := x_i y_i$ and $[n] := \{1, \ldots, n\}$. We study the empirical logistic loss

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(z_i^\top w), \qquad \text{where} \qquad \ell(u) = \log(1 + e^{-u}). \tag{1}$$

Write $g(w) := \nabla L(w)$, $L_t := L(w_t)$, and $g_t := g(w_t)$. Define the unnormalized margin $\rho(w) := \min_{i\in[n]} z_i^\top w$ and the normalized p-margin $\widehat{\gamma}_p(w) := \rho(w)/\|w\|_p$ for $p \in \{2, \infty\}$. The optimal max-norm and Euclidean data margins are

$$\gamma_\infty := \max_{\|u\|_\infty \le 1} \min_{i\in[n]} z_i^\top u, \qquad \gamma_2 := \max_{\|u\|_2 \le 1} \min_{i\in[n]} z_i^\top u. \tag{2}$$

We analyze deterministic, bias-corrected Adam with a fixed stability constant $\epsilon > 0$ (Kingma & Ba, 2015). Initialize $\bar{m}_{-1} := 0$ and $\bar{v}_{-1} := 0$, and choose $\beta_1, \beta_2 \in [0, 1)$. The iterates are

$$\begin{aligned} \bar{m}_t &= \beta_1 \bar{m}_{t-1} + (1-\beta_1) g_t, \qquad \bar{v}_t = \beta_2 \bar{v}_{t-1} + (1-\beta_2) g_t^2, \\ m_t &= \bar{m}_t / (1-\beta_1^{t+1}), \qquad v_t = \bar{v}_t / (1-\beta_2^{t+1}), \\ w_{t+1} &= w_t - \eta_t D_t, \qquad D_t := m_t / (\sqrt{v_t} + \epsilon \mathbf{1}), \end{aligned} \tag{3}$$

where all operations are entrywise. The default PyTorch choice for ϵ is 10^{-8} (Paszke et al., 2019).

We assume that the data is separable: there exists $w \in \mathbb{R}^d$ such that $\min_{i \in [n]} z_i^\top w > 0$, which is a standard assumption in the analyses of optimization implicit bias (Soudry et al., 2018; Ji & Telgarsky, 2019; Nacson et al., 2019; Wu et al., 2023; Zhang et al., 2024). Thus, both $\gamma_\infty > 0$ and $\gamma_2 > 0$ hold.

We use the decreasing stepsize schedule $\eta_t = \eta_0 (t+1)^{-a}$, where $\eta_0 > 0$ and $a \in (1/3, 1]$, as in prior analyses of Adam and steepest descent (Nacson et al., 2019; Zhang et al., 2024; Fan et al., 2025; Baek et al., 2026; Li et al., 2026). In addition to satisfying $\sum_{t=0}^{\infty} \eta_t = \infty$, this schedule also satisfies, for every $q \in (0,1)$ and $c > 0$, $\sum_{r=0}^{t} q^r [\exp(c \sum_{s=t-r}^{t-1} \eta_s) - 1] \le C_{\mathrm{geo}} \eta_t, \ t \ge t_0$, where $t_0 = t_0(q, c, a, \eta_0)$ and $C_{\mathrm{geo}} = C_{\mathrm{geo}}(q, c, a, \eta_0)$ (Zhang et al., 2024); see Proposition 9 in App. B.

For the analysis, we use the gradient-scale proxy introduced by Zhang et al. (2024): $G(w) := -\frac{1}{n} \sum_{i=1}^{n} \ell'(z_i^\top w)$. It controls the gradient norms through the data margins:

$$\gamma_\infty G(w) \le \|g(w)\|_1 \le R_1 G(w), \qquad \gamma_2 G(w) \le \|g(w)\|_2 \le R_2 G(w).$$

Write $G_t := G(w_t)$. We can control the bias-corrected moment errors in terms of G_t: $\max\{\|m_t - g_t\|_1, \|\sqrt{v_t} - |g_t|\|_1\} \lesssim \eta_t G_t$ after some t_{ema} burn-in period (as shown in Lemmas 15 and 16 in App. B). Thus, the bias-corrected moments track g_t and $|g_t|$ at scale $O(\eta_t G_t)$. In the idealized case where $m_t[j] \approx g_t[j]$ and $\sqrt{v_t[j]} \approx |g_t[j]|$, the update is approximated by the soft-sign map $\sigma_\epsilon(g)[j] := \frac{g[j]}{|g[j]| + \epsilon}$, $j \in [d]$. This map satisfies (see Lemma 18 in App. B)

$$\langle g, \sigma_\epsilon(g) \rangle = \sum_j \frac{g[j]^2}{|g[j]| + \epsilon} \ge \|g\|_1 - d\epsilon, \qquad \langle g, \sigma_\epsilon(g) \rangle \ge \frac{\|g\|_1^2}{\|g\|_1 + d\epsilon}. \tag{4}$$

The first inequality is informative when $\epsilon \ll \|g\|_1$, whereas the second applies throughout the trajectory. Both are useful for studying Adam's trajectories. However, we must still control the discrepancy between D_t and $\sigma_\epsilon(g_t)$ for any $\epsilon > 0$, as shown in Lemma 17 in App. B:

$$\max\{\|g_t \odot (D_t - \sigma_\epsilon(g_t))\|_1, \epsilon \|D_t - \sigma_\epsilon(g_t)\|_2, |\langle g_t, D_t - \sigma_\epsilon(g_t) \rangle|\} \le C_{\mathrm{ema}} \eta_t G_t, \ t \ge t_{\mathrm{ema}}.$$

Combining this with a Taylor expansion yields the basic loss recursion (see Proposition 19 in App. B)

$$L_{t+1} \le L_t - \eta_t \langle g_t, \sigma_\epsilon(g_t) \rangle + O(\eta_t^2 G_t), \qquad t \ge t_{\mathrm{ema}}.$$

Margin analysis additionally requires a bound on the accumulated displacement. The unit coefficient on the accumulated step size in the next result is essential: applying the triangle inequality using only the pointwise bound $\|D_t\|_\infty \le C_D := \sqrt{(1-\beta_1)/(1-\beta_2)}$ from Lemma 11 in App. B would not recover the $1/\gamma_\infty$ coefficient in Theorem 4 or the max-norm optimality in Theorem 3 (see below).

Lemma 1. *Suppose that $0 \le \beta_1 \le \beta_2 < 1$. If there exists a constant B_g such that $B_g \ge \|g(w)\|_\infty$ for every $w \in \mathbb{R}^d$, then the following holds for every integer r with $0 \le r < T$ and $\epsilon > 0$:*

$$\left\| \sum_{t=r}^{T-1} \eta_t D_t \right\|_\infty \le \sum_{t=r}^{T-1} \eta_t + C_{\mathrm{rad}} \eta_r \left[1 + \log \frac{B_g^2 + \epsilon^2}{(1-\beta_2)\epsilon^2} \right], \tag{5}$$

where the constant C_{rad} depends only on β_1 and β_2.

For $\epsilon = 0$, Zhang et al. (2024) obtained a similar bound under the additional condition $|g_0[j]| \ge \rho$ for every $j \in [d]$ and some $\rho > 0$, together with a ρ-dependent burn-in time. Because $\epsilon > 0$ keeps the update well defined, Lemma 1 permits arbitrary parameter initialization w_0, applies to every interval $[r, T)$, and accounts for bias correction. The subtle point is controlling momentum carried into $[r, T)$ while keeping the leading coefficient 1; see App. D for details. Moreover, Lemmas 1 and 11 rely on the condition $\beta_1 \le \beta_2$, which is also assumed in prior analyses of Adam's implicit bias (Zhang et al.,

2024; Baek et al., 2026). We assume that $0 \le \beta_1 \le \beta_2 < 1$ throughout the following sections. It remains open whether our results extend beyond this condition.

Next, we introduce the gradient and sign residuals:

$$\lambda_t := \frac{G_t}{\epsilon}, \qquad r_t^{\mathrm{sgn}} := \frac{\|g_t \odot (D_t - \mathrm{sign}(g_t))\|_1}{\|g_t\|_1}, \qquad r_t^{\mathrm{grad}} := \frac{\|\epsilon D_t - g_t\|_2}{\|g_t\|_2}.$$

The order parameter λ_t separates the two regimes. Under the heuristic $D_t \approx \sigma_\epsilon(g_t)$, the bound $\|g_t\|_1 \le R_1 G_t$ gives $|g_t[j]|/\epsilon \le R_1 \lambda_t$ for every $j \in [d]$. Thus, when $\lambda_t \ll 1$, every coordinate is small relative to ϵ and $D_t \approx g_t/\epsilon$. When $\lambda_t \gg 1$, some coordinates may remain small, so coordinatewise closeness to $\mathrm{sign}(g_t)$ need not hold. We instead measure sign-likeness through relative first-order loss decrease. Using (4) and $\|g_t\|_1 \ge \gamma_\infty G_t = \gamma_\infty \epsilon \lambda_t$, we have $1 - \frac{\langle g_t, \sigma_\epsilon(g_t)\rangle}{\|g_t\|_1} \le \frac{d}{\gamma_\infty \lambda_t}$. Hence, as $\lambda_t \to \infty$, the soft-sign direction achieves the same first-order loss decrease as sign descent. This motivates r_t^{sgn}, since $\left|\frac{\langle g_t, D_t\rangle}{\|g_t\|_1} - 1\right| \le r_t^{\mathrm{sgn}}$, whereas r_t^{grad} directly measures the discrepancy between D_t and g_t/ϵ. See Proposition C.1 in App. C for a general characterization of these residuals.

Finally, we establish an ϵ-uniform reference time T_* after which every trajectory lies in the low-loss regime. Proposition 2.1 shows that G_t and L_t are uniformly comparable for $t \ge T_*$, with $G_{T_*} \ge g_* > 0$. Since the crossover thresholds defined below vanish as $\epsilon \downarrow 0$ and $G_t \to 0$, each threshold lies below G_{T_*} for all sufficiently small ϵ and is reached at a finite time after T_*.

Proposition 2.1. *In the same setting as Lemma 1, the loss and the proxy converge to zero. Moreover, there exist a time $T_* > t_{\mathrm{ema}}$ and a constant $g_* > 0$ such that for every $\epsilon \in (0, 1]$, we have $G_{T_*} \ge g_*$, and $L_t \le \log 2/n$ and $1/2 L_t \le G_t \le L_t$ for every $t \ge T_*$.*

3 Main Results

We introduce three stopping times that mark the geometric transition of the update. Recall $\ell_\epsilon := \log(1/\epsilon)$. Fix $b, \delta \in (0, 1)$, and let T_* be the time from Proposition 2.1. For $\epsilon \in (0, 1)$, we define

$$T_1^-(\epsilon; b) := \inf\left\{t \ge T_* : G_t \le \epsilon \ell_\epsilon^b\right\}, \quad T_1^\delta(\epsilon) := \inf\left\{t \ge T_* : G_t \le \delta\epsilon\right\},$$

and

$$T_1^+(\epsilon; b) := \inf\left\{t \ge T_* : G_t \le \epsilon \ell_\epsilon^{-b}\right\}.$$

Here, $\ell_\epsilon^b \to \infty$ and $\ell_\epsilon^{-b} \to 0$ as $\epsilon \downarrow 0$. For simplicity, we write $T_1^- := T_1^-(\epsilon; b)$, $T_1 := T_1^\delta(\epsilon)$, and $T_1^+ := T_1^+(\epsilon; b)$. These stopping times correspond to $\lambda_t = G_t/\epsilon$ first falling below the levels ℓ_ϵ^b, δ, and ℓ_ϵ^{-b}, respectively. We have the ordering $T_1^- < T_1 < T_1^+$ for sufficiently small ϵ, as shown by Proposition C.2 in App. C. Thus, T_1^- and T_1^+ bracket the update crossover, while $T_1 = T_1^\delta(\epsilon)$ is the reference crossover time. Theorem 2 shows that for small ϵ, Adam's local update is sign-like at T_1^- and gradient-like at T_1^+, as indicated by the corresponding small sign and gradient residuals. Moreover, at T_1, the gradient residual is bounded by $O(\delta)$ plus a term that vanishes with ϵ. Corollary 23 in App. E strengthens this conclusion: for suitably chosen δ and sufficiently small ϵ, every $t \ge T_1$ satisfies $r_t^{\mathrm{grad}} \le p$ and $r_t^{\mathrm{sgn}} \ge q$, where $p, q \in (0, 1)$ and $p + q = 1$. Thus, T_1 marks permanent entry into the gradient-like regime.

Theorem 2. *For any $b, \delta \in (0, 1)$, there exists a constant $\tilde{\epsilon}_{\mathrm{res}}(b, \delta)$ such that for every $0 < \epsilon \le \tilde{\epsilon}_{\mathrm{res}}(b, \delta)$, we have*

$$\max\left\{r^{\mathrm{sgn}}_{T_1^-(\epsilon;b)}, r^{\mathrm{grad}}_{T_1^+(\epsilon;b)}\right\} = \begin{cases} O\left(\ell_\epsilon^{-b} + \ell_\epsilon^{-\frac{a}{1-a}}\right), & a \in (1/3, 1), \\ O\left(\ell_\epsilon^{-b}\right), & a = 1. \end{cases}$$

Denote $K_0 = R_1 C_D$. We also have

$$r^{\mathrm{grad}}_{T_1^\delta(\epsilon)} \le \frac{R_1\delta}{1 + R_1\delta} + \begin{cases} O\left(\ell_\epsilon^{-\frac{a}{1-a}}\right), & a \in (1/3, 1), \\ O\left(\epsilon^{\frac{1}{2K_0\eta_0}}\right), & a = 1. \end{cases}$$

We next turn to the classifier geometry, beginning with a few key definitions:

$$\mathcal{M}_\infty := \{u : \|u\|_\infty \le 1, \rho(u) = \gamma_\infty\}, \quad \gamma_{2|\infty} := \max_{u \in \mathcal{M}_\infty} \frac{\rho(u)}{\|u\|_2}, \qquad \Delta_{\text{geom}} := \gamma_2 - \gamma_{2|\infty}.$$

Here, $\gamma_{2|\infty}$ is the largest Euclidean-normalized margin attained by a max-norm-optimal classifier. Hence, $\Delta_{\text{geom}} \ge 0$ measures the separation between the two optimal geometries. The condition $\Delta_{\text{geom}} > 0$ rules out any direction that is optimal under both norm geometries. The following theorem shows that, for small ϵ, the ℓ_∞-margin gap is small at T_1^-, T_1, and T_1^+ despite the change in update geometry, whereas the ℓ_2-margin gap is bounded below by Δ_{geom} up to a vanishing error.

Theorem 3. *Let $\mathcal{T}_\epsilon(b,\delta) := \{T_1^-(\epsilon;b), T_1^\delta(\epsilon), T_1^+(\epsilon;b)\}$. For any $\delta, b \in (0,1)$, there exists a constant $\tilde{\epsilon}_2(\delta,b)$ such that for every $0 < \epsilon \le \tilde{\epsilon}_2(\delta,b)$ and every $T \in \mathcal{T}_\epsilon(b,\delta)$, we have*

$$\gamma_\infty - \widehat{\gamma}_\infty(w_T) \le O(r_\epsilon), \qquad \gamma_2 - \widehat{\gamma}_2(w_T) \ge \Delta_{\text{geom}} - O(r_\epsilon), \qquad r_\epsilon := \begin{cases} \ell_\epsilon^{-a}, & a \in (1/3,1), \\ \log \ell_\epsilon / \ell_\epsilon, & a = 1. \end{cases}$$

We discuss four ingredients in the proof of Theorem 3: **1. How to obtain the constant γ_∞; 2. How to derive the sharp weight-norm bound; 3. How to establish the lower bound on the ℓ_2-margin gap; 4. How to transfer the results at T_1 to T_1^- and T_1^+**; see App. F for details. Since the ℓ_∞-margin is already nearly optimal at the post-crossover time T_1^+, we then ask **how much additional training is required to attain a prescribed Euclidean-margin accuracy**. Importantly, addressing these questions clarifies how the various results of this work fit together.

1. The γ_∞ constant We first discuss why the direct application of the bounds in (4) fails. The first bound in (4) gives $\langle g_t, \sigma_\epsilon(g_t)\rangle \ge \|g_t\|_1 - d\epsilon \ge (\gamma_\infty - d\epsilon/G_t)\, G_t$. If we worked only with T_1, then we have $G_t > \delta\epsilon$ for $t < T_1$, which leads to $d\epsilon/G_t < d/\delta$. Consequently, we only obtain $\langle g_t, \sigma_\epsilon(g_t)\rangle \ge (\gamma_\infty - d/\delta)\, G_t$. The coefficient may be nonpositive and, even when positive, does not converge to γ_∞ as $\epsilon \downarrow 0$. Thus, this estimate cannot recover the sharp bound $S_{T_1} - S_{T_*} \le \ell_\epsilon/\gamma_\infty +$ lower-order terms. We could use the other result in (4): $\langle g_t, \sigma_\epsilon(g_t)\rangle \ge \kappa_\delta G_t$ where $\kappa_\delta := \gamma_\infty^2/(R_1 + d/\delta)$. However, κ_δ can be much smaller than γ_∞, yielding the leading term $\ell_\epsilon/\kappa_\delta$ instead of the sharp term $\ell_\epsilon/\gamma_\infty$.

To obtain a refined analysis of the trajectory up to T_1, we define another stopping time

$$\bar{T}_\epsilon := \inf\{t \ge T_* : G_t \le \epsilon \Lambda_\epsilon\},$$

where Λ_ϵ (to be specified next) satisfies $\Lambda_\epsilon \to \infty$ and $\epsilon\Lambda_\epsilon \to 0$. Importantly, $\bar{T}_\epsilon$ is an auxiliary proof device rather than a new transition time. For sufficiently small ϵ, we have $\Lambda_\epsilon > \delta$ and consequently $\bar{T}_\epsilon \le T_1$. Therefore, the use of $\bar{T}_\epsilon$ splits the trajectory into

$$\underbrace{T_* \le t < \bar{T}_\epsilon}_{\text{Dominant part}} \quad + \quad \underbrace{\bar{T}_\epsilon \le t < T_1}_{\text{Short transition period}}.$$

Its purpose is to isolate the part of the trajectory where the first bound in (4) is sufficiently accurate to recover the exact leading coefficient $1/\gamma_\infty$.

Concretely, if we consider $t < \bar{T}_\epsilon$, we have $G_t > \epsilon\Lambda_\epsilon$ and therefore $d\epsilon/G_t < d/\Lambda_\epsilon = o(1)$. Consequently, it holds that $\langle g_t, \sigma_\epsilon(g_t)\rangle \ge (\gamma_\infty - d/\Lambda_\epsilon)\, G_t$. Therefore, the effective descent coefficient is $\gamma_\infty - o(1)$, which preserves the sharp leading constant. Summing the loss recursion gives

$$S_{\bar{T}_\epsilon} - S_{T_*} \le \frac{\ell_\epsilon}{\gamma_\infty} + O\left(\frac{\ell_\epsilon}{\Lambda_\epsilon} + \sum_{t < \bar{T}_\epsilon} \eta_t^2 + 1\right).$$

Define

$$\omega_a(\ell) := \begin{cases} \ell^{1-a}, & a \in (1/3,1), \\ \log \ell, & a = 1. \end{cases}, \qquad \Lambda_\epsilon := \begin{cases} \ell_\epsilon^a, & a \in (1/3,1), \\ \ell_\epsilon / \log \ell_\epsilon, & a = 1. \end{cases}$$

We have $\ell_\epsilon/\Lambda_\epsilon = \omega_a(\ell_\epsilon)$. Using $\sum_{t<\bar{T}_\epsilon} \eta_t^2 = O\big(\omega_a(\ell_\epsilon)\big)$, we obtain $S_{\bar{T}_\epsilon} - S_{T_*} \le \ell_\epsilon/\gamma_\infty + O\big(\omega_a(\ell_\epsilon)\big)$.

Finally, between $\bar{T}_\epsilon$ and T_1, the gradient proxy only needs to decrease from $G_t \asymp \epsilon\Lambda_\epsilon$ to $G_t \asymp \delta\epsilon$. This corresponds to a multiplicative factor of order Λ_ϵ/δ. The weaker but uniformly positive descent

coefficient κ_δ is sufficient over this shorter interval and gives $S_{T_1} - S_{\bar{T}_\epsilon} = O_\delta\big(\log \Lambda_\epsilon\big)$. The choice of Λ_ϵ ensures that $\log \Lambda_\epsilon = O\big(\omega_a(\ell_\epsilon)\big)$. Therefore, the transition period only contributes a lower-order term and we obtain the desired

$$S_{T_1} - S_{T_*} \le \frac{\ell_\epsilon}{\gamma_\infty} + O\big(\omega_a(\ell_\epsilon)\big). \tag{6}$$

A similar approach can be used to obtain the scaling behavior of S_{T_1} stated in the following theorem. Its proof is simpler than that of Theorem 3 and no additional stopping time is required. It shows that $S_{T_1^-} \sim S_{T_1} \sim S_{T_1^+} \sim (1/\gamma_\infty)\ell_\epsilon$ as $\epsilon \downarrow 0$.

Theorem 4. *For every $b, \delta \in (0,1)$, we have $S_{T_1^\delta(\epsilon)} - S_{T_1^-(\epsilon;b)} = O(\log \ell_\epsilon)$, $S_{T_1^+(\epsilon;b)} - S_{T_1^\delta(\epsilon)} = \Theta(\ell_\epsilon^b)$, and $S_{T_1^\delta(\epsilon)} - S_{T_*} = \frac{1}{\gamma_\infty}\ell_\epsilon + o(\ell_\epsilon)$.*

2. The weight-norm bound Our next goal is to use (6) to obtain a sharp bound on the norm of w_{T_1}. Based on Lemma 1, for any $0 \le r < T_1$, we have

$$\left\| \sum_{t=r}^{T_1-1} \eta_t D_t \right\|_\infty \le S_{T_1} - S_r + C_{\mathrm{rad}} \eta_r H_\epsilon, \tag{7}$$

where $H_\epsilon := 1 + \log\left(\frac{R_1^2+\epsilon^2}{(1-\beta_2)\epsilon^2}\right) = O(\ell_\epsilon)$. If we apply Lemma 1 directly with $r = T_*$, then we would only obtain $\eta_{T_*} H_\epsilon = O(\ell_\epsilon)$, which is of the same order as the leading term $\ell_\epsilon/\gamma_\infty$. Therefore, this approach does not give the desired lower-order remainder $O\big(\omega_a(\ell_\epsilon)\big) = o(\ell_\epsilon)$.

Instead, we split the trajectory at $k_\epsilon := \lceil \ell_\epsilon \rceil$ and write $w_{T_1} = w_{k_\epsilon} - \sum_{t=k_\epsilon}^{T_1-1} \eta_t D_t$. We can show $S_{k_\epsilon} = O\big(\omega_a(\ell_\epsilon)\big) = o(\ell_\epsilon)$ and $S_{T_1} = \Omega(\ell_\epsilon)$. Hence, for sufficiently small ϵ, we have $k_\epsilon \ge T_*$ and $S_{k_\epsilon} < S_{T_1}$, which implies $k_\epsilon < T_1$. This is needed so that Lemma 1 can be applied to the interval $[k_\epsilon, T_1)$. The choice of k_ϵ balances the two terms $\|w_{k_\epsilon}\|_\infty$ and $\eta_{k_\epsilon} H_\epsilon$. Concretely, the contribution of the trajectory up to k_ϵ is small: $\|w_{k_\epsilon}\|_\infty \le \|w_0\|_\infty + C_D S_{k_\epsilon} = O\big(\omega_a(\ell_\epsilon)\big)$, while the $\eta_{k_\epsilon} H_\epsilon$ term in Lemma 1 is also small: $\eta_{k_\epsilon} H_\epsilon = O\big(\omega_a(\ell_\epsilon)\big)$. Consequently, using (7) with $r = k_\epsilon$, we obtain

$$\|w_{T_1}\|_\infty \le \|w_{k_\epsilon}\|_\infty + \left\| \sum_{t=k_\epsilon}^{T_1-1} \eta_t D_t \right\|_\infty \le S_{T_1} - S_{k_\epsilon} + O\big(\omega_a(\ell_\epsilon)\big).$$

Finally, using (6) and $k_\epsilon \ge T_*$ (hence, $S_{T_1} - S_{k_\epsilon} \le S_{T_1} - S_{T_*}$), we obtain

$$\|w_{T_1}\|_\infty \le \frac{\ell_\epsilon}{\gamma_\infty} + O(\omega_a(\ell_\epsilon)). \tag{8}$$

3. From the ℓ_∞-margin to the ℓ_2-margin Define $x_\epsilon := w_{T_1}/\|w_{T_1}\|_\infty$. We first complete the argument for the ℓ_∞-margin gap at T_1 in Theorem 3. At T_1, we have $G_{T_1} \le \delta\epsilon$, and $L_{T_1} \le 2G_{T_1} \le 2\delta\epsilon$ by Proposition 2.1. The loss-margin relationship in Lemma 8 (App. B) also gives $L(w_{T_1}) \ge \frac{\log 2}{n} e^{-\rho(w_{T_1})}$. Therefore, it holds that $\rho(w_{T_1}) \ge \ell_\epsilon - O(1)$. Dividing this lower bound by the radius upper bound in (8) and using the positive homogeneity of ρ, we obtain

$$\widehat{\gamma}_\infty(w_{T_1}) = \frac{\rho(w_{T_1})}{\|w_{T_1}\|_\infty} = \rho(x_\epsilon) \ge \frac{\ell_\epsilon - O(1)}{\ell_\epsilon/\gamma_\infty + O(\omega_a(\ell_\epsilon))} = \gamma_\infty - O(r_\epsilon), \tag{9}$$

where $r_\epsilon = \omega_a(\ell_\epsilon)/\ell_\epsilon$. Next, we show how to derive the second inequality of Theorem 3 based on this.

Set $A := [I_d; -I_d; -Z] \in \mathbb{R}^{(2d+n)\times d}$ where the i-th row of Z is z_i^T, and $b := [\mathbf{1}_d; \mathbf{1}_d; -\gamma_\infty \mathbf{1}_n] \in \mathbb{R}^{2d+n}$ where $\mathbf{1}$ is a vector of all ones. It then holds that $\mathcal{M}_\infty = \big\{u \in \mathbb{R}^d : Au \le b\big\}$. For this set, applying Hoffman's error bound in convex analysis (Hoffman, 1952) (see Lemma 26 in App. F), we obtain for any $x \in \mathbb{R}^d$ that

$$\mathrm{dist}_2(x, \mathcal{M}_\infty) \le H_{2,\infty}(A) \, \|(Ax - b)_+\|_\infty \,, \tag{10}$$

where $\mathrm{dist}_2(x, \mathcal{M}_\infty) := \min_{u \in \mathcal{M}_\infty} \|x - u\|_2$, $[y]_+ := (\max\{y_j, 0\})_j$, and the constant $H_{2,\infty}(A)$ depends only on the constraint matrix A (which depends on the data). We can further show $\|(Ax_\epsilon - b)_+\|_\infty = \gamma_\infty - \rho(x_\epsilon)$. By using $\gamma_\infty - \rho(x_\epsilon) \le O(r_\epsilon)$ (obtained from (9)) and (10), we

obtain $\mathrm{dist}_2(x_\epsilon, \mathcal{M}_\infty) \le O(r_\epsilon)$. Equivalently, there exists $u_\epsilon \in \mathcal{M}_\infty$ such that $\|x_\epsilon - u_\epsilon\|_2 \le O(r_\epsilon)$. This is the central idea of this part of the proof. A small max-margin deficit only shows that $\rho(x_\epsilon) \approx \gamma_\infty$. The Hoffman bound strengthens this scalar statement into the geometric consequence: $x_\epsilon \approx u_\epsilon$ for some $u_\epsilon \in \mathcal{M}_\infty$. This allows us to compare the Euclidean-normalized margins. Define $F(v) := {}^{\rho(v)}/_{\|v\|_2}$, $v \neq 0$. Since $u_\epsilon \in \mathcal{M}_\infty$, it follows from the definition of $\gamma_{2|\infty}$ that $F(u_\epsilon) \le \gamma_{2|\infty}$. The Lipschitz property of the function ρ in Lemma 27 (App. F) then gives

$$F(x_\epsilon) \le F(u_\epsilon) + (R_2 + \gamma_\infty)\|x_\epsilon - u_\epsilon\|_2.$$

Using the bounds on $F(u_\epsilon)$ and $\|x_\epsilon - u_\epsilon\|_2$, we obtain $F(x_\epsilon) \le \gamma_{2|\infty} + O_\delta(r_\epsilon)$. Finally, by the positive homogeneity of ρ, it holds that $F(x_\epsilon) = \widehat{\gamma}_2(w_{T_1})$, which gives the desired inequality.

4. Transfer to T_1^- and T_1^+ The core idea in transferring the margin results at T_1 to T_1^- is to compare $\|w_{T_1} - w_{T_1^-}\|_\infty$ and $\|w_{T_1^-}\|_\infty$. If the former is small relative to the latter, then the change in the normalized margin is also small. Concretely, Lemma 28 (App. F) gives $|\hat{\gamma}_p(x) - \hat{\gamma}_p(y)| \lesssim \|x - y\|_p / \|y\|_p$, $p \in \{\infty, 2\}$. After establishing $\|w_{T_1^-}\|_p \gtrsim \ell_\epsilon$ and $\|w_{T_1} - w_{T_1^-}\|_p \lesssim \log \ell_\epsilon$ for $p \in \{\infty, 2\}$, we obtain $|\hat{\gamma}_p(w_{T_1}) - \hat{\gamma}_p(w_{T_1^-})| = O(\log \ell_\epsilon / \ell_\epsilon)$. The desired results then follow from the corresponding margin bounds at T_1 and the triangle inequality.

However, the same argument does not generally recover the desired $O(r_\epsilon)$ rate when transferring the margin bounds from T_1 to T_1^+. Since $S_{T_1^+} - S_{T_1} = O(\ell_\epsilon^b)$ (given in Theorem 4), we obtain $\|w_{T_1^+} - w_{T_1}\|_p \lesssim \ell_\epsilon^b$ for $p \in \{\infty, 2\}$ using the triangle inequality and the uniform update bound in Lemma 11 (App. B). Since $\|w_{T_1}\|_p \gtrsim \ell_\epsilon$, $p \in \{\infty, 2\}$, Lemma 28 (App. F) gives $|\hat{\gamma}_p(w_{T_1^+}) - \hat{\gamma}_p(w_{T_1})| = O(\ell_\epsilon^{b-1})$, $p \in \{\infty, 2\}$. Consequently, the resulting rate is $O(r_\epsilon + \ell_\epsilon^{b-1})$, rather than $O(r_\epsilon)$. For $a \in (1/3, 1)$, ℓ_ϵ^{b-1} is worse than r_ϵ if $b > 1 - a$; and for $a = 1$, it is always worse. Therefore, our next goal is to give a tighter analysis of the trajectory after T_1.

Since Adam has entered the gradient-like regime by T_1, a natural idea is to decompose its update into a gradient direction and a residual direction for $t \ge T_1$. Specifically, we ensure that the gradient direction is of unit norm and the norm of the residual direction is exactly the gradient residual. Let $u_t := g_t / \|g_t\|_2$. We write Adam's iterate as

$$w_{t+1} = w_t - \theta_t(u_t + e_t), \quad \text{where} \quad e_t := \frac{\epsilon D_t - g_t}{\|g_t\|_2}, \quad \theta_t := \frac{\eta_t \|g_t\|_2}{\epsilon}. \tag{11}$$

We regard θ_t as the effective step size and define the accumulated original and effective step sizes by

$$V_t^\delta(\epsilon) := \sum_{s=T_1^\delta(\epsilon)}^{t-1} \eta_s, \qquad H_t^\delta(\epsilon) := \sum_{s=T_1^\delta(\epsilon)}^{t-1} \theta_s.$$

For simplicity, we let $V_t := V_t^\delta(\epsilon)$ and $H_t := H_t^\delta(\epsilon)$. Proposition 3.1 relates these quantities and shows that the post-T_1 trajectory is a normalized-gradient trajectory with a summable perturbation.

Proposition 3.1. *For every $\delta \in (0, 1)$, there exists a constant $\tilde{\epsilon}_1(\delta)$ such that, for every $0 < \epsilon \le \tilde{\epsilon}_1(\delta)$ and every $t \ge T_1^\delta(\epsilon)$, we have $L_t \asymp G_t \asymp \frac{\epsilon}{1+V_t^\delta(\epsilon)}$, $H_t^\delta(\epsilon) \asymp \log(1 + V_t^\delta(\epsilon))$, and $\sum_{s=T_1^\delta(\epsilon)}^{\infty} \theta_s \|e_s\|_2 \le E_0$, where E_0 is a constant that depends on the data and the step-size schedule.*

Using Proposition 3.1 instead, we obtain $\|w_{T_1^+} - w_{T_1}\|_p \le H_{T_1^+} + E_0$, $p \in \{2, \infty\}$. Since $V_{T_1^+} = O(\ell_\epsilon^b)$, it holds that $H_{T_1^+} = O(\log \ell_\epsilon)$ by $H_{T_1^+} \lesssim \log(1 + V_{T_1^+})$. Consequently, we have $\|w_{T_1^+} - w_{T_1}\|_p = O(\log \ell_\epsilon)$, $p \in \{2, \infty\}$, which leads to $|\widehat{\gamma}_p(w_{T_1^+}) - \widehat{\gamma}_p(w_{T_1})| = O\left({}^{\log \ell_\epsilon}/_{\ell_\epsilon}\right) = O(r_\epsilon)$. The margin gap results at T_1^+ in Theorem 3 then follow. We generalize this result to any $t \ge T_1$ in the following proposition. Essentially, the margin gap is controlled by two terms: r_ϵ and ${}^{H_t}/_{\ell_\epsilon}$. The former controls the gap that is already present at T_1, while the latter measures the post-crossover movement relative to the classifier's existing size.

Proposition 3.2. *For every $\delta \in (0, 1)$, there exists a constant $\tilde{\epsilon}_3(\delta)$ such that for every $0 < \epsilon \le \tilde{\epsilon}_3(\delta)$ and $t \ge T_1^\delta(\epsilon)$, it holds that*

$$\gamma_\infty - \hat{\gamma}_\infty(w_t) \le O\left(r_\epsilon + \frac{H_t^\delta(\epsilon)}{\ell_\epsilon}\right), \qquad \gamma_2 - \widehat{\gamma}_2(w_t) \ge \Delta_{\text{geom}} - O\left(r_\epsilon + \frac{H_t^\delta(\epsilon)}{\ell_\epsilon}\right).$$

Euclidean Optimality We quantify the movement required to approach the ℓ_2-optimal geometry. Assume $\Delta_{\text{geom}} > 0$ and fix $\zeta \in (0, \Delta_{\text{geom}})$. By Theorem 3, we have for all sufficiently small ϵ that

$$\gamma_2 - \widehat{\gamma}_2\big(w_{T_1}\big) \geq \Delta_{\text{geom}} - C_\delta r_\epsilon \geq {}^{(\Delta_{\text{geom}}+\zeta)}/_2 > \zeta, \tag{12}$$

where $C_\delta > 0$ is some constant. Therefore, the Euclidean-margin gap has not yet fallen below ζ at T_1. Consequently, the following hitting time records a genuine post-crossover change:

$$\tau_\zeta^\delta(\epsilon) := \inf\left\{t \geq T_1^\delta(\epsilon) : \gamma_2 - \widehat{\gamma}_2(w_t) \leq \zeta\right\}.$$

The second inequality of Proposition 3.2 suggests that reducing the Euclidean-margin gap below ζ requires effective movement of order ℓ_ϵ. Theorem 5 sharpens this observation by establishing a two-sided comparison with the classifier norm inherited at the crossover.

Theorem 5. *Suppose that $\Delta_{\text{geom}} > 0$. For any $\delta \in (0,1)$ and $\zeta \in (0, \Delta_{\text{geom}})$, there exists a constant $\tilde{\epsilon}_4(\delta, \zeta)$ such that the following statements hold for every $0 < \epsilon \leq \tilde{\epsilon}_4(\delta, \zeta)$:*

(i) For every $t \geq T_1^\delta(\epsilon)$, we have $\gamma_2 - \hat{\gamma}_2(w_t) \leq O\big(\frac{1+\|w_{T_1^\delta(\epsilon)}\|_2}{1+H_t^\delta(\epsilon)}\big)$.

(ii) There exist constants $0 < c_\zeta \leq C_\zeta < \infty$ such that $c_\zeta \|w_{T_1^\delta(\epsilon)}\|_2 \leq H^\delta_{\tau_\zeta^\delta(\epsilon)}(\epsilon) \leq C_\zeta(1 + \|w_{T_1^\delta(\epsilon)}\|_2)$. Consequently, it holds that $H^\delta_{\tau_\zeta^\delta(\epsilon)}(\epsilon) = \Theta(\log \frac{1}{\epsilon})$ and $V_{\tau_\zeta^\delta(\epsilon)} = \epsilon^{-\Theta(1)}$.

We sketch the proof of Part (ii) (see App. G for details). Let $\tau_\zeta := \tau_\zeta^\delta(\epsilon)$ and $R_{T_1} := \|w_{T_1}\|_2$. Part (i) shows that after T_1, the Euclidean-margin gap is controlled by the ratio between the classifier norm inherited at the crossover and the subsequently accumulated effective step size H_t. Since $H_t \to \infty$ by Proposition 3.1, the gap is at most ζ once H_t is a sufficiently large ζ-dependent multiple of R_{T_1}. Accounting for the bounded one-step overshoot gives $H_{\tau_\zeta} \lesssim 1 + R_{T_1}$.

For the lower bound, we have $\gamma_2 - \widehat{\gamma}_2(w_{T_1}) \geq {}^{(\Delta_{\text{geom}}+\zeta)}/_2$ for all sufficiently small ϵ (see (12)). At τ_ζ, the Euclidean-margin gap is at most ζ. Together, they imply $\widehat{\gamma}_2(w_{\tau_\zeta}) - \widehat{\gamma}_2(w_{T_1}) \geq {}^{(\Delta_{\text{geom}}-\zeta)}/_2$, and Lemma 28 (App. F) therefore gives ${}^{(\Delta_{\text{geom}}-\zeta)}/_2 \leq {}^{2R_2\|w_{\tau_\zeta} - w_{T_1}\|_2}/_{R_{T_1}}$. Using the representation in (11) and the summability bound in Proposition 3.1, we obtain $\|w_{\tau_\zeta} - w_{T_1}\|_2 \leq H_{\tau_\zeta} + E_0$. Combining these inequalities yields $H_{\tau_\zeta} \geq \frac{\Delta_{\text{geom}}-\zeta}{4R_2} R_{T_1} - E_0$. Since $R_{T_1} \to \infty$, the fixed constant E_0 can be absorbed for sufficiently small ϵ, giving $H_{\tau_\zeta} \gtrsim R_{T_1}$. The lower and upper bounds together imply $H_{\tau_\zeta} \asymp \|w_{T_1}\|_2 = \Theta(\ell_\epsilon)$ (see App. F for details). Combining the S-time characterization in Theorem 4 with Theorem 5 gives the corresponding iteration-count scales.

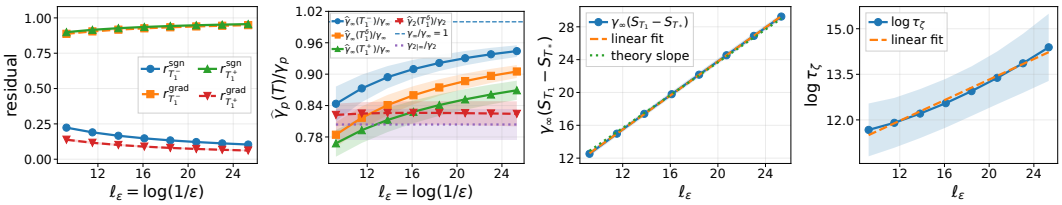
Corollary 6. *Suppose that $\Delta_{\text{geom}} > 0$. Fix any $\zeta \in (0, \Delta_{\text{geom}})$ and $\delta \in (0,1)$. Under the schedule $\eta_t = \eta_0(t+1)^{-a}$, the following holds as $\epsilon \downarrow 0$:*

$$\begin{cases} T_1^\delta(\epsilon) = \Theta\big((\log(1/\epsilon))^{1/(1-a)}\big), \\ \tau_\zeta^\delta(\epsilon) = \epsilon^{-\Theta(1)/(1-a)}, \end{cases} \quad a \in (1/3, 1); \qquad \begin{cases} T_1^\delta(\epsilon) = \epsilon^{-1/(\eta_0\gamma_\infty)+o(1)}, \\ \tau_\zeta^\delta(\epsilon) = \exp\big(\epsilon^{-\Theta(1)}\big), \end{cases} \quad a = 1.$$

Using the gradient proxy for the cross-entropy loss introduced by Fan et al. (2025), all the results in this section can be extended to the separable multiclass classification setting (details in App. I).

Experiments We test our predictions on eight balanced Gaussian-mixture datasets with $n = 64$ and $d = 128$. For seed s, we draw $\widetilde{\nu}_s \sim \mathcal{N}(0, I_d)$, set $\nu_{\star,s} = \widetilde{\nu}_s / \|\widetilde{\nu}_s\|_2$, and sample $X \mid Y = y \sim \mathcal{N}\big(8y\nu_{\star,s}, D^2\big)$ where $D_{jj} = 3^{(j-1)/(d-1)}$. Thus, the coordinatewise standard deviations range from 1 to 3. For each $\epsilon \in \{10^{-4}, \ldots, 10^{-11}\}$, we run Adam from $w_0 = 0$ with $(\beta_1, \beta_2) = (0.9, 0.99)$ and $\eta_t = 10^{-3}(t+1)^{-1/2}$. Define $T_\star := \max_\epsilon \inf\{t : L_t \leq \log(2)/n\}$. We take $b = 0.9$, $\delta = 0.2$, and $\zeta = 0.75\Delta_{\text{geom}}$. All curves report the mean and one standard deviation across the seeds.

Figures 1a and 1b confirm the separation between the update and classifier transitions. As $\epsilon \downarrow 0$, the ordered residual pair $(r^{\text{sgn}}, r^{\text{grad}})$ approaches $(0,1)$ at T_1^- and $(1,0)$ at T_1^+. Meanwhile, the max-norm margin ratios at T_1^-, T_1, and T_1^+ all approach one, whereas the Euclidean-margin ratio at T_1 remains close to $\gamma_{2|\infty}/\gamma_2 < 1$. At the smallest tested value of ϵ, we have $r^{\text{grad}}_{T_1^+} = 0.062 \pm 0.005$ while ${}^{(\gamma_2 - \widehat{\gamma}_2(w_{T_1^+}))}/_{\Delta_{\text{geom}}} = 0.831 \pm 0.030$. Thus, more than 83% of the max-norm-to-Euclidean gap remains even after the update has become strongly gradient-like. Figures 1c and 1d support the



(a) Update residuals **(b)** Classifier margins **(c)** Local-update crossover **(d)** Classifier transition

Figure 1: **Update–classifier separation and transition scales.** **(a)** Sign and gradient residuals at T_1^- and T_1^+ versus $\ell_\epsilon = \log(1/\epsilon)$. **(b)** Max-norm margin ratios at T_1^-, T_1, and T_1^+, together with the Euclidean-margin ratio at T_1. The horizontal lines indicate 1 and the across-seed mean of $\gamma_{2|\infty}/\gamma_2$. **(c)** $\gamma_\infty\big(S_{T_1} - S_{T_\star}\big)$ versus ℓ_ϵ. The fitted slope is 1.036, close to the predicted unit slope. Since $\eta_t = \eta_0(t+1)^{-1/2}$ implies $S_t = 2\eta_0\sqrt{t} + O(1)$, this also supports $T_1 = \Theta(\ell_\epsilon^2)$. **(d)** $\log \tau_\zeta$ versus ℓ_ϵ, consistent with $\tau_\zeta = \epsilon^{-\Theta(1)}$. See App. A.1/App. A.2 for more binary/nonlinear experiments.

predicted separation in time scales. The fitted slope of $\gamma_\infty(S_{T_1} - S_{T_\star})$ against ℓ_ϵ is 1.036, close to the predicted unit slope. Since $a = 1/2$, the approximately linear dependence of $\sqrt{T_1}$ and $\log \tau_\zeta$ on ℓ_ϵ agrees with $T_1 = \Theta(\ell_\epsilon^2)$ and $\tau_\zeta = \epsilon^{-\Theta(1)}$. The experiments therefore reproduce the predicted polylogarithmic-versus-polynomial separation between the local-update and classifier transitions.

4 Related Work

For separable linear classification, gradient descent converges in direction to the Euclidean max-margin classifier (Soudry et al., 2018; Ji & Telgarsky, 2019; Nacson et al., 2019). Its margin gap decays at a rate $O(1/\log t)$, while normalized gradient descent and accelerated methods improve this rate to $O(1/t)$ and $O(\log t/t^2)$, respectively (Ji & Telgarsky, 2021; Ji et al., 2021). Related results cover arbitrary constant step sizes (Wu et al., 2023), multiclass classification (Soudry et al., 2018; Ravi et al., 2024), homogeneous neural networks (Lyu & Li, 2019), matrix factorization (Gunasekar et al., 2017; Li et al., 2020), and self-attention (Tarzanagh et al., 2023); see Vardi (2023) for a survey. Implicit bias under more general steepest-descent geometries has also been studied (Fan et al., 2025; Tsilivis et al., 2025; Li et al., 2026), with sign and spectral descent as prominent examples (Bernstein et al., 2018; Carlson et al., 2015; Bernstein & Newhouse, 2024). The implicit bias of mirror descent has also been investigated (Sun et al., 2022; 2023).

For Adam on separable linear data, a fixed positive stability constant leads to the Euclidean max-margin direction, whereas removing it yields convergence of the normalized margin to the optimal max-norm margin (Wang et al., 2022; Zhang et al., 2024). In the per-sample setting without a stability constant, Baek et al. (2026) show that incremental sampling can alter this bias and, in an extreme case, recover the Euclidean max-margin direction. Related analyses connect AdamW to ℓ_∞-constrained optimization (Xie & Li, 2024) and study the implicit bias of adaptive methods in homogeneous neural networks (Wang et al., 2021; Gronich & Vardi, 2026). In another related work, Wang et al. (2026) study memoryless smoothed-sign descent—the $\beta_1 = \beta_2 = 0$ special case of Adam—under an annealed stability constant. They show that the annealing rate selects a continuum of limiting classifiers along a Burg-type barrier path. Their analysis compares endpoint geometries across annealing schedules; we instead study the within-trajectory evolution of the update and classifier geometries for bias-corrected Adam with a fixed stability constant.

5 Conclusion

We studied deterministic, full-batch Adam with a fixed $\epsilon > 0$ on linearly separable classification problems. At the update crossover, Adam is already gradient-like, yet its classifier remains nearly max-norm-margin optimal and, when $\Delta_{\text{geom}} > 0$, bounded away from Euclidean optimality. Closing this gap requires $\Theta(\log(1/\epsilon))$ additional effective movement, yielding polylogarithmic-versus-polynomial transition times for $a \in (1/3, 1)$ and polynomial-versus-exponential times for $a = 1$. Important directions for future work include extending the current analysis to stochastic Adam, Adam with constant step sizes, and neural networks.

AI USE STATEMENT

We used generative AI tools to correct grammatical errors. All experimental code was implemented by AI tools under the authors' guidance and checked by the authors. We also used AI tools to check mathematical notation, identify typographical errors, and assist in extending the binary results to the multiclass setting. The use of Hoffman's error bound in the proof of Theorem 29 was suggested by AI tools, while the proof itself was developed by the authors. We have not used generative AI tools for any other tasks requiring disclosure. We have reviewed and validated all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Beomhan Baek, Minhak Song, and Chulhee Yun. Implicit bias of per-sample adam on separable data: Departure from the full-batch regime. In *International Conference on Learning Representations*, volume 2026, pp. 120180–120229, 2026.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International conference on machine learning*, pp. 560–569. PMLR, 2018.

David Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. *Advances in neural information processing systems*, 28, 2015.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Eitan Gronich and Gal Vardi. The implicit bias of adam and muon on smooth homogeneous neural networks. *arXiv preprint arXiv:2602.16340*, 2026.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

Alan J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4):263–265, 1952. doi: 10.6028/jres.049.027.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2019.

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

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

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

Jichu Li, Xuan Tang, and Difan Zou. The implicit bias of steepest descent with mini-batch stochastic gradient. *arXiv preprint arXiv:2602.11557*, 2026.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *arXiv preprint arXiv:2012.09839*, 2020.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Du, Jason Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/909c8fef63e1cede406ce9e6794f99a2-Abstract-Conference.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 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3420–3428, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Javier F. Peña, Juan C. Vera, and Luis F. Zuluaga. New characterizations of hoffman constants for systems of linear constraints. *Mathematical Programming*, 187:79–109, 2021. doi: 10.1007/s10107-020-01473-6.

Hrithik Ravi, Clayton Scott, Daniel Soudry, and Yutong Wang. The implicit bias of gradient descent on separable multiclass data. *Advances in neural information processing systems*, 37:81324–81359, 2024.

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.

Haoyuan Sun, Kwangjun Ahn, Christos Thrampoulidis, and Navid Azizan. Mirror descent maximizes generalized margin and can be implemented efficiently. *Advances in Neural Information Processing Systems*, 35:31089–31101, 2022.

Haoyuan Sun, Khashayar Gatmiry, Kwangjun Ahn, and Navid Azizan. A unified approach to controlling implicit regularization via mirror descent. *Journal of Machine Learning Research*, 24 (393):1–58, 2023. URL http://jmlr.org/papers/v24/23-0836.html.

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023.

Nikolaos Tsilivis, Eitan Gronich, Julia Kempe, and Gal Vardi. Flavors of margin: Implicit bias of steepest descent in homogeneous neural networks, 2025. URL https://arxiv.org/abs/2410.22069.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6): 86–93, 2023.

Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10849–10858, 2021.

Bohan Wang, Qi Meng, Huishuai Zhang, Ruoyu Sun, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. Does momentum change the implicit regularization on separable data? *Advances in Neural Information Processing Systems*, 35:26764–26776, 2022.

Xiangwu Wang, Chengwei Cao, Yicheng Song, Ran Bi, and Peilin Yu. Stability annealing selects the implicit bias of smoothed sign descent: A rate-indexed barrier path on separable data. *arXiv preprint arXiv:2607.06013*, 2026.

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

Shuo Xie and Zhiyuan Li. Implicit bias of adamw: ℓ_∞ norm constrained optimization, 2024. URL https://arxiv.org/abs/2404.04454.

Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

A Experiments **13**
A.1 Binary Experiments 13
A.2 Nonlinear Experiments 14

B Preliminaries **17**
B.1 The loss and the proxy 17
B.2 Adam's Update and EMA 22

C Sign and Gradient Residual Characterization **30**

D Proof of Lemma 1 **35**

E Adam's Local Update Transition **42**
E.1 Proof of Theorem 4 42
E.2 Implications of Theorem 4 50
E.3 Proof of Proposition 3.1 51

F Adam's ℓ_∞-margin optimality **55**
F.1 Margin Gap at T_1 57
F.2 Margin Gap at T_1^- 66
F.3 Margin Gap at T_1^+ 68

G Adam's Euclidean Optimality **71**

H Proof of Corollary 6 **76**

I Multiclass Extension **78**
I.1 Setup and notation 78
I.2 Section B Analogous Lemmas 79
I.3 Section F Analogous Lemmas 87
I.4 Main Results 89

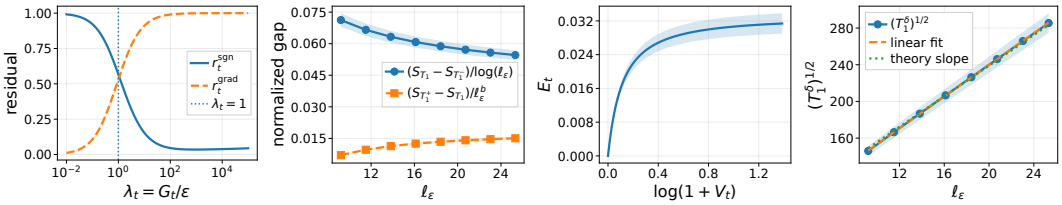
A Experiments

A.1 Binary Experiments

We present additional results from the binary experiments beyond those in Figure 1. Recall that $T_1^- := T_1^-(\epsilon; b)$, $T_1 := T_1^\delta(\epsilon)$, and $T_1^+ := T_1^+(\epsilon; b)$. All curves report the mean and one standard deviation across the seeds. For $\epsilon = 10^{-11}$, Figure 2a shows that the sign and gradient residuals converge to $(0, 1)$ and $(1, 0)$ when $\lambda_t = G_t/\epsilon$ become large and small, respectively. This demonstrates that λ_t serves as an effective order parameter and supports the claims in Proposition C.1 in App. C. Figure 2b shows that $(S_{T_1} - S_{T_1^-})/\log \ell_\epsilon$ is bounded above (in fact, decreases as ℓ_ϵ increases), supporting the claim $S_{T_1} - S_{T_1^-} = O(\log \ell_\epsilon)$ in Theorem 4. Moreover, as ℓ_ϵ increases, the curve of $(S_{T_1^+} - S_{T_1})/\ell_\epsilon^b$ levels off after an initial increase, supporting the claim $S_{T_1^+} - S_{T_1} = \Theta(\ell_\epsilon^b)$ in Theorem 4. Figure 2c shows that the accumulated gradient residual E_t levels off, while the accumulated effective step

size $H_t \asymp \log(1+V_t)$ diverges, supporting the claim of Proposition 3.1. Figure 2d shows a linear relationship between $\sqrt{T_1}$ and ℓ_ϵ, further supporting the scaling behavior claimed in Corollary 6.

For each dataset, we compute γ_∞ by linear programming using SciPy's HiGHS solver. We then minimize $\frac{1}{2}\|u\|_2^2$ subject to $\|u\|_\infty \le 1$ and $Zu \ge \gamma_\infty \mathbf{1}$. Here, $Z \in \mathbb{R}^{n\times d}$ has i-th row $z_i^\top = y_i x_i^\top$. Denoting the resulting solution by $u_\star$, we compute $\gamma_{2|\infty} = \gamma_\infty / \|u_\star\|_2$.



(a) Residual vs. λ_t **(b)** S-time difference **(c)** Accumulated Residual **(d)** $\sqrt{T_1}$ vs. ℓ_ϵ

Figure 2: Additional Binary Experiments. **(a)** Fix $\epsilon = 10^{-11}$. Sign (r_t^{sgn}) and gradient residuals (r_t^{grad}) against the order parameter $\lambda_t = G_t/\epsilon$. **(b)** $(S_{T_1} - S_{T_1^-})/\log \ell_\epsilon$ and $(S_{T_1^+} - S_{T_1})/\ell_\epsilon^b$ against ℓ_ϵ. (c) $E_t := \sum_{s=T_1}^{t-1} \theta_s \|e_s\|_2$ against $\log(1+V_t)$. (d) $\sqrt{T_1}$ against ℓ_ϵ.

A.2 Nonlinear Experiments

These experiments provide a qualitative extension of the linear results. Because the network is trained with a constant stepsize, we do not use them to test the iteration-time exponents derived for linear models. Moreover, exact max-margin computation is challenging, so we use empirical reference margins as approximations.

Experimental setting. We adopt the experimental setup of Tsilivis et al. (2025). For each of the eight seeds, we sample without replacement 50 images of the digit 3 and 50 images of the digit 6 from the MNIST training set. We assign labels -1 and $+1$, respectively, and shuffle the resulting $n = 100$ examples. Each image is flattened into $x_i \in \mathbb{R}^{784}$ and divided by 255 without additional preprocessing. Testing uses all 1,968 images labeled 3 or 6 in the official MNIST test set. The test data are used only to report loss and accuracy.

We train a two-layer ReLU network of width $m = 128$, given by

$$f_\theta(x) = \sum_{r=1}^{m} a_r \,\mathrm{ReLU}(w_r^\top x), \qquad \theta = (W, a), \tag{13}$$

where $W = (w_1^\top, \dots, w_m^\top) \in \mathbb{R}^{m\times d}$ and both layers are trained. The network has no bias terms and no $1/\sqrt{m}$ output normalization. We use the inactive ReLU subgradient at zero. With $\alpha = 0.05$, the initialization is

$$W_{rj,0} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{\alpha^2}{d}\right), \qquad a_{r,0} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \alpha^2). \tag{14}$$

The signed logits and the logistic loss are

$$q_i(\theta) := y_i f_\theta(x_i), \quad L(\theta) := \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-q_i(\theta)}\right). \tag{15}$$

We further define

$$G(\theta) := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+e^{q_i(\theta)}}, \qquad s_i(\theta) := \frac{1}{1+\exp(q_i(\theta))},\ i \in [n]. \tag{16}$$

Denote $G_t := G(\theta_t)$ and $g_t := \nabla L(\theta_t)$. The quantity $s_i(\theta)$ is the magnitude of the derivative of the logistic loss with respect to the signed logit $q_i(\theta)$. Therefore, G_t is the average of this loss-derivative magnitude across samples. However, it does not track the actual parameter gradient scale. The actual parameter gradient is $g_t = -\frac{1}{n}\sum_{i=1}^{n} y_i s_i(\theta_t) \nabla_\theta f_{\theta_t}(x_i)$, which involves the changing Jacobians $\nabla_\theta f_{\theta_t}(x_i)$. Consequently, unlike the linear setting, G_t/ϵ is only an approximate order parameter in

the neural-network experiments. For Adam's hyperparameters, we set $(\beta_1, \beta_2) = (0.9, 0.99)$ and use a constant stepsize $\eta = 3 \times 10^{-3}$. We consider $\epsilon \in \{10^{-4}, 10^{-8}, 10^{-12}\}$ and run each trajectory for 100,000 updates, recording 480 logarithmically spaced checkpoints.

For geometric references, we run paired GD and sign descent (SignGD) trajectories for at most 20,000 iterations using the same data, initialization, and stepsize. Because GD's update magnitude shrinks with the gradient, its late trajectory is stable. We define its reference margin as the median over the final 10% of saved checkpoints. In contrast, because SignGD's update magnitude remains nearly constant, it can continue to move even after the loss becomes very small. We stop it at the first checkpoint at which the training loss falls below 10^{-7} and use its margin at that iterate. Because exact max-margin computation is intractable, we use the GD reference margin as the empirical proxy for ℓ_2-max-margin. We denote the reference margins of SignGD and GD by $\bar{\gamma}_{2,s}^{\mathrm{Sign}}$ and $\bar{\gamma}_{2,s}^{\mathrm{GD}}$ for each seed s. The average values across seeds and standard deviations are: $\bar{\gamma}_2^{\mathrm{Sign}} = 0.026472 \pm 0.014431$ and $\bar{\gamma}_2^{\mathrm{GD}} = 0.507090 \pm 0.038898$. For every seed s, the value of $\bar{\gamma}_{2,s}^{\mathrm{GD}} - \bar{\gamma}_{2,s}^{\mathrm{Sign}}$ is positive.

Update and margin geometry. After flattening W and concatenating it with a, we measure the sign and gradient residuals

$$r_t^{\mathrm{sgn}} := \frac{\big\| g_t \odot \big(D_t - \mathrm{sign}(g_t)\big)\big\|_1}{\|g_t\|_1}, \qquad r_t^{\mathrm{grad}} := \frac{\|\epsilon D_t - g_t\|_2}{\|g_t\|_2}. \tag{17}$$

A small r_t^{sgn} indicates a sign-like direction, whereas a small r_t^{grad} indicates $D_t \approx g_t/\epsilon$ and hence a gradient-like direction. We also compute the gradient residual separately on the W and a blocks, denoted as $r_{t,W}^{\mathrm{grad}}$ and $r_{t,a}^{\mathrm{grad}}$, respectively. The raw training margin is $\rho(\theta) := \min_{i\in[n]} y_i f_\theta(x_i)$. Because the network is two-homogeneous under the joint scaling $\theta \mapsto c\theta$, we use the scale-invariant margin:

$$\widehat{\gamma}_2(\theta) := \frac{\rho(\theta)}{\|\theta\|_2^2}, \qquad \|\theta\|_2^2 := \|W\|_F^2 + \|a\|_2^2.$$

This normalized margin is invariant to joint scaling but not to the function-preserving layer rebalancing $(W, a) \mapsto (cW, a/c)$. We define the following quantity to measure the classifier's geometric transition:

$$R_{2,s}(t) := \frac{\bar{\gamma}_{2,s}^{\mathrm{GD}} - \widehat{\gamma}_{2,s}^{\mathrm{Adam}}(t)}{\bar{\gamma}_{2,s}^{\mathrm{GD}} - \bar{\gamma}_{2,s}^{\mathrm{Sign}}}. \tag{18}$$

It is equal to one when Adam's ℓ_2-margin is replaced by the SignGD reference. For $R_2(t)$, the denominator is the total empirical Euclidean-margin gap between GD and SignGD, while the numerator is the part of this gap that Adam has not yet closed. Therefore, $R_2(t)$ is the fraction of the empirical SignGD-to-GD Euclidean-margin gap that remains, while $1 - R_2(t)$ measures the progress towards the GD reference. Moreover, we can rewrite (18) as

$$\widehat{\gamma}_{2,s}^{\mathrm{Adam}}(t) = R_{2,s}(t)\bar{\gamma}_{2,s}^{\mathrm{Sign}} + \big(1 - R_{2,s}(t)\big)\bar{\gamma}_{2,s}^{\mathrm{GD}}.$$

Therefore, when $R_{2,s}(t) \in [0, 1]$, it is an affine coefficient between the two corresponding scalar reference margins.

Empirical transition times. We define a few time points that mark the geometric transitions of the update and the classifier. They are

$$\begin{aligned}
T_* &:= \text{first of five consecutive checkpoints with } 100\% \text{ training accuracy},\\
T_G^{0.1} &:= \text{first of five consecutive checkpoints at or after } T_* \text{ with } \frac{G_t}{\epsilon} \le 0.1,\\
T_{\mathrm{upd}}^{0.30} &:= \text{first of five consecutive checkpoints at or after } T_* \text{ with } \max\left\{r_{t,W}^{\mathrm{grad}}, r_{t,a}^{\mathrm{grad}}\right\} \le 0.30,\\
\tau_{0.5}^{\mathrm{emp}} &:= \text{first checkpoint at or after } T_{\mathrm{upd}}^{0.30} \text{ from which } R_2(t) \le 0.5 \text{ holds at every remaining checkpoint},
\end{aligned}$$

where the last definition requires at least five remaining checkpoints. The threshold 0.30 in $T_{\mathrm{upd}}^{0.30}$ means that in each layer, the mismatch between ϵD_t and g_t is at most 30% of the gradient norm. It signals that the update has entered a gradient-like regime where $\epsilon D_t \approx g_t$. The half-gap condition

$R_2(t) \le 0.5$ in the definition of $\tau_{0.5}^{\text{emp}}$ is equivalent to $\hat{\gamma}_2^{\text{Adam}}(t) \ge \frac{\bar{\gamma}_2^{\text{GD}} + \bar{\gamma}_2^{\text{Sign}}}{2}$. Thus, at $\tau_{0.5}^{\text{emp}}$, Adam has traversed at least half of the empirical Euclidean-margin interval from SignGD to GD. Moreover, this time is restricted to $t \ge T_{\text{upd}}^{0.30}$. The experiment asks how long the classifier takes to change after the local update has already become gradient-like. It can happen that $R_2(t) > 0.5$ throughout the trajectory and $\tau_{0.5}^{\text{emp}}$ is not reached by the end of training.

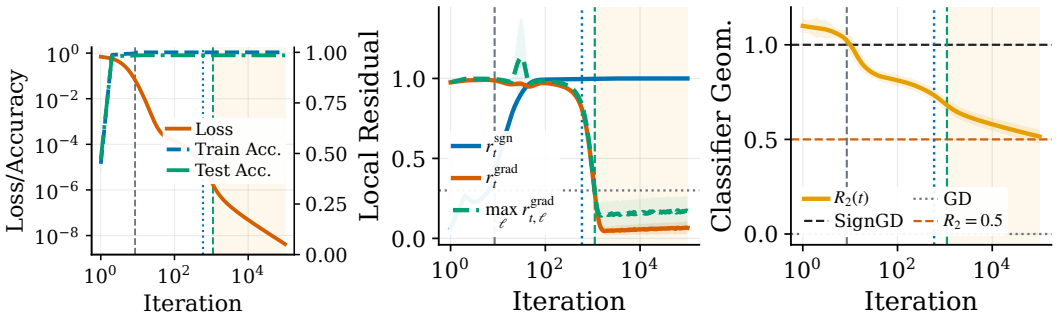
Results. Figures 3 to 5 show the results for $\epsilon = 10^{-4}$, $\epsilon = 10^{-8}$, and $\epsilon = 10^{-12}$, respectively. Each curve shows the mean over the eight seeds, and shaded bands show one sample standard deviation. The gray dashed, blue dotted, and green dashed vertical lines in each figure mark the medians of T_*, $T_G^{0.1}$, and $T_{\text{upd}}^{0.30}$, respectively. Their values together with the first and third quartiles are reported in the following table for the three choices of ϵ. We first note that the training accuracy saturates very early: the median values of T_* are 8, 5, and 5, while the logistic loss continues to decrease by many orders of magnitude. Thus, predictive accuracy alone does not reveal either geometric transition.

Table 1: Event times T_*, $T_G^{0.1}$, $T_{\text{upd}}^{0.30}$, and $\tau_{0.5}^{\text{emp}}$ for $\epsilon \in \{10^{-4}, 10^{-8}, 10^{-12}\}$, reported as median $[Q_1, Q_3]$ over eight seeds and rounded to the nearest iteration, where Q_1 and Q_3 are the first and third quartiles.

ϵ	T_*	$T_G^{0.1}$	$T_{\text{upd}}^{0.30}$	$\tau_{0.5}^{\text{emp}}$
10^{-4}	8 [4, 8]	592 [588, 610]	1,106 [1,100, 1,106]	64,821 [56,686, 73,648] (3/8 confirmed)
10^{-8}	5 [2, 8]	2,120 [2,070, 2,172]	2,970 [2,952, 2,970]	Not observed
10^{-12}	5 [2, 8]	3,779 [3,756, 3,871]	4,808 [4,808, 4,808]	Not observed

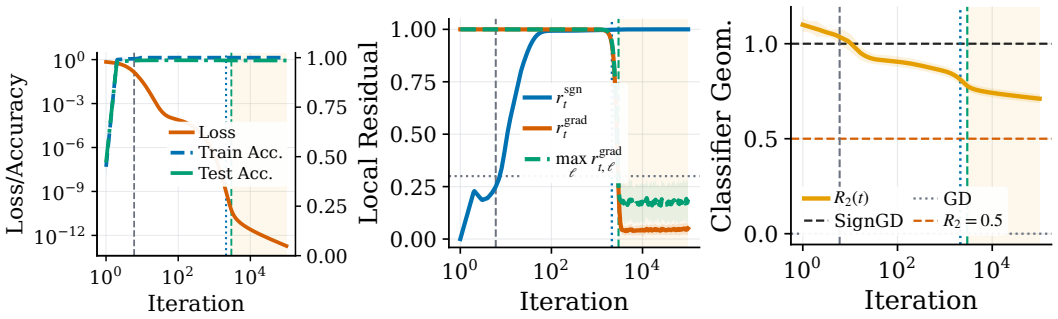
For $\epsilon \in \{10^{-4}, 10^{-8}, 10^{-12}\}$, the corresponding medians of $T_{\text{upd}}^{0.3}$ are 1106, 2970, and 4808. Therefore, the geometric transition is delayed as ϵ decreases. While the median sign residual is approximately one, the median gradient residual at $T_{\text{upd}}^{0.30}$ decreases from 0.280 to 0.241 and then to 0.212. Thus, the updates in all three cases have entered the gradient-like regime and are no longer sign-like.

The classifier transition is substantially slower. Every trajectory satisfies $R_2(T_{\text{upd}}^{0.30}) > 0.5$, with median values 0.677, 0.765, 0.818 for $\epsilon = 10^{-4}, 10^{-8}$ and 10^{-12}, respectively. At the last iteration, the corresponding median values remain 0.500, 0.700, and 0.776. The defining event of $\tau_{0.5}^{\text{emp}}$ is confirmed in only three of eight seeds for $\epsilon = 10^{-4}$ and in none of the seeds for $\epsilon = 10^{-8}$ or 10^{-12}. Therefore, the transition toward Euclidean optimality occurs much later than the update's geometric transition. Moreover, decreasing ϵ leaves a larger empirical Euclidean-margin gap in the gradient-like regime.



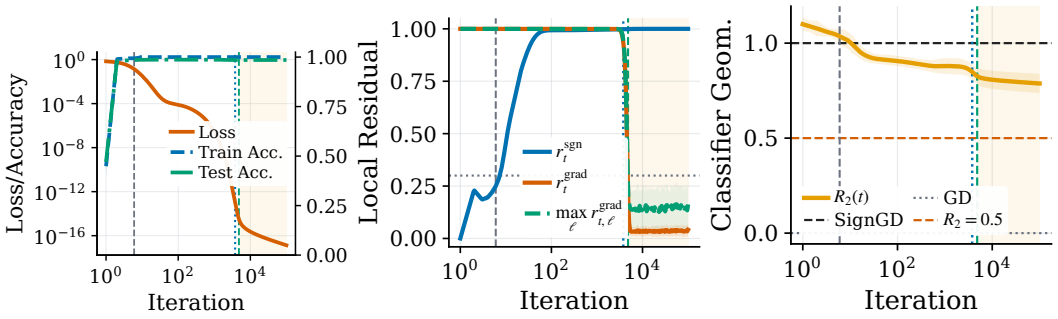
(a) Loss and accuracy. (b) Update residuals. (c) Margin geometry.

Figure 3: Nonlinear experiment with $\epsilon = 10^{-4}$. Curves show the mean over eight paired seeds, and shaded bands show one sample standard deviation. (a) Logistic loss and training and test accuracy against iterations. (b) The sign and gradient residuals and the maximum layerwise gradient residual against iterations. (c) $R_2(t)$ against iterations; the SignGD and GD references are one and zero, respectively, and $R_2 = 0.5$ marks half of the empirical Euclidean-margin gap.



(a) Loss and accuracy. (b) Update residuals. (c) Margin geometry.

Figure 4: Nonlinear experiment with $\epsilon = 10^{-8}$. The plottings are the same as in Figure 3.



(a) Loss and accuracy. (b) Update residuals. (c) Margin geometry.

Figure 5: Nonlinear experiment with $\epsilon = 10^{-12}$. The plottings are the same as in Figure 3.

B PRELIMINARIES

Our objective is $L(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(z_i^T w)$ where $\ell(u) := \log(1 + e^{-u})$. Recall that

$$a_i(w) := -l'(z_i^T w) = \frac{1}{1 + e^{z_i^T w}}, \quad G(w) := \frac{1}{n}\sum_{i=1}^{n} a_i(w), \quad g(w) := \nabla L(w) = -\frac{1}{n}\sum_{i=1}^{n} a_i(w) z_i.$$

For simplicity, we let $G_t := G(w_t)$, $L_t := L(w_t)$, and $g_t := g(w_t)$. We use the decreasing step size schedule $\eta_t = \eta_0(1+t)^{-a}$ where $\eta_0 > 0$ and $a \in (1/3, 1]$. We assume that the data is separable and consequently $\gamma_\infty, \gamma_2 > 0$. We further assume that β_1, β_2, and ϵ are chosen such that $0 \le \beta_1 \le \beta_2 < 1$ and $\epsilon \in (0, 1]$. Recall that $R_1 := \max_{i\in[n]} \|x_i\|_1$ and $R_2 := \max_{i\in[n]} \|x_i\|_2$.

B.1 THE LOSS AND THE PROXY

The lower bound on the ℓ_1-norm of $g(w)$ was proved in Zhang et al. (2024, Lemma A.1). We give an extension to the ℓ_2-norm case following the same approach.

Lemma 7. *It holds for every $w \in \mathbb{R}^d$ that*

$$\gamma_\infty G(w) \le \|g(w)\|_1 \le R_1 G(w), \qquad \gamma_2 G(w) \le \|g(w)\|_2 \le R_2 G(w). \tag{19}$$

In particular, we have $G(w) > 0$ and $g(w) \neq 0$ at every finite w.

Proof. Recall that $z_i := x_i y_i$. We let $\phi(u) := \min_{1\le i\le n} z_i^T u$. This function is continuous because it is the pointwise minimum of finitely many continuous functions. Hence, the maxima are obtained when maximizing $\phi(u)$ over the compact feasible sets $\{u : \|u\|_\infty \le 1\}$ and $\{u : \|u\|_2 \le 1\}$. We denote the maximizers by u_∞ and u_2 respectively and they satisfy:

$$\|u_\infty\|_\infty \le 1, \min_i z_i^T u_\infty = \gamma_\infty \quad \text{and} \quad \|u_2\|_2 \le 1, \min_i z_i^T u_2 = \gamma_2.$$

Consequently, we have $z_i^T u_\infty \geq \gamma_\infty, \forall i$ and $z_i^T u_2 \geq \gamma_2, \forall i$ (note that γ_∞ and γ_2 are positive given the data is separable). This implies the following

$$-\langle g(w), u_\infty \rangle = \frac{1}{n}\sum_{i=1}^n a_i(w) z_i^T u_\infty \geq \frac{1}{n}\sum_{i=1}^n a_i(w)\gamma_\infty = \gamma_\infty G(w).$$

Consequently, we have $\gamma_\infty G(w) \leq -\langle g(w), u_\infty \rangle \leq |\langle g(w), u_\infty \rangle| \leq \|g(w)\|_1 \|u_\infty\|_\infty \leq \|g(w)\|_1$. Similarly for the Euclidean norm, it holds that

$$-\langle g(w), u_2 \rangle = \frac{1}{n}\sum_{i=1}^n a_i(w) z_i^T u_2 \geq \frac{1}{n}\sum_{i=1}^n a_i(w)\gamma_2 = \gamma_2 G(w).$$

Consequently, we have $\gamma_2 G(w) \leq -\langle g(w), u_2 \rangle \leq |\langle g(w), u_2 \rangle| \leq \|g(w)\|_2 \|u_2\|_2 \leq \|g(w)\|_2$. For the upper bounds, we have via triangle inequality that

$$\|g(w)\|_2 = \|-\frac{1}{n}\sum_{i=1}^n a_i(w) z_i\|_2 \leq \frac{1}{n}\sum_{i=1}^n a_i(w)\|z_i\|_2 \leq R_2 G(w),$$

$$\|g(w)\|_1 = \|-\frac{1}{n}\sum_{i=1}^n a_i(w) z_i\|_1 \leq \frac{1}{n}\sum_{i=1}^n a_i(w)\|z_i\|_1 \leq R_1 G(w).$$

Finally, since $e^{z_i^T w} \in (0, \infty)$, it holds that $a_i(w) = \frac{1}{1+e^{z_i^T w}} \in (0,1)$ and consequently $G(w) = \frac{1}{n}\sum_{i=1}^n a_i(w) > 0$, and the conclusion $g(w) \neq 0$ for every finite w follows as $\|g(w)\|_1 > 0$ ($\|g(w)\|_2 > 0$). □

The lower bound on $\frac{G(w)}{L(w)}$ and the conclusion $\rho(w) \geq 0$ are proved in Zhang et al. (2024, Lemma C.4,C.7). We provide the proof here for completeness.

Lemma 8. *For every $w \in \mathbb{R}^d$, we have*

$$G(w) \leq L(w), \qquad \frac{G(w)}{L(w)} \geq 1 - \frac{nL(w)}{2}. \tag{20}$$

If either $G(w) \leq 1/(2n)$ or $L(w) \leq \log 2/n$ holds, then we have $L(w) \leq 2G(w)$. Define $\rho(w) := \min_{i\in[n]} z_i^T w$. The inequality $L(w) \leq \log 2/n$ implies that

$$\rho(w) \geq 0 \quad \textit{and} \quad L(w) \geq \frac{\log 2}{n} e^{-\rho(w)}. \tag{21}$$

Proof. Let $\ell_i := \ell(z_i^\top w)$. Since $e^{-\ell_i} = 1/(1+e^{-z_i^\top w})$, it holds that

$$a_i(w) = \frac{1}{1+e^{z_i^\top w}} = 1 - e^{-\ell_i}.$$

For $x \geq 0$, we have $1 - e^{-x} \leq x$ and $1 - e^{-x} \geq x - x^2/2$. Averaging these inequalities gives

$$G(w) \leq L(w), \qquad G(w) \geq L(w) - \frac{1}{2n}\sum_i \ell_i^2 \overset{(a)}{\geq} L(w) - \frac{nL(w)^2}{2},$$

where (a) follows from $\sum_i \ell_i^2 \leq (\sum_i \ell_i)^2 = n^2 L(w)^2$. Rearranging leads to the desired.

If $G(w) \leq 1/(2n)$, we have $a_i(w) \leq nG(w) \leq 1/2, \forall i$. Since $\ell_i = -\log(1-a_i)$ and $-\log(1-x) \leq 2x$ on $[0, 1/2]$, we obtain

$$L(w) = \frac{1}{n}\sum_i \ell_i = \frac{1}{n}\sum_i -\log(1-a_i(w)) \leq \frac{2}{n}\sum_i a_i(w) = 2G(w).$$

If instead $L(w) \leq \log 2/n$, then each individual loss satisfies $\ell_i \leq nL(w) \leq \log 2$, hence $a_i = 1 - e^{-\ell_i} \leq 1/2$, and the same argument applies.

Given $L(w) \leq \frac{\log 2}{n}$, we have $\ell_i \leq \log 2, \forall i$. The function $\ell(u) = \log(1 + e^{-u})$ is strictly decreasing as $\ell'(u) = -\frac{1}{1+e^u} < 0$. Thus, given $z_i^T w < 0$, we have that $\ell_i = \ell(z_i^T w) > l(0) = \log 2$, which is a contradiction. Since $z_i^T w \geq 0, \forall i$, we conclude that $\rho(w) \geq 0$ by taking the minimum over i.

Let i^* be the index such that $\rho(w) = z_{i^*}^T$. We then have

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-z_i^T w}) \geq \frac{1}{n} \log(1 + e^{-z_{i^*}^T w}) = \frac{1}{n} \log(1 + e^{-\rho(w)}). \tag{22}$$

Define $x := e^{-\rho(w)}$. We have $0 < x \leq 1$ given $\rho(w) \geq 0$. The function $f(x) = \log(1 + x)$ is concave on $[0, 1]$ because $f''(x) = -\frac{1}{(1+x)^2} < 0$. By concavity, we have

$$\log(1 + x) = f(x) = f((1 - x)0 + 1x) \geq (1 - x)f(0) + xf(1) = x \log 2.$$

Substituting $x = e^{-\rho(w)}$, we obtain $\log(1 + e^{-\rho(w)}) \geq (\log 2)e^{-\rho(w)}$. Combining this with (22) finishes the proof. □

The result in the following proposition was shown to hold for the step size schedule $\eta_t = (t + 2)^{-a}$ in Zhang et al. (2024, Lemma C.1.). We provide a proof here for completeness and show it holds for the schedule $\eta_t = \eta_0(t + 1)^{-a}$.

Proposition 9. *Let $\eta_t = \eta_0(t + 1)^{-a}$ with $a \in (0, 1]$ and $\eta_0 > 0$. For every $q \in [0, 1)$ and $c > 0$, there exist a time $t_0 = t_0(q, c, a, \eta_0)$ and a constant $C_{\text{geo}} = C_{\text{geo}}(q, c, a, \eta_0)$ such that*

$$\sum_{r=0}^{t} q^r \left[\exp\left(c \sum_{s=t-r}^{t-1} \eta_s \right) - 1 \right] \leq C_{\text{geo}} \eta_t, \qquad t \geq t_0. \tag{23}$$

Proof. If $q = 0$, then the sum is zero. In this case, we can take $t_0 = 0$ and $C_{\text{geo}} = 0$. Henceforth, we assume $q \in (0, 1)$. We first prove

$$\sum_{r=0}^{\infty} q^r r^k \leq \frac{k!}{(1 - q)^{k+1}}, \tag{24}$$

where k is an integer with $k \geq 1$. For every integer $r \geq 0$, it holds that $r^k \leq (r+1)(r+2) \cdots (r+k) = k!\binom{r+k}{k}$. Therefore, we have $\sum_{r=0}^{\infty} q^r r^k \leq k! \sum_{r=0}^{\infty} \binom{r+k}{k} q^r$. To evaluate the last series, we can differentiate $\sum_{n=0}^{\infty} q^n = \frac{1}{1-q}$ exactly k times. Since $0 < q < 1$, the power series can be differentiated term by term and we obtain $\sum_{n=k}^{\infty} \frac{n!}{(n-k)!} q^{n-k} = \frac{k!}{(1-q)^{k+1}}$. Setting $r = n - k$ and dividing by $k!$, it holds that $\sum_{r=0}^{\infty} \binom{r+k}{k} q^r = \frac{1}{(1-q)^{k+1}}$. Substituting this into the previous inequality finishes the proof of (24).

Define $H_{t,r} := \sum_{s=t-r}^{t-1} \eta_s, 0 \leq r \leq t$ and $\mathcal{E}_t = \sum_{r=0}^{t} q^r (e^{cH_{t,r}} - 1)$. We derive a bound on $H_{t,r}$. Changing variables to $j = s + 1$, we can write $H_{t,r} = \eta_0 \sum_{j=t-r+1}^{t} j^{-a}$. For every integer $j \geq 1$, it holds that

$$j^{-a} = \left(\frac{j + 1}{j} \right)^a (j + 1)^{-a} \leq 2^a (j + 1)^{-a}.$$

Moreover, because x^{-a} is decreasing, we have $(j + 1)^{-a} \leq \int_j^{j+1} x^{-a}\, dx$. Combining these two inequalities leads to $H_{t,r} \leq 2^a \eta_0 \int_{t-r+1}^{t+1} x^{-a}\, dx$, $0 \leq r \leq t$. Summing over $j = t - r + 1, ..., t$, we have

$$H_{t,r} \leq 2^a \eta_0 \int_{t-r+1}^{t+1} x^{-a}\, dx, \qquad 0 \leq r \leq t. \tag{25}$$

We first consider the case where $0 < a < 1$. Evaluating the integral in (25), we obtain

$$H_{t,r} \leq \frac{2^a \eta_0}{1 - a} \left[(t + 1)^{1-a} - (t - r + 1)^{1-a} \right]. \tag{26}$$

Define $u := \frac{t-r+1}{t+1} \in (0, 1]$. Since $0 < 1 - a < 1$, it holds that $u^{1-a} \geq u$. Consequently, we have

$$\begin{aligned}(t+1)^{1-a} - (t-r+1)^{1-a} &= (t+1)^{1-a}\left(1 - u^{1-a}\right) \leq (t+1)^{1-a}(1-u) \\ &= (t+1)^{1-a}\frac{r}{t+1} = \frac{r}{(t+1)^a}.\end{aligned}$$

Substituting this into (26) gives $H_{t,r} \leq \frac{2^a\eta_0}{1-a}\frac{r}{(t+1)^a} = \frac{2^a}{1-a}\, r\eta_t$. Define $d_t := \frac{2^a c}{1-a}\eta_t$. We have $cH_{t,r} \leq d_t r$ and consequently $\mathcal{E}_t \leq \sum_{r=0}^{t} q^r(e^{d_t r} - 1)$.

Using $e^{d_t r} - 1 = \sum_{k=1}^{\infty} \frac{d_t^k r^k}{k!}$, we obtain

$$\mathcal{E}_t \leq \sum_{r=0}^{t} q^r \sum_{k=1}^{\infty} \frac{d_t^k r^k}{k!} = \sum_{k=1}^{\infty} \frac{d_t^k}{k!} \sum_{r=0}^{t} q^r r^k \leq \sum_{k=1}^{\infty} \frac{d_t^k}{k!} \sum_{r=0}^{\infty} q^r r^k.$$

The exchange of the sums is justified because all terms are nonnegative. Applying (24), we further obtain

$$\mathcal{E}_t \leq \sum_{k=1}^{\infty} \frac{d_t^k}{k!} \frac{k!}{(1-q)^{k+1}} = \frac{1}{1-q} \sum_{k=1}^{\infty} \left(\frac{d_t}{1-q}\right)^k.$$

Next, we choose t_0 to be

$$t_0 := \left\lceil \left(\frac{2^{a+1}c\eta_0}{(1-a)(1-q)}\right)^{1/a} \right\rceil. \tag{27}$$

Then, for every $t \geq t_0$, it holds that $(t+1)^a \geq \frac{2^{a+1}c\eta_0}{(1-a)(1-q)}$, or equivalently, $\frac{d_t}{1-q} \leq \frac{1}{2}$. For any $x \in [0, 1/2]$, we have $\sum_{k=1}^{\infty} x^k = \frac{x}{1-x} \leq 2x$. We apply this to $\frac{d_t}{1-q}$ to obtain $\mathcal{E}_t \leq \frac{1}{1-q}\frac{2d_t}{1-q} = \frac{2d_t}{(1-q)^2}$. Substituting the definition of d_t, we obtain

$$\mathcal{E}_t \leq \frac{2^{a+1}c}{(1-a)(1-q)^2}\eta_t.$$

Consequently, for $0 < a < 1$, we can take $C_{\mathrm{geo}} = \frac{2^{a+1}c}{(1-a)(1-q)^2}$ and t_0 to be the one in (27).

Next, we consider the case where $a = 1$. In this case, we have $\eta_t = \frac{\eta_0}{t+1}$. From (25), it holds that

$$H_{t,r} \leq 2\eta_0 \int_{t-r+1}^{t+1} \frac{dx}{x} = 2\eta_0 \log\left(\frac{t+1}{t-r+1}\right) = 2\eta_0 \log\left(1 + \frac{r}{t-r+1}\right).$$

Set $m := \lceil 2c\eta_0 \rceil$. We then obtain $e^{cH_{t,r}} \leq \left(1 + \frac{r}{t-r+1}\right)^{2c\eta_0} \leq \left(1 + \frac{r}{t-r+1}\right)^m$. From this, applying the binomial theorem leads to

$$e^{cH_{t,r}} - 1 \leq \sum_{k=1}^{m} \binom{m}{k} \left(\frac{r}{t-r+1}\right)^k.$$

Multiplying by q^r and summing gives

$$\mathcal{E}_t \leq \sum_{k=1}^{m} \binom{m}{k} U_{t,k}, \qquad U_{t,k} := \sum_{r=0}^{t} q^r \left(\frac{r}{t-r+1}\right)^k. \tag{28}$$

We show each $U_{t,k}$ where $1 \leq k \leq m$ is bounded by a constant multiple of $1/t$. For $t \geq 1$, we decompose $U_{t,k}$ into $U_{t,k} = U_{t,k}^{\mathrm{s}} + U_{t,k}^{\mathrm{l}}$ where

$$U_{t,k}^{\mathrm{s}} := \sum_{r=0}^{\lfloor t/2 \rfloor} q^r \left(\frac{r}{t-r+1}\right)^k, \qquad U_{t,k}^{\mathrm{l}} := \sum_{r=\lfloor t/2 \rfloor + 1}^{t} q^r \left(\frac{r}{t-r+1}\right)^k.$$

If $r \le \lfloor t/2 \rfloor$, then it holds that $t - r + 1 \ge \frac{t}{2}$ and consequently $\frac{r}{t-r+1} \le \frac{2r}{t}$. Given this, we obtain for $U^{\mathrm{s}}_{t,k}$ that

$$U^{\mathrm{s}}_{t,k} \le \left(\frac{2}{t}\right)^k \sum_{r=0}^{\lfloor t/2 \rfloor} q^r r^k \le \left(\frac{2}{t}\right)^k \sum_{r=0}^{\infty} q^r r^k \le \frac{2^k k!}{(1-q)^{k+1}}\, t^{-k}.$$

For $t \ge 1$ and $k \ge 1$, it holds that $t^{-k} \le t^{-1}$. Therefore, we conclude that

$$U^{\mathrm{s}}_{t,k} \le \frac{2^k k!}{(1-q)^{k+1}}\, \frac{1}{t}, \qquad \text{when } t \ge 1,\ 1 \le k \le m. \tag{29}$$

For $r > \lfloor t/2 \rfloor$, we have $r > t/2$ and consequently $q^r \le q^{t/2}$. Moreover, given $t - r + 1 \ge 1$, it holds that $(\frac{r}{t-r+1})^k \le r^k \le t^k$. Since there are at most t terms in the sum $U^{\mathrm{l}}_{t,k}$, we have $U^{\mathrm{l}}_{t,k} \le q^{t/2} t^{k+1}$. To make this at most $1/t$ for every $1 \le k \le m$, it is sufficient to require $q^{t/2} t^{m+2} \le 1$. Let $\lambda := -\log q > 0$. This condition is equivalent to $(m+2)\log t \le \frac{\lambda t}{2}$, which is guaranteed whenever $(m+2)\sqrt{t} \le \frac{\lambda t}{2}$ given $\log t \le \sqrt{t}$ for $t \ge 1$. Therefore, we choose t_0 in this case to be

$$t_0 := \left\lceil \left(\frac{2(m+2)}{-\log q} \right)^2 \right\rceil, \qquad m = \lceil 2c\eta_0 \rceil. \tag{30}$$

For every $t \ge t_0$, it holds that $q^{t/2} t^{m+2} \le 1$. Therefore, we obtain for $1 \le k \le m$ that

$$U^{\mathrm{l}}_{t,k} \le q^{t/2} t^{k+1} = \left(q^{t/2} t^{m+2}\right) t^{k-m-1} \le t^{k-m-1} \le \frac{1}{t}. \tag{31}$$

Combining (29) and (31), we obtain

$$U_{t,k} \le \left[\frac{2^k k!}{(1-q)^{k+1}} + 1\right] \frac{1}{t}, \qquad t \ge t_0.$$

Substituting this into (28), we obtain

$$\mathcal{E}_t \le \frac{1}{t} \sum_{k=1}^{m} \binom{m}{k} \left[\frac{2^k k!}{(1-q)^{k+1}} + 1\right], \qquad t \ge t_0.$$

Define $A_m(q) := \sum_{k=1}^{m} \binom{m}{k} [\frac{2^k k!}{(1-q)^{k+1}} + 1]$. Since $t \ge t_0 \ge 1$, we have $\frac{1}{t} \le \frac{2}{t+1} = \frac{2}{\eta_0}\eta_t$. Therefore, we conclude that $\mathcal{E}_t \le \frac{2A_m(q)}{\eta_0}\eta_t$. For $a = 1$, we can take C_{geo} to be

$$C_{\mathrm{geo}} = \frac{2m}{\eta_0} \max_{1 \le k \le m} \binom{m}{k} \left[\frac{2^m m!}{(1-q)^{m+1}} + 1\right],$$

and t_0 to be the one in (30). □

Lemma 10. *For any $u, v \in \mathbb{R}$, we have*

$$e^{-|u-v|} \le \frac{-\ell'(u)}{-\ell'(v)} \le e^{|u-v|}. \tag{32}$$

Consequently, it holds that $\left|\frac{-\ell'(u)}{-\ell'(v)} - 1\right| \le e^{|u-v|} - 1$.

Proof. Define $\phi(x) := \log(-\ell'(x))$. Since $-\ell'(x) = \frac{1}{1+e^x}$, we have $\phi(x) = -\log(1+e^x)$ and differentiating it leads to $\phi'(x) = -\frac{e^x}{1+e^x}$. Because $0 < \frac{e^x}{1+e^x} < 1$, we have $|\phi'(x)| \le 1$. Therefore, the function ϕ is globally 1-Lipschitz. This implies that $|\phi(u) - \phi(v)| \le |u - v|$. Substituting $\phi(x) = \log(-\ell'(x))$, we obtain

$$|\log(-\ell'(u)) - \log(-\ell'(v))| \le |u - v|.$$

This is equivalent to $-|u-v| \le \log \frac{-\ell'(u)}{-\ell'(v)} \le |u-v|$. Exponentiating gives $e^{-|u-v|} \le \frac{-\ell'(u)}{-\ell'(v)} \le e^{|u-v|}$.

Let $h := |u - v| \geq 0$ and $R := \frac{-\ell'(u)}{-\ell'(v)} \geq 0$. We have already shown that $e^{-h} \leq R \leq e^h$. Next, we apply this to show $|R - 1| \leq e^h - 1$. To do this, we consider two cases where $R \geq 1$ and $R < 1$. In the former case, we have

$$|R - 1| = R - 1 \leq e^h - 1.$$

In the latter case, we have

$$|R - 1| = 1 - R \leq 1 - e^{-h} = \frac{e^h - 1}{e^h} \leq e^h - 1,$$

where the last inequality is by $e^h \geq 1$. Combining both cases finishes the proof. □

B.2 Adam's Update and EMA

In this section, we first state and prove a few results on Adam's update. They ultimately lead to the proof of the loss recursion (44) in Proposition 19. In the end, we prove Proposition 2.1.

Lemma 11. *For every $t \geq 0$ and $\epsilon > 0$, Adam's update in (3) satisfies*

$$\|D_t\|_\infty \leq C_D \quad \textit{with} \quad C_D := \sqrt{\frac{1 - \beta_1}{1 - \beta_2}}. \tag{33}$$

Proof. In the case where $\beta_1 = \beta_2 = 0$, we have $|D_t[j]| = \frac{|g_t[j]|}{|g_t[j]+\epsilon|} \leq 1 = C_D$. From now on, we assume that $\beta_2 > 0$. Fix a coordinate j and let $x_s := g_s[j]$ for $s = 0, ..., t$. Given $\bar{m}_{-1} = \bar{v}_{-1} = 0$, we have that $\bar{m}_t[j] = (1 - \beta_1) \sum_{s=0}^t \beta_1^{t-s} x_s$ and $\bar{v}_t[j] = (1 - \beta_2) \sum_{s=0}^t \beta_2^{t-s} x_s^2$. After bias correction, we have that

$$m_t[j] = \frac{1 - \beta_1}{1 - \beta_1^{t+1}} \sum_{s=0}^t \beta_1^{t-s} x_s \quad \text{and} \quad v_t[j] = \frac{1 - \beta_2}{1 - \beta_2^{t+1}} \sum_{s=0}^t \beta_2^{t-s} x_s^2$$

Define $a_{t,s} := \frac{(1-\beta_1)\beta_1^{t-s}}{1-\beta_1^{t+1}}$ and $b_{t,s} := \frac{(1-\beta_2)\beta_2^{t-s}}{1-\beta_2^{t+1}}$, then we have $m_t[j] = \sum_{s=0}^t a_{t,s} x_s$ and $v_t[j] = \sum_{s=0}^t b_{t,s} x_s^2$. Note that $a_{t,s} \geq 0, b_{t,s} \geq 0, \sum_{s=0}^t a_{t,s} = 1$ and $\sum_{s=0}^t b_{t,s} = 1$. When $\beta_1 = 0$, the interpretation is that $a_{t,t} = 1$ and $a_{t,s} = 0$ for $s < t$ (the factor 0^0 is treated as 1). By Jensen's inequality, we have

$$m_t[j]^2 = (\sum_{s=0}^t a_{t,s} x_s)^2 \leq \sum_{s=0}^t a_{t,s} x_s^2.$$

For every $s \in \{0, ..., t\}$, the ratio $\frac{a_{t,s}}{b_{t,s}}$ satisfies

$$\frac{a_{t,s}}{b_{t,s}} = \frac{1 - \beta_1}{1 - \beta_2} \frac{1 - \beta_2^{t+1}}{1 - \beta_1^{t+1}} (\frac{\beta_1}{\beta_2})^{t-s} \leq \frac{1 - \beta_1}{1 - \beta_2} = C_D^2,$$

where we have used $\frac{1-\beta_2^{t+1}}{1-\beta_1^{t+1}} \leq 1$. Consequently, it holds that $a_{t,s} \leq C_D^2 b_{t,s}$. Using this, we have

$$m_t[j]^2 \leq \sum_{s=0}^t a_{t,s} x_s^2 \leq C_D^2 \sum_{s=0}^t b_{t,s} x_s^2 = C_D^2 v_t[j].$$

Taking the square root leads to $|m_t[j]| \leq C_D \sqrt{v_t[j]}$. Consequently, we have

$$|D_t[j]| = \frac{|m_t[j]|}{\sqrt{v_t[j]} + \epsilon} \leq C_D \frac{\sqrt{v_t[j]}}{\sqrt{v_t[j]} + \epsilon} \leq C_D, \forall j.$$

The desired follows by taking the maximum over all the coordinates. □

Lemma 12. *For all $u, v \in \mathbb{R}^d$, we have*

$$e^{-R_1 \|u-v\|_\infty} G(v) \leq G(u) \leq e^{R_1 \|u-v\|_\infty} G(v). \tag{34}$$

Further assume that $0 \leq \beta_1 \leq \beta_2 < 1$. It holds for every $0 \leq r \leq t$ that

$$e^{-R_1 C_D \sum_{s=r}^{t-1} \eta_s} G_r \leq G_t \leq e^{R_1 C_D \sum_{s=r}^{t-1} \eta_s} G_r.$$

Proof. Fix any $i \in [n]$. Define $p_i := z_i^T u$ and $q_i := z_i^T v$. By definition, we have $a_i(u) = -\ell'(p_i)$ and $a_i(v) = -\ell'(q_i)$. Applying Lemma 10 with $a = p_i$ and $b = q_i$ gives

$$e^{-|z_i^T(u-v)|} \le \frac{a_i(u)}{a_i(v)} \le e^{|z_i^T(u-v)|}.$$

Multiplying both sides by $a_i(v) = \frac{1}{1+e^{z_i^T v}} > 0$ leads to

$$e^{-|z_i^T(u-v)|} a_i(v) \le a_i(u) \le e^{|z_i^T(u-v)|} a_i(v).$$

By $|z_i^T(u-v)| \le \|z_i\|_1 \|u-v\|_\infty \le R_1 \|u-v\|_\infty$, we further obtain

$$e^{-R_1\|u-v\|_\infty} a_i(v) \le a_i(u) \le e^{R_1\|u-v\|_\infty} a_i(v).$$

Summing this inequality over i, dividing by n, and using the definition of the function G leads to (34). Applying (34) with $u = w_t$ and $v = w_r$ gives

$$e^{-R_1\|w_t-w_r\|_\infty} G_r \le G_t \le e^{R_1\|w_t-w_r\|_\infty} G_r.$$

By Lemma 11, it holds that for any $t \ge r$

$$\|w_t - w_r\|_\infty = \|-\sum_{s=r}^{t-1} \eta_s D_s\|_\infty \le C_D \sum_{s=r}^{t-1} \eta_s. \tag{35}$$

Consequently, we have $R_1\|w_t - w_r\|_\infty \le R_1 C_D \sum_{s=r}^{t-1} \eta_s$. Combining the preceding bound with (35) finishes the proof. □

Lemma 13. *For every $t \ge 0$ and every $\epsilon > 0$, we have*

$$L(w_{t+1}) \le L(w_t) - \eta_t \langle g_t, D_t \rangle + C_H \eta_t^2 G_t, \tag{36}$$

where $C_H = \frac{R_1^2}{2} \frac{1-\beta_1}{1-\beta_2} \exp(R_1 \eta_0 \sqrt{\frac{1-\beta_1}{1-\beta_2}})$.

Proof. The gradient of the loss is $\nabla L(w) = \frac{1}{n} \sum_{i=1}^n \ell'(z_i^T w) z_i = -\frac{1}{n} \sum_{i=1}^n a_i(w) z_i$, differentiating again leads to the Hessian

$$\nabla^2 L(w) = \frac{1}{n} \sum_i \ell''(z_i^\top w) z_i z_i^\top \quad \text{where} \quad 0 \le \ell''(u) = \frac{e^u}{(1+e^u)^2} \le \frac{1}{1+e^u} = -\ell'(u).$$

Let $\widetilde{w}_\xi := w_t - \xi \eta_t D_t$ for some $\xi \in [0,1]$. By Lemma 11, we have $|z_i^\top(\widetilde{w}_\xi - w_t)| \le R_1 C_D \eta_t \le R_1 C_D \eta_0$. Therefore, Lemma 10 gives

$$-\ell'(z_i^\top \widetilde{w}_\xi) \le e^{|z_i^T \widetilde{w}_\xi - z_i^T w_t|} [-\ell'(z_i^T w_t)] \le e^{R_1 C_D \eta_0} a_i(w_t),$$

where we have used $-\ell'(z_i^T w_t) = a_i(w_t)$. Moreover, it holds that $|z_i^\top D_t| \le \|z_i\|_1 \|D_t\|_\infty \le R_1 C_D$. Consequently, we have

$$D_t^\top \nabla^2 L(\widetilde{w}_\xi) D_t \le \frac{1}{n} \sum_i \big[-\ell'(z_i^\top \widetilde{w}_\xi)\big] (z_i^\top D_t)^2 \le e^{R_1 C_D \eta_0} R_1^2 C_D^2 G_t = M_H G_t, \tag{37}$$

where we have let $M_H := e^{R_1 C_D \eta_0} R_1^2 C_D^2$. Define $\phi(\omega) := L(w_t - \omega \eta_t D_t)$ with $\omega \in [0,1]$. By Taylor's theorem, there exists $\xi_t \in (0,1)$ such that

$$\phi(1) = \phi(0) + \phi'(0) + \frac{1}{2}\phi''(\xi_t).$$

Substituting $\phi(0) = L(w_t)$, $\phi(1) = L(w_{t+1})$, $\phi'(0) = -\eta_t \langle g_t, D_t \rangle$, and $\phi''(\xi_t) = \eta_t^2 D_t^T \nabla^2 L(\widetilde{w}_{\xi_t}) D_t$, we obtain

$$\begin{aligned} L(w_{t+1}) &= L(w_t) - \eta_t \langle g_t, D_t \rangle + \frac{\eta_t^2}{2} D_t^T \nabla^2 L(\widetilde{w}_{\xi_t}) D_t \\ &\le L(w_t) - \eta_t \langle g_t, D_t \rangle + \frac{M_H}{2} \eta_t^2 G_t. \end{aligned}$$

Letting $C_H = \frac{M_H}{2}$ and substituting the definition of C_D finish the proof. □

Lemma 14. *For every $0 \le r \le t$, we have*

$$\|g_{t-r} - g_t\|_1 \le R_1 G_t \left[\exp\left(R_1 C_D \sum_{s=t-r}^{t-1} \eta_s\right) - 1\right]. \tag{38}$$

Proof. By Lemma 11, it holds that

$$\|w_t - w_{t-r}\|_\infty \le \sum_{s=t-r}^{t-1} \eta_s \|D_s\|_\infty \le C_D \sum_{s=t-r}^{t-1} \eta_s.$$

Hence, for every sample $i \in [n]$, we have

$$|z_i^\top (w_t - w_{t-r})| \le \|z_i\|_1 \|w_t - w_{t-r}\|_\infty \le R_1 C_D \sum_{s=t-r}^{t-1} \eta_s =: \Delta_{t,r}.$$

Applying Lemma 10 with $a = z_i^T w_{t-r}$ and $b = z_i^T w_t$, we obtain

$$\left|\frac{-\ell'(z_i^T w_{t-r})}{-\ell'(z_i^T w_t)} - 1\right| = \left|\frac{a_i(w_{t-r})}{a_i(w_t)} - 1\right| \le e^{|z_i^\top (w_t - w_{t-r})|} - 1 \le e^{\Delta_{t,r}} - 1,$$

where we have used $a_i(w_{t-r}) = -\ell'(z_i^T w_{t-r})$ and $a_i(w_t) = -\ell'(z_i^T w_t)$. Consequently, it holds that $|a_i(w_{t-r}) - a_i(w_t)| \le a_i(w_t)(e^{\Delta_{t,r}} - 1)$. By the definition of g, with $g_t := g(w_t), \forall t$, we have

$$\begin{aligned}\|g_{t-r} - g_t\|_1 &= \|-\frac{1}{n}\sum_{i=1}^n [a_i(w_{t-r}) - a_i(w_t)] z_i\|_1 \le \frac{1}{n}\sum_{i=1}^n |a_i(w_{t-r}) - a_i(w_t)| \|z_i\|_1 \\ &\le \frac{R_1}{n}\sum_{i=1}^n |a_i(w_{t-r}) - a_i(w_t)| \le R_1 (e^{\Delta_{t,r}} - 1)\frac{1}{n}\sum_{i=1}^n a_i(w_t) = R_1 G_t (e^{\Delta_{t,r}} - 1).\end{aligned}$$

Substituting $\Delta_{t,r} := R_1 C_D \sum_{s=t-r}^{t-1} \eta_s$ into this inequality finishes the proof. □

Lemma 15. *Let $C_{\text{geo}}(q, c, a, \eta_0)$ be the constant in Proposition 9. There exists a time $t_m = t_m(q = \beta_1, c = R_1 C_D, a, \eta_0)$ such that for all $t \ge t_m$, we have*

$$\|m_t - g_t\|_1 \le C_m \eta_t G_t,$$

where $C_m = R_1 C_{\text{geo}}(q = \beta_1, c = R_1 C_D, a, \eta_0)$.

Proof. If $\beta_1 = 0$, then $m_t = g_t$ and the bound clearly holds. Suppose that $\beta_1 \in (0, 1)$. Similar to Lemma 11, we can write $m_t = \sum_{r=0}^t a_{t,r} g_{t-r}$ with $a_{t,r} = \frac{(1-\beta_1)\beta_1^r}{1-\beta_1^{t+1}} \ge 0$ and $\sum_{r=0}^t a_{t,r} = 1$. Since $1 - \beta_1^{t+1} \ge 1 - \beta_1$, we have $a_{t,r} \le \beta_1^r$. These lead to

$$\begin{aligned}\|m_t - g_t\|_1 &= \|\sum_{r=0}^t a_{t,r}(g_{t-r} - g_t)\|_1 \le \sum_{r=0}^t a_{t,r}\|g_{t-r} - g_t\|_1 \\ &\le \sum_{r=0}^t \beta_1^r \|g_{t-r} - g_t\|_1.\end{aligned}$$

Applying Lemma 14, we obtain

$$\|m_t - g_t\|_1 \le R_1 G_t \sum_{r=0}^t \beta_1^r [\exp(R_1 C_D \sum_{s=t-r}^{t-1} \eta_s) - 1] \le C_m \eta_t G_t, \quad \forall t \ge t_m.$$

where $C_m = R_1 C_{\text{geo}}(q, c, a, \eta_0)$ with $q = \beta_1$ and $c = R_1 C_D$. □

Lemma 16. *Let $C_{\text{geo}}(q, c, a, \eta_0)$ be the constant in Proposition 9. There exists a time $t_v = t_v(q = \sqrt{\beta_2}, c = R_1 C_D, a, \eta_0)$ such that for all $t \ge t_v$, we have*

$$\|\sqrt{v_t} - |g_t|\|_1 \le C_v \eta_t G_t.$$

where $C_v = R_1 C_{\text{geo}}(q = \sqrt{\beta_2}, c = R_1 C_D, a, \eta_0)$.

Proof. If $\beta_2 = 0$, then $v_t = g_t^{\odot 2}$ and the left side is zero and the bound trivially holds. Suppose that $\beta_2 \in (0, 1)$. We define the normalized EMA weights

$$b_{t,r} := \frac{(1-\beta_2)\beta_2^r}{1-\beta_2^{t+1}}, \qquad r = 0, \dots, t,$$

and note that $\sum_{r=0}^{t} b_{t,r} = 1$. Moreover, it holds that $\bar{v}_t = \beta_2 \bar{v}_{t-1} + (1-\beta_2) g_t^{\odot 2} = (1-\beta_2)\sum_{r=0}^{t} \beta_2^r g_{t-r}^{\odot 2}$. After bias-correction, this becomes $v_t = \sum_{r=0}^{t} b_{t,r} g_{t-r}^{\odot 2}$. For a fixed feature coordinate j, we introduce two vectors in $\mathbb{R}^{t+1}$,

$$X^{(j)} = (\sqrt{b_{t,r}} g_{t-r}[j])_{r=0}^{t}, \qquad Y^{(j)} = (\sqrt{b_{t,r}} g_t[j])_{r=0}^{t}.$$

Then, we have

$$\|X^{(j)}\|_2^2 = \sum_{r=0}^{t} b_{t,r} g_{t-r}[j]^2 = v_t[j], \quad \|Y^{(j)}\|_2^2 = \sum_{r=0}^{t} b_{t,r} g_t[j]^2 = g_t[j]^2 \sum_{r=0}^{t} b_{t,r} = g_t[j]^2.$$

Consequently, it holds that $\|X^{(j)}\|_2 = \sqrt{v_t[j]}$ and $\|Y^{(j)}\|_2 = |g_t[j]|$. Reverse triangle inequality gives

$$\begin{aligned}
\left| \sqrt{v_t[j]} - |g_t[j]| \right| &\le \|X^{(j)} - Y^{(j)}\|_2 \\
&= \left(\sum_{r=0}^{t} b_{t,r} |g_{t-r}[j] - g_t[j]|^2 \right)^{1/2} \\
&\le \sum_{r=0}^{t} \sqrt{b_{t,r}} |g_{t-r}[j] - g_t[j]|,
\end{aligned}$$

where the last inequality is by applying the inequality $(\sum_{r=0}^{t} x_r^2)^{1/2} \le \sum_{r=0}^{t} x_r$ (for any nonnegative numbers $x_0, ..., x_t$) with $x_r = \sqrt{b_{t,r}} |g_{t-r}[j] - g_t[j]|$. Because $1 - \beta_2^{t+1} \ge 1 - \beta_2$, we have $\sqrt{b_{t,r}} = \sqrt{\frac{1-\beta_2}{1-\beta_2^{t+1}}} \beta_2^{r/2} \le \beta_2^{r/2}$. Summing the preceding inequality over $j = 1, ..., d$ yields

$$\begin{aligned}
\|\sqrt{v_t} - |g_t|\|_1 &= \sum_{j=1}^{d} \left| \sqrt{v_t[j]} - |g_t[j]| \right| \le \sum_{j=1}^{d} \sum_{r=0}^{t} \beta_2^{r/2} |g_{t-r}[j] - g_t[j]| \\
&= \sum_{r=0}^{t} \beta_2^{r/2} \sum_{j=1}^{d} |g_{t-r}[j] - g_t[j]| = \sum_{r=0}^{t} \beta_2^{r/2} \|g_{t-r} - g_t\|_1.
\end{aligned}$$

Now applying Lemma 14 and Proposition 9 (with $q = \sqrt{\beta_2}$ and $c = R_1 C_D$), we obtain

$$\|\sqrt{v_t} - |g_t|\|_1 \le R_1 G_t \sum_{r=0}^{t} (\sqrt{\beta_2})^r \left[\exp(R_1 C_D \sum_{s=t-r}^{t-1} \eta_s) - 1 \right] \le C_v \eta_t G_t, \quad \forall t \ge t_v,$$

where $C_v = R_1 C_{\text{geo}}(q = \sqrt{\beta_2}, c = R_1 C_D, a, \eta_0)$. □

Lemma 17. *Let $t_{\text{ema}} := \max\{t_m, t_v\}$ where t_m and t_v are the times in Lemma 15 and Lemma 16 respectively. For all $t \ge t_{\text{ema}}$ and all $\epsilon > 0$, we have*

$$\|g_t \odot (D_t - \sigma_\epsilon(g_t))\|_1 \le C_{\text{ema}} \eta_t G_t, \tag{39}$$

$$\epsilon \|D_t - \sigma_\epsilon(g_t)\|_2 \le C_{\text{ema}} \eta_t G_t, \tag{40}$$

$$|\langle g_t, D_t - \sigma_\epsilon(g_t) \rangle| \le C_{\text{ema}} \eta_t G_t, \tag{41}$$

where $C_{\text{ema}} = \max\{C_m + C_D C_v, C_m + C_v\}$. Recall that C_D,C_m,C_v are the constants in Lemmas 11, 15 and 16 respectively.

Proof. Fix a coordinate j, we let $a := g_t[j]$, $b := m_t[j]$, and $c := \sqrt{v_t[j]}$. We have $\sigma_\epsilon(g_t)[j] = \frac{a}{|a|+\epsilon}$, $D_t[j] = \frac{b}{c+\epsilon}$, and we further define $\triangle_j := D_t[j] - \sigma_\epsilon(g_t)[j] = \frac{b}{c+\epsilon} - \frac{a}{|a|+\epsilon}$. We first have the

following equality

$$\frac{a}{|a|+\epsilon} - \frac{b}{c+\epsilon} = \frac{a(c+\epsilon) - b(|a|+\epsilon)}{(|a|+\epsilon)(c+\epsilon)} = \frac{(a-b)(c+\epsilon) + b(c-|a|)}{(|a|+\epsilon)(c+\epsilon)}.$$

Multiplying by $|a|$, applying triangle inequality, and bounding the two terms separately gives

$$|a|\left|\frac{a}{|a|+\epsilon} - \frac{b}{c+\epsilon}\right| \le \frac{|a|}{|a|+\epsilon}|a-b| + \frac{|a|}{|a|+\epsilon}\frac{|b|}{c+\epsilon}|c-|a|| \le |a-b| + C_D|c-|a||,$$

where the second inequality is by $0 \le \frac{|a|}{|a|+\epsilon} \le 1$ and $\frac{|b|}{c+\epsilon} = |D_t[j]| \le C_D$ by Lemma 11. Consequently, we have $|a||\triangle_j| \le |a-b| + C_D\big|c-|a|\big|$. Applying this inequality, we obtain

$$\begin{aligned}
\|g_t \odot (D_t - \sigma_\epsilon(g_t))\|_1 &= \sum_{j=1}^d |g_t[j]||D_t[j] - \sigma_\epsilon(g_t)[j]| \\
&\le \sum_{j=1}^d |g_t[j] - m_t[j]| + C_D \sum_{j=1}^d \big|\sqrt{v_t[j]} - |g_t[j]|\big| \\
&= \|m_t - g_t\|_1 + C_D\|\sqrt{v_t} - |g_t|\|_1 \\
&\le C_m\eta_t G_t + C_D C_v \eta_t G_t \\
&= (C_m + C_D C_v)\eta_t G_t, \quad \forall t \ge t_{\text{ema}},
\end{aligned}$$

where we have applied Lemmas 15 and 16 and let $t_{\text{ema}} := \max\{t_v, t_m\}$. The inner-product inequality (41) follows immediately from

$$|\langle g_t, D_t - \sigma_\epsilon(g_t)\rangle| \le \|g_t \odot (D_t - \sigma_\epsilon(g_t))\|_1.$$

Next, we prove a different coordinatewise inequality for (40)

$$\frac{b}{c+\epsilon} - \frac{a}{|a|+\epsilon} = \frac{b-a}{c+\epsilon} + [\frac{a}{c+\epsilon} - \frac{a}{|a|+\epsilon}] = \frac{b-a}{c+\epsilon} + \frac{a(|a|-c)}{(c+\epsilon)(|a|+\epsilon)}.$$

Multiplying by ϵ, taking absolute values, and using the triangle inequality gives

$$\epsilon\left|\frac{b}{c+\epsilon} - \frac{a}{|a|+\epsilon}\right| \le \frac{\epsilon}{c+\epsilon}|b-a| + \frac{\epsilon}{c+\epsilon}\frac{|a|}{|a|+\epsilon}|c-|a|| \le |b-a| + |c-|a||,$$

where in the last inequality we used $0 \le \frac{\epsilon}{c+\epsilon} \le 1$ and $0 \le \frac{|a|}{|a|+\epsilon} \le 1$. By substituting the definition of $\triangle_j$, we have $\epsilon|\triangle_j| \le |b-a| + |c-|a||$. Based on this, we obtain

$$\begin{aligned}
\epsilon\|D_t - \sigma_\epsilon(g_t)\|_2 \le \epsilon\|D_t - \sigma_\epsilon(g_t)\|_1 &= \sum_{j=1}^d \epsilon|D_t[j] - \sigma_\epsilon(g_t)[j]| \\
&\le \sum_{j=1}^d |m_t[j] - g_t[j]| + \sum_{j=1}^d \big|\sqrt{v_t[j]} - |g_t[j]|\big| \\
&= \|m_t - g_t\|_1 + \|\sqrt{v_t} - |g_t|\|_1 \\
&\le C_m\eta_t G_t + C_v\eta_t G_t = (C_m + C_v)\eta_t G_t, \quad \forall t \ge t_{\text{ema}}.
\end{aligned}$$

Letting $C_{\text{ema}} = \max\{C_m + C_D C_v, C_m + C_v\}$ finishes the proof. □

In Lemma 18, (42) is used in Part 1 of the proofs of Theorem 4 and Theorem 29, where $G_t/\epsilon \to \infty$. Its purpose is to obtain the sharp coefficient γ_∞, whereas (43) is used in Part 2 of both proofs.

Lemma 18. *For every $g \in \mathbb{R}^d$ and $\epsilon > 0$, we have*

$$\langle g, \sigma_\epsilon(g) \rangle \geq \|g\|_1 - d\epsilon, \tag{42}$$

$$\langle g, \sigma_\epsilon(g) \rangle \geq \frac{\|g\|_1^2}{\|g\|_1 + d\epsilon}. \tag{43}$$

Proof. When $\epsilon = 0$, we have $\sigma_0(g) = \mathrm{sign}(g)$ gives $\langle g, \sigma_0(g) \rangle = \|g\|_1$. Hence, both inequalities hold. Suppose now that $\epsilon > 0$. For each coordinate, we have

$$\frac{g_j^2}{|g_j| + \epsilon} = \frac{|g_j|(|g_j| + \epsilon) - \epsilon|g_j|}{|g_j| + \epsilon} = |g_j| - \epsilon \frac{|g_j|}{|g_j| + \epsilon}.$$

Summing over j gives

$$\begin{aligned}\langle g, \sigma_\epsilon(g) \rangle &= \sum_{j=1}^d g_j \sigma_\epsilon(g)_j = \sum_{j=1}^d \frac{g_j^2}{|g_j| + \epsilon} \\ &= \sum_{j=1}^d |g_j| - \epsilon \sum_{j=1}^d \frac{|g_j|}{|g_j| + \epsilon} = \|g\|_1 - \epsilon \sum_{j=1}^d \frac{|g_j|}{|g_j| + \epsilon}.\end{aligned}$$

For every j, it holds that $0 \leq \frac{|g_j|}{|g_j|+\epsilon} \leq 1$; therefore, we have $\langle g, \sigma_\epsilon(g) \rangle \geq \|g\|_1 - d\epsilon$. Let $x_j := |g_j|$ and $y_j := |g_j| + \epsilon$ (note that $y_j > 0, \forall j \in [d]$). From $\sum_{j=1}^d x_j = \sum_{j=1}^d \frac{x_j}{\sqrt{y_j}} \sqrt{y_j}$, we have $(\sum_{j=1}^d x_j)^2 \leq (\sum_{j=1}^d \frac{x_j^2}{y_j})(\sum_{j=1}^d y_j)$ by Cauchy-Schwarz. Rearranging this leads to $\sum_{j=1}^d \frac{x_j^2}{y_j} \geq \frac{(\sum_j x_j)^2}{\sum_j y_j}$. Substituting the definitions of x_j and y_j, we obtain

$$\sum_j \frac{|g_j|^2}{|g_j| + \epsilon} \geq \frac{(\sum_j |g_j|)^2}{\sum_j (|g_j| + \epsilon)} = \frac{\|g\|_1^2}{\|g\|_1 + d\epsilon}.$$

□

Proposition 19. *The following holds for every $\epsilon > 0$,*

$$L_{t+1} \leq L_t - \eta_t \langle g_t, \sigma_\epsilon(g_t) \rangle + C_{\mathrm{md}} \eta_t^2 G_t, \quad t \geq t_{\mathrm{ema}}, \tag{44}$$

where t_{ema} is defined in Lemma 17 and $C_{\mathrm{md}} := C_H + C_{\mathrm{ema}}$. Recall that C_H and C_{ema} are the constants in Lemmas 13 and 17 respectively.

Proof. Let $\delta_t := \langle g_t, D_t - \sigma_\epsilon(g_t) \rangle$. By Lemma 17, we have $|\delta_t| \leq C_{\mathrm{ema}} \eta_t G_t, \forall t \geq t_{\mathrm{ema}}$. Consequently, it holds that $\delta_t \geq -C_{\mathrm{ema}} \eta_t G_t$ which implies

$$\begin{aligned}\langle g_t, D_t \rangle &= \langle g_t, \sigma_\epsilon(g_t) + D_t - \sigma_\epsilon(g_t) \rangle = \langle g_t, \sigma_\epsilon(g_t) \rangle + \delta_t \\ &\geq \langle g_t, \sigma_\epsilon(g_t) \rangle - C_{\mathrm{ema}} \eta_t G_t, \quad \forall t \geq t_{\mathrm{ema}},\end{aligned}$$

From this, Lemma 13 gives

$$\begin{aligned}L_{t+1} &\leq L_t - \eta_t \langle g_t, D_t \rangle + C_H \eta_t^2 G_t \\ &\leq L_t - \eta_t \langle g_t, \sigma_\epsilon(g_t) \rangle + C_{\mathrm{ema}} \eta_t^2 G_t + C_H \eta_t^2 G_t \\ &= L_t - \eta_t \langle g_t, \sigma_\epsilon(g_t) \rangle + (C_H + C_{\mathrm{ema}}) \eta_t^2 G_t, \quad \forall t \geq t_{\mathrm{ema}}.\end{aligned}$$

Letting $C_{\mathrm{md}} := C_H + C_{\mathrm{ema}}$ finishes the proof. □

Proof of Proposition 2.1. Fix any $\epsilon \in (0, 1]$. Combining (19) and (43), we obtain

$$\langle g_t, \sigma_\epsilon(g_t) \rangle \geq \frac{\|g_t\|_1^2}{\|g_t\|_1 + d\epsilon} \geq \frac{\gamma_\infty^2}{R_1 + d} G_t^2,$$

where the last step uses $G_t \leq 1$ and $\epsilon \leq 1$. Define $\kappa := \frac{\gamma_\infty^2}{R_1 + d}$. Substituting this into (44), we have

$$L_{t+1} \leq L_t - \kappa \eta_t G_t^2 + C_{\mathrm{md}} \eta_t^2 G_t.$$

By Young's inequality, we have $C_{\mathrm{md}}\eta_t^2 G_t = (\sqrt{\kappa\eta_t}G_t)(\frac{C_{\mathrm{md}}}{\sqrt{\kappa}}\eta_t^{3/2}) \le \frac{\kappa}{2}\eta_t G_t^2 + \frac{C_{\mathrm{md}}^2}{2\kappa}\eta_t^3$. Consequently, it holds that

$$L_{t+1} \le L_t - c\eta_t G_t^2 + K\eta_t^3, \qquad \forall t \ge t_{\mathrm{ema}}, \tag{45}$$

where $c := \frac{\kappa}{2} = \frac{\gamma_\infty^2}{2(R_1+d)}$ and $K := \frac{C_{\mathrm{md}}^2}{2\kappa} = \frac{(R_1+d)C_{\mathrm{md}}^2}{2\gamma_\infty^2}$. Since $\eta_t = \frac{\eta_0}{(t+1)^a}$ with $a \in (\frac{1}{3}, 1]$, we have $\sum_{t=0}^\infty \eta_t^3 = \eta_0^3 \sum_{t=0}^\infty (t+1)^{-3a} < \infty$. Define $R_t := \sum_{s=t}^\infty \eta_s^3$. We then have $R_{t+1} = R_t - \eta_t^3$ and $R_t \downarrow 0$. This holds because

$$R_t = \eta_0^3 \sum_{s=t}^\infty \frac{1}{(s+1)^{3a}} \le \eta_0^3 \int_t^\infty x^{-3a} dx = \frac{\eta_0^3}{3a-1} t^{1-3a},$$

and $t^{1-3a} \to 0$ as $1 - 3a < 0$. Further define $\Phi_t := L_t + KR_t, \forall t$. Using (45), we obtain

$$\begin{aligned}
\Phi_{t+1} &= L_{t+1} + KR_{t+1} \\
&\le L_t - c\eta_t G_t^2 + K\eta_t^3 + K(R_t - \eta_t^3) \\
&= L_t + KR_t - c\eta_t G_t^2 \\
&= \Phi_t - c\eta_t G_t^2, \qquad \forall t \ge t_{\mathrm{ema}}.
\end{aligned} \tag{46}$$

Note that Φ_t is non-increasing for $t \ge t_{\mathrm{ema}}$. Rearranging and summing from $t = t_{\mathrm{ema}}$ to $T-1$, we obtain

$$c \sum_{t=t_{\mathrm{ema}}}^{T-1} \eta_t G_t^2 \le \Phi_{t_{\mathrm{ema}}} - \Phi_T \le \Phi_{t_{\mathrm{ema}}}. \tag{47}$$

Taking $T \to \infty$ leads to $\sum_{t=t_{\mathrm{ema}}}^\infty \eta_t G_t^2 \le \frac{\Phi_{\mathrm{ema}}}{c} < \infty$. The prefix from $t = 0$ to $t_{\mathrm{ema}} - 1$ contains only finitely many terms. Therefore, we conclude that $\sum_{t=0}^\infty \eta_t G_t^2 < \infty$. This holds for every $\epsilon \in (0, 1]$.

Next, we show that $G_t \to 0$. As $a \le 1$, we have $\sum_{t=0}^\infty \eta_t = \infty$. Suppose that $\liminf_{t\to\infty} G_t > 0$. Then there exists $\delta > 0$ and t_δ such that $G_t \ge \delta$ for $t \ge t_\delta$. Consequently we have $\sum_{t=t_\delta}^\infty \eta_t G_t^2 \ge \delta^2 \sum_{t=t_\delta}^\infty \eta_t = \infty$, which contradicts $\sum_{t=0}^\infty \eta_t G_t^2 < \infty$. Hence, we must have $\liminf_{t\to\infty} G_t = 0$. Since Φ_t is non–increasing and bounded below by zero, there exists $\Phi_\infty \ge 0$ such that $\Phi_t \to \Phi_\infty$. Since $R_t \to 0$, we further have $L_t = \Phi_t - KR_t \to \Phi_\infty$. Therefore, L_t has a limit, which we denote by L_∞ (note that $L_\infty = \Phi_\infty$). Given $\liminf_{t\to\infty} G_t = 0$, there exists a subsequence $t_k \to \infty$ such that $G_{t_k} \to 0$. Hence, we have $G_{t_k} \le \frac{1}{2n}$ for all sufficiently large k. By Lemma 8, we have $L_{t_k} \le 2G_{t_k}$. Therefore, $L_{t_k} \to 0$ given $G_{t_k} \to 0$. Given the entire sequence L_t has the limit L_∞, every subsequence must have the same limit. This implies that $L_\infty = 0$. Therefore, we conclude that $L_t \to 0$. Moreover, given $0 < G_t \le L_t$ by Lemma 8, we further obtain $G_t \to 0$.

Next, we construct a time T_* that works for every $\epsilon \in (0, 1]$. Let $t_0 := t_{\mathrm{ema}}$. By Lemma 11, we have $\|w_{t_0}\|_\infty = \|w_0 - \sum_{s=0}^{t_0-1} \eta_s D_s\|_\infty \le \|w_0\|_\infty + C_D \sum_{s=0}^{t_0-1} \eta_s$. Define $M_0 := \|w_0\|_\infty + C_D S_{t_0}$. Then, we have $\|w_{t_0}\|_\infty \le M_0$ uniformly in ϵ. Note that $S_{t_0} \le \eta_0 t_0$ and $M_0 \le \|w_0\|_\infty + C_D\eta_0 t_0$. For each sample, we have $|z_i^T w_{t_0}| \le \|z_i\|_1 \|w_{t_0}\|_\infty \le R_1 M_0$. This leads to $l(z_i^T w_{t_0}) \le \log(1 + e^{R_1 M_0})$ and by averaging, we further obtain $L_{t_0} \le B_L$ where $B_L := \log(1 + e^{R_1 M_0})$. Therefore, we obtain

$$\Phi_{t_0} = L_{t_0} + KR_{t_0} \le B, \qquad B := B_L + KR_{t_0} < \infty. \tag{48}$$

Importantly, B is independent of ϵ. Let $\ell_i := \ell(z_i^T w)$ and $a_i(w) := 1 - e^{-\ell_i}$. Recall that $L(w) = \frac{1}{n}\sum_{i=1}^n \ell_i$ and $G(w) = \frac{1}{n}\sum_{i=1}^n a_i(w)$. Given $L(w)$ is the average of the ℓ_i's, there is at least one index i_* such that $\ell_{i_*} \ge L(w)$. Given the function $x \to 1 - e^{-x}$ is increasing, we have

$$a_{i_*}(w) = 1 - e^{-\ell_{i_*}} \ge 1 - e^{-L(w)}.$$

Therefore, we have

$$G(w) = \frac{1}{n}\sum_{i=1}^n a_i(w) \ge \frac{1}{n} a_{i_*}(w) \ge \frac{1}{n}(1 - e^{-L(w)}). \tag{49}$$

Define $h := \frac{\log 2}{n}$. We can choose a deterministic time $t_1 \geq t_0$ such that $KR_{t_1} \leq \frac{h}{2}$. Concretely, we have via integral approximation that

$$R_t = \eta_0^3 \sum_{s=t}^{\infty} \frac{1}{(s+1)^{3a}} = \eta_0^3 \sum_{k=t+1}^{\infty} k^{-3a} \leq \eta_0^3 \int_t^{\infty} x^{-3a} dx = \frac{\eta_0^3}{3a-1} t^{1-3a}, \qquad t \geq 1.$$

Therefore, $KR_t \leq \frac{h}{2}$ is guaranteed whenever $t^{3a-1} \geq \frac{2K\eta_0^3}{(3a-1)h}$. We define $t_1 := \max\{t_0, 1, \lceil (\frac{2K\eta_0^3}{(3a-1)h})^{\frac{1}{3a-1}} \rceil\}$. Then, it holds that $KR_t \leq \frac{h}{2}, \forall t \geq t_1$. Note that t_1 does not depend on ϵ.

Suppose that $t \geq t_1$ and $\Phi_t > h$. Then, we have

$$L_t = \Phi_t - KR_t > h - KR_{t_1} \geq \frac{h}{2},$$

where we used $R_t \leq R_{t_1}$. By (49), we have

$$G_t \geq \frac{1}{n}(1 - e^{-h/2}) = \frac{1}{n}(1 - \exp(-\frac{\log 2}{2n})) =: \bar{g} > 0.$$

Therefore, $\Phi_t > h, t \geq t_1$ implies that $G_t \geq \bar{g}$. By (46), we have the following whenever $t \geq t_1$ and $\Phi_t > h$,

$$\Phi_{t+1} \leq \Phi_t - c\bar{g}^2 \eta_t. \tag{50}$$

Since $\sum_t \eta_t = \infty$, we can choose a deterministic time $T_* > t_1$ such that $c\bar{g}^2 \sum_{t=t_1}^{T_*-1} \eta_t > B$. Concretely, in the case where $1/3 < a < 1$, we have

$$\sum_{t=t_1}^{T-1} \eta_t = \eta_0 \sum_{k=t_1+1}^{T} k^{-a} \geq \eta_0 \int_{t_1+1}^{T+1} x^{-a} dx = \frac{\eta_0}{1-a}[(T+1)^{1-a} - (t_1+1)^{1-a}].$$

Therefore, it is sufficient to require $\frac{c\bar{g}^2\eta_0}{1-a}[(T_*+1)^{1-a} - (t_1+1)^{1-a}] > B$. Solving for T_* gives

$$T_* = \lceil [(t_1+1)^{1-a} + \frac{(1-a)B}{c\bar{g}^2\eta_0}]^{\frac{1}{1-a}} \rceil, \qquad \frac{1}{3} < a < 1.$$

For the case of $a = 1$, we have

$$\sum_{t=t_1}^{T-1} \eta_t = \eta_0 \sum_{k=t_1+1}^{T} \frac{1}{k} \geq \eta_0 \int_{t_1+1}^{T+1} \frac{dx}{x} = \eta_0 \log(\frac{T+1}{t_1+1}).$$

Therefore, it is sufficient to require $c\bar{g}^2\eta_0 \log(\frac{T^*+1}{t_1+1}) > B$. Solving for T_* leads to

$$T_* = \lceil (t_1+1)\exp(\frac{B}{c\bar{g}^2\eta_0}) \rceil, \qquad a = 1.$$

We now show that Φ_t must fall below h by time T_*. Suppose that $\Phi_t > h$ for $t = t_1, ..., T_*$. Then iterating (50), we obtain

$$\Phi_{T_*} \leq \Phi_{t_1} - c\bar{g}^2 \sum_{t=t_1}^{T_*-1} \eta_t \leq B - c\bar{g}^2 \sum_{t=t_1}^{T_*-1} \eta_t < 0,$$

which is impossible because $\Phi_{T_*} \geq 0$. Therefore, there exists some $\tau \leq T_*$ such that $\Phi_\tau \leq h$. Since Φ_t is non-increasing, we have $\Phi_t \leq h = \frac{\log 2}{n}, t \geq T_*$. Because $L_t \leq \Phi_t$, we obtain

$$L_t \leq \frac{\log 2}{n}, \qquad t \geq T_*. \tag{51}$$

Importantly, T_* is independent of ϵ and it works for every $\epsilon \in (0, 1]$. By Lemma 8, we always have $G_t \leq L_t$. By (51) and Lemma 8, we have $L_t \leq 2G_t$. Consequently, we have $\frac{1}{2}L_t \leq G_t \leq L_t$ for every $t \geq T_*$.

Finally, Lemma 12 gives $G_t \geq G_r \exp(-R_1 C_D \sum_{s=r}^{t-1} \eta_s)$ for $0 \leq r \leq t$. Applying this with $r = 0$ and $t = T_*$, we obtain

$$G_{T_*} \geq G_0 \exp(-R_1 C_D S_{T_*}) =: g_* > 0.$$

The last inequality holds because $G_0 > 0$. Note that g_* is independent of ϵ. □

C Sign and Gradient Residual Characterization

Throughout this section, we assume data separability, $0 \le \beta_1 \le \beta_2 < 1$, and $\epsilon \in (0, 1]$. Recall that $R_1 := \max_{i\in[n]} \|x_i\|_1$ and $R_2 := \max_{i\in[n]} \|x_i\|_2$. The step size schedule is $\eta_t = \eta_0(t+1)^{-a}$ with $\eta_0 > 0$ and $a \in (1/3, 1]$.

Proposition C.1. *Denote* $C_{\mathrm{loc}} := C_{\mathrm{ema}} \max\{\frac{1}{\gamma_\infty}, \frac{1}{\gamma_2}\}$. *The following holds for every* $\epsilon > 0$ *and* $t \ge t_{\mathrm{ema}}$

$$\left[\frac{1}{1+R_1\lambda_t} - C_{\mathrm{loc}}\eta_t\right]_+ \le \mathfrak{r}_t^{\mathrm{sgn}} \le \min\left\{1, \frac{d}{\gamma_\infty \lambda_t}\right\} + C_{\mathrm{loc}}\eta_t, \tag{52}$$

$$\left[1 - \frac{\sqrt{d}}{\gamma_2\lambda_t} - C_{\mathrm{loc}}\eta_t\right]_+ \le \mathfrak{r}_t^{\mathrm{grad}} \le \frac{R_1\lambda_t}{1+R_1\lambda_t} + C_{\mathrm{loc}}\eta_t, \tag{53}$$

where $[x]_+ := \max\{x, 0\}$. *Consequently, for every family* $t = t(\epsilon)$ *with* $t(\epsilon) \to \infty$ *as* $\epsilon \downarrow 0$, *we have* $\lambda_{t(\epsilon)} \to \infty$ *implying* $(\mathfrak{r}_{t(\epsilon)}^{\mathrm{sgn}}, \mathfrak{r}_{t(\epsilon)}^{\mathrm{grad}}) \to (0, 1)$, *and* $\lambda_{t(\epsilon)} \to 0$ *implying* $(\mathfrak{r}_{t(\epsilon)}^{\mathrm{sgn}}, \mathfrak{r}_{t(\epsilon)}^{\mathrm{grad}}) \to (1, 0)$.

Proof of Proposition C.1. Fix any $\epsilon > 0$ and $t \ge t_{\mathrm{ema}}$, and abbreviate $g := g_t$, $D := D_t$, $G := G_t$, and $\lambda := \lambda_t = G_t/\epsilon$. We introduce the corresponding soft-sign residuals

$$\bar{\mathfrak{r}}_t^{\mathrm{sgn}} := \frac{\|g \odot (\sigma_\epsilon(g) - \mathrm{sign}(g))\|_1}{\|g\|_1}, \qquad \bar{\mathfrak{r}}_t^{\mathrm{grad}} := \frac{\|\epsilon\sigma_\epsilon(g) - g\|_2}{\|g\|_2}.$$

Let $A := g \odot (D - \mathrm{sign}(g))$ and $B := g \odot (\sigma_\epsilon(g) - \mathrm{sign}(g))$. Then we have $A - B = g \odot (D - \sigma_\epsilon(g))$, $\frac{\|A\|_1}{\|g\|_1} = \mathfrak{r}_t^{\mathrm{sgn}}$, and $\frac{\|B\|_1}{\|g\|_1} = \bar{\mathfrak{r}}_t^{\mathrm{sgn}}$. The reverse triangle inequality $\big|\|A\|_1 - \|B\|_1\big| \le \|A - B\|_1$ gives

$$|\mathfrak{r}_t^{\mathrm{sgn}} - \bar{\mathfrak{r}}_t^{\mathrm{sgn}}| = \frac{|\|A\|_1 - \|B\|_1|}{\|g\|_1} \le \frac{\|g \odot (D - \sigma_\epsilon(g))\|_1}{\|g\|_1} \le \frac{C_{\mathrm{ema}}}{\gamma_\infty}\eta_t,$$

where we used $\|g\|_1 \ge \gamma_\infty G_t$ and $\|g_t \odot (D_t - \sigma_\epsilon(g_t))\|_1 \le C_{\mathrm{ema}}\eta_t G_t$ by Lemma 17. Similarly, we let $\tilde{A} := \epsilon D - g$ and $\tilde{B} := \epsilon\sigma_\epsilon(g) - g$. Then, we have $\tilde{A} - \tilde{B} = \epsilon(D - \sigma_\epsilon(g))$, $\frac{\|\tilde{A}\|_2}{\|g\|_2} = \mathfrak{r}_t^{\mathrm{grad}}$, and $\frac{\|\tilde{B}\|_2}{\|g\|_2} = \bar{\mathfrak{r}}_t^{\mathrm{grad}}$. The reverse triangle inequality $\big|\|\tilde{A}\|_2 - \|\tilde{B}\|_2\big| \le \|\tilde{A} - \tilde{B}\|_2$ gives

$$|\mathfrak{r}_t^{\mathrm{grad}} - \bar{\mathfrak{r}}_t^{\mathrm{grad}}| = \frac{|\|\tilde{A}\|_2 - \|\tilde{B}\|_2|}{\|g\|_2} \le \frac{\epsilon\|D - \sigma_\epsilon(g)\|_2}{\|g\|_2} \le \frac{C_{\mathrm{ema}}}{\gamma_2}\eta_t,$$

where we used $\|g\|_2 \ge \gamma_2 G_t$ and $\epsilon\|D_t - \sigma_\epsilon(g_t)\|_2 \le C_{\mathrm{ema}}\eta_t G_t$ by Lemma 17. Combining the previous bounds, we obtain

$$\max\{|\mathfrak{r}_t^{\mathrm{sgn}} - \bar{\mathfrak{r}}_t^{\mathrm{sgn}}|, |\mathfrak{r}_t^{\mathrm{grad}} - \bar{\mathfrak{r}}_t^{\mathrm{grad}}|\} \le C_0\eta_t \quad \text{where} \quad C_0 := C_{\mathrm{ema}} \max\{\frac{1}{\gamma_\infty}, \frac{1}{\gamma_2}\}. \tag{54}$$

Define $C_{\mathrm{loc}} := C_0 = C_{\mathrm{ema}} \max\{\frac{1}{\gamma_\infty}, \frac{1}{\gamma_2}\}$. Next, we upper and lower bound $\bar{\mathfrak{r}}_t^{\mathrm{sgn}}$ (and consequently $\mathfrak{r}_t^{\mathrm{sgn}}$ by (54)). For a coordinate with $g_j \ne 0$, we have $\sigma_\epsilon(g)_j = \mathrm{sign}(g_j)\frac{|g_j|}{|g_j|+\epsilon}$. This gives $\sigma_\epsilon(g)_j - \mathrm{sign}(g_j) = -\mathrm{sign}(g_j)\frac{\epsilon}{|g_j|+\epsilon}$ and consequently $|g_j||\sigma_\epsilon(g)_j - \mathrm{sign}(g_j)| = \frac{\epsilon|g_j|}{|g_j|+\epsilon}$. The last equality also holds when $g_j = 0$. Summing over j gives $\bar{\mathfrak{r}}_t^{\mathrm{sgn}} = \frac{\epsilon}{\|g\|_1}\sum_{j=1}^d \frac{|g_j|}{|g_j|+\epsilon}$. Let $p_j := \frac{|g_j|}{\|g\|_1}$. Note that $p_j \ge 0, \forall j$ and $\sum_{j=1}^d p_j = 1$ as $g \ne 0$. From this, we obtain

$$\bar{\mathfrak{r}}_t^{\mathrm{sgn}} = \sum_{j=1}^d p_j \frac{\epsilon}{|g_j| + \epsilon} \ge \sum_{j=1}^d p_j \frac{\epsilon}{\|g\|_\infty + \epsilon} = \frac{\epsilon}{\|g\|_\infty + \epsilon}.$$

From $\|g\|_\infty \le \|g\|_1 \le R_1 G$, we further obtain $\bar{\mathfrak{r}}_t^{\mathrm{sgn}} \ge \frac{\epsilon}{R_1 G + \epsilon} = \frac{1}{1+R_1\lambda}$. Next, from $0 \le \frac{\epsilon}{|g_j|+\epsilon} \le 1$, it holds that $\bar{\mathfrak{r}}_t^{\mathrm{sgn}} \le 1$. Moreover, we have $\bar{\mathfrak{r}}_t^{\mathrm{sgn}} = \frac{\epsilon}{\|g\|_1}\sum_{j=1}^d \frac{|g_j|}{|g_j|+\epsilon} \le \frac{d\epsilon}{\|g\|_1} \le \frac{d\epsilon}{\gamma_\infty G} = \frac{d}{\gamma_\infty\lambda}$ where the last inequality is by $\|g\|_1 \ge \gamma_\infty G$. Putting things together, we have

$$\frac{1}{1+R_1\lambda} \le \bar{\mathfrak{r}}_t^{\mathrm{sgn}} \le \min\{1, \frac{d}{\gamma_\infty\lambda}\}.$$

From (54), we have $\bar{\mathfrak{r}}_t^{\mathrm{sgn}} - C_0\eta_t \le \mathfrak{r}_t^{\mathrm{sgn}} \le \bar{\mathfrak{r}}_t^{\mathrm{sgn}} + C_0\eta_t$. This leads to

$$\frac{1}{1+R_1\lambda} - C_0\eta_t \le \mathfrak{r}_t^{\mathrm{sgn}} \le \min\{1, \frac{d}{\gamma_\infty \lambda}\} + C_0\eta_t. \tag{55}$$

Given $\mathfrak{r}_t^{\mathrm{sgn}} \ge 0$, the lower bound becomes $\mathfrak{r}_t^{\mathrm{sgn}} \ge \left[\frac{1}{1+R_1\lambda} - C_0\eta_t\right]_+$ where $[x]_+ := \max\{x, 0\}$.

Next, we focus on bounding $\bar{\mathfrak{r}}_t^{\mathrm{grad}}$ (and consequently $\mathfrak{r}_t^{\mathrm{grad}}$ by (54)). From the coordinatewise equality $\epsilon\sigma_\epsilon(g)_j - g_j = -g_j \frac{|g_j|}{|g_j|+\epsilon}$, we have $\epsilon\sigma_\epsilon(g) - g = -g \odot \frac{|g|}{|g|+\epsilon}$. Therefore, the gradient residual $\bar{\mathfrak{r}}_t^{\mathrm{grad}}$ is $\bar{\mathfrak{r}}_t^{\mathrm{grad}} = \frac{\|g \odot \frac{|g|}{|g|+\epsilon}\|_2}{\|g\|_2}$. Given the scalar function $x \to \frac{x}{x+\epsilon}$ is increasing for $x \ge 0$, it holds that $\frac{|g_j|}{|g_j|+\epsilon} \le \frac{\|g\|_\infty}{\|g\|_\infty+\epsilon}$ and consequently $\|g \odot \frac{|g|}{|g|+\epsilon}\|_2 \le \frac{\|g\|_\infty}{\|g\|_\infty+\epsilon}\|g\|_2$. From this, we conclude that

$$\bar{\mathfrak{r}}_t^{\mathrm{grad}} = \frac{\|g \odot \frac{|g|}{|g|+\epsilon}\|_2}{\|g\|_2} \le \frac{\frac{\|g\|_\infty}{\|g\|_\infty+\epsilon}\|g\|_2}{\|g\|_2} = \frac{\|g\|_\infty}{\|g\|_\infty + \epsilon}.$$

Using $\|g\|_\infty \le \|g\|_1 \le R_1 G$ and applying the monotonicity of $\frac{x}{x+\epsilon}$, we obtain

$$\bar{\mathfrak{r}}_t^{\mathrm{grad}} \le \frac{R_1 G}{R_1 G + \epsilon} = \frac{R_1\lambda}{1+R_1\lambda}.$$

For the lower bound, using the reverse triangle inequality $\|\epsilon\sigma_\epsilon(g) - g\|_2 \ge \|g\|_2 - \epsilon\|\sigma_\epsilon(g)\|_2$ and dividing by $\|g\|_2$, we obtain $\bar{\mathfrak{r}}_t^{\mathrm{grad}} \ge 1 - \frac{\epsilon\|\sigma_\epsilon(g)\|_2}{\|g\|_2}$. From $|\sigma_\epsilon(g)_j| = \frac{|g_j|}{|g_j|+\epsilon} \le 1$, it holds $\|\sigma_\epsilon(g)\|_2 \le \sqrt{d}$. Moreover, we also have $\|g\|_2 \ge \gamma_2 G$. Putting them together, we conclude

$$\bar{\mathfrak{r}}_t^{\mathrm{grad}} \ge 1 - \frac{\epsilon\|\sigma_\epsilon(g)\|_2}{\|g\|_2} \ge 1 - \frac{\epsilon\sqrt{d}}{\gamma_2 G} = 1 - \frac{\sqrt{d}}{\gamma_2\lambda}.$$

From (54), we have $\bar{\mathfrak{r}}_t^{\mathrm{grad}} - C_0\eta_t \le \mathfrak{r}_t^{\mathrm{grad}} \le \bar{\mathfrak{r}}_t^{\mathrm{grad}} + C_0\eta_t$. This leads to

$$1 - \frac{\sqrt{d}}{\gamma_2\lambda} - C_0\eta_t \le \mathfrak{r}_t^{\mathrm{grad}} \le \frac{R_1 G}{R_1 G + \epsilon} + C_0\eta_t = \frac{R_1\lambda}{1+R_1\lambda} + C_0\eta_t. \tag{56}$$

Given $\mathfrak{r}_t^{\mathrm{grad}}$ is non-negative, the lower bound becomes $\mathfrak{r}_t^{\mathrm{grad}} \ge \left[1 - \frac{\sqrt{d}}{\gamma_2\lambda} - C_0\eta_t\right]_+$.

Let $t(\epsilon)$ be any family such that $t(\epsilon) \to \infty$ and $\lambda_{t(\epsilon)} \to \infty$ as $\epsilon \downarrow 0$. Since the learning rate is $\eta_t = \eta_0(t+1)^{-a}$ with $a > 0$, we have $\eta_{t(\epsilon)} \to 0$. By (55), we have $0 \le \mathfrak{r}_{t(\epsilon)}^{\mathrm{sgn}} \le \min\{1, \frac{d}{\gamma_\infty\lambda_{t(\epsilon)}}\} + C_0\eta_{t(\epsilon)}$. Hence, we conclude that $\mathfrak{r}_{t(\epsilon)}^{\mathrm{sgn}} \to 0$. For $\mathfrak{r}_{t(\epsilon)}^{\mathrm{grad}}$, given $1 - \frac{\sqrt{d}}{\gamma_2\lambda_{t(\epsilon)}} - C_0\eta_{t(\epsilon)} \to 1$ and $\frac{R_1\lambda_{t(\epsilon)}}{1+R_1\lambda_{t(\epsilon)}} + C_0\eta_{t(\epsilon)} \to 1$, we conclude that $\mathfrak{r}_{t(\epsilon)}^{\mathrm{grad}} \to 1$ by (56).

Let $t(\epsilon)$ be any family such that $t(\epsilon) \to \infty$ and $\lambda_{t(\epsilon)} \to 0$ as $\epsilon \downarrow 0$. We again have $\eta_{t(\epsilon)} \to 0$. From (56), we have that $0 \le \mathfrak{r}_{t(\epsilon)}^{\mathrm{grad}} \le \frac{R_1\lambda_{t(\epsilon)}}{1+R_1\lambda_{t(\epsilon)}} + C_0\eta_{t(\epsilon)}$. Consequently, it holds that $\mathfrak{r}_{t(\epsilon)}^{\mathrm{grad}} \to 0$. For $\mathfrak{r}_{t(\epsilon)}^{\mathrm{sgn}}$, given $\frac{1}{1+R_1\lambda_{t(\epsilon)}} - C_0\eta_{t(\epsilon)} \to 1$ and $\frac{d}{\gamma_\infty\lambda_{t(\epsilon)}} \to \infty$, we conclude from (55) that $\mathfrak{r}_{t(\epsilon)}^{\mathrm{sgn}} \to 1$. □

Proposition C.2. *For every $b, \delta \in (0,1)$, there exists a constant $\epsilon_0(b,\delta)$ such that*

$$T_1^-(\epsilon; b) < T_1^\delta(\epsilon) < T_1^+(\epsilon; b),$$

for every $0 < \epsilon \le \epsilon_0(b,\delta)$. Moreover, it holds that $T_1^-(\epsilon;b) \to \infty$, $\lambda_{T_1^-(\epsilon;b)} \to \infty$, $T_1^+(\epsilon;b) \to \infty$, and $\lambda_{T_1^+(\epsilon;b)} \to 0$ as $\epsilon \downarrow 0$. Consequently, Proposition C.1 gives

$$\left(\mathfrak{r}^{\mathrm{sgn}}_{T_1^-(\epsilon;b)}, \mathfrak{r}^{\mathrm{grad}}_{T_1^-(\epsilon;b)}\right) \to (0,1), \quad \textit{and} \quad \left(\mathfrak{r}^{\mathrm{sgn}}_{T_1^+(\epsilon;b)}, \mathfrak{r}^{\mathrm{grad}}_{T_1^+(\epsilon;b)}\right) \to (1,0) \qquad \textit{as } \epsilon \downarrow 0.$$

Moreover, it holds for $T_1^\delta(\epsilon)$ that $\limsup_{\epsilon\downarrow 0} \max\left\{\left|r^{\mathrm{sgn}}_{T_1^\delta(\epsilon)} - 1\right|, r^{\mathrm{grad}}_{T_1^\delta(\epsilon)}\right\} \le \frac{R_1\delta}{1+R_1\delta}$.

Proof of Proposition C.2. Fix any $b, \delta \in (0,1)$. For simplicity, we let $T_\epsilon^- := T_1^-(\epsilon;b)$, $T_\epsilon^\delta := T_1^\delta(\epsilon)$, and $T_\epsilon^+ := T_1^+(\epsilon;b)$. Recall that $\ell_\epsilon = \log\frac{1}{\epsilon}$. Define $h_\epsilon^- := \epsilon\ell_\epsilon^b$, $h_\epsilon^\delta := \delta\epsilon$, and $h_\epsilon^+ := \epsilon\ell_\epsilon^{-b}$.

As $\epsilon \downarrow 0$, it holds that $h_\epsilon^- \to 0$, $h_\epsilon^\delta \to 0$, and $h_\epsilon^+ \to 0$. For each fixed ϵ, given $G_t \to 0$, we have $T_\epsilon^-, T_\epsilon^\delta, T_\epsilon^+ < \infty$. As $\ell_\epsilon^b \to \infty$ and $\ell_\epsilon^{-b} \to 0$, we have $\ell_\epsilon^b > \delta > \ell_\epsilon^{-b}$ (or equivalently, $h_\epsilon^- > h_\epsilon^\delta > h_\epsilon^+$) for sufficiently small ϵ. Concretely, we can let $\bar{\epsilon}_0(b,\delta) := \exp(-2\delta^{-1/b})$ and choose any ϵ such that $0 < \epsilon \le \bar{\epsilon}_0(b,\delta)$. This implies that

$$\{t \ge T_* : G_t \le h_\epsilon^+\} \subseteq \{t \ge T_* : G_t \le h_\epsilon^\delta\} \subseteq \{t \ge T_* : G_t \le h_\epsilon^-\}.$$

Consequently, we have $T_\epsilon^- \le T_\epsilon^\delta \le T_\epsilon^+$.

Let $S_t := \sum_{s=0}^{t-1} \eta_s$ and $K := R_1 C_D$. By Lemma 12, we have $G_t \ge G_0 e^{-KS_t}$. Consequently, it holds that $G_0 e^{-KS_{T_\epsilon^-}} \le G_{T_\epsilon^-} \le \epsilon \ell_\epsilon^b$, where the second inequality is by the definition of T_ϵ^-. From $G_0 e^{-KS_{T_\epsilon^-}} \le \epsilon \ell_\epsilon^b$, we have $S_{T_\epsilon^-} \ge \frac{1}{K} \log \frac{G_0}{\epsilon \ell_\epsilon^b}$ by taking logarithms and rearranging. Since $\log \frac{G_0}{\epsilon \ell_\epsilon^b} = \ell_\epsilon - b \log \ell_\epsilon + \log G_0 \to \infty$ as $\epsilon \downarrow 0$, we conclude that $S_{T_\epsilon^-} \to \infty$ and consequently $T_\epsilon^- \to \infty$.

Since $T_\epsilon^- \to \infty$, we have $T_\epsilon^- > T_*$ for sufficiently small ϵ. Concretely, we set $\epsilon_{\text{burn}}(b,\delta) := \min\{e^{-2}, \frac{g_*^2}{4}\}$ (recall that g_* is the constant in Proposition 2.1) and choose $0 < \epsilon \le \epsilon_{\text{burn}}$. For $\epsilon \le e^{-2}$, we have $\ell_\epsilon \ge 2$. Since $0 < b < 1$, we have $\ell_\epsilon^b \le \ell_\epsilon$. For $\ell_\epsilon \ge 2$, we also have $\ell_\epsilon \le e^{\ell_\epsilon/2}$. Consequently, it holds that

$$\epsilon \ell_\epsilon^b = e^{-\ell_\epsilon} \ell_\epsilon^b \le e^{-\ell_\epsilon} e^{\ell_\epsilon/2} = e^{-l_\epsilon/2} = \sqrt{\epsilon} \le \frac{g_*}{2} < g_*.$$

Therefore, we must have $T_\epsilon^- > T_*$ given $G_{T_*} \ge g_* > \epsilon \ell_\epsilon^b$.

By minimality of T_ϵ^-, we have $G_{T_\epsilon^- - 1} > \epsilon \ell_\epsilon^b$ and $G_{T_\epsilon^-} \le \epsilon \ell_\epsilon^b$. These inequalities and Lemma 12 lead to

$$\epsilon \ell_\epsilon^b \ge G_{T_\epsilon^-} \ge e^{-K\eta_{T_\epsilon^- - 1}} G_{T_\epsilon^- - 1} > e^{-K\eta_{T_\epsilon^- - 1}} \epsilon \ell_\epsilon^b. \tag{57}$$

Therefore, we obtain after rearranging that

$$e^{-K\eta_{T_\epsilon^- - 1}} < \frac{G_{T_\epsilon^-}}{\epsilon \ell_\epsilon^b} \le 1.$$

Given $T_\epsilon^- \to \infty$, we have $\eta_{T_\epsilon^- - 1} \to 0$ and consequently $e^{-K\eta_{T_\epsilon^- - 1}} \to 1$. This gives $\frac{G_{T_\epsilon^-}}{\epsilon \ell_\epsilon^b} \to 1$. It follows that

$$\lambda_{T_\epsilon^-} = \frac{G_{T_\epsilon^-}}{\epsilon} = \ell_\epsilon^b \frac{G_{T_\epsilon^-}}{\epsilon \ell_\epsilon^b} \to \infty.$$

Set $N := \max\{T_*, \lceil (\frac{K\eta_0}{\log 2})^{1/a} \rceil\}$. Then we have $K\eta_N < \log 2$. Consequently, it holds that $e^{-K\eta_N} > e^{-\log 2} = \frac{1}{2}$. Define $c_N := G_0 e^{-KS_N} > 0$. Lemma 12 gives $G_t \ge G_0 e^{-KS_t} \ge c_N$ when $0 \le t \le N$. Set $x_0(b,\delta) := \max\{2, (2\delta)^{1/b}, 2\log \frac{2}{c_N}\}$ and choose $0 < \epsilon \le e^{-x_0}$. Letting $x = \log(\frac{1}{\epsilon})$, we have $x \ge x_0 \ge 2$ and $x^b \le e^{x/2}$ since $b \in (0,1)$. Consequently, we obtain

$$\epsilon \ell_\epsilon^b = e^{-x} x^b \le e^{-x/2} \le e^{-x_0/2} \le \frac{c_N}{2} < c_N.$$

Combining this with $G_t \ge c_N, 0 \le t \le N$, we conclude that $T_\epsilon^- > N$ (as $G_t > \epsilon \ell_\epsilon^b, 0 \le t \le N$). Consequently, it holds that $\eta_{T_\epsilon^- - 1} \le \eta_N$. Combining this with (57), we obtain

$$\frac{G_{T_\epsilon^-}}{\epsilon} > e^{-K\eta_{T_\epsilon^- - 1}} \ell_\epsilon^b \ge e^{-K\eta_N} \ell_\epsilon^b > \frac{1}{2} l_\epsilon^b = \frac{1}{2} x^b \ge \delta, \tag{58}$$

where in the last inequality we used $x^b \ge x_0^b \ge 2\delta$. We have already shown $T_\epsilon^- \le T_\epsilon^\delta$. Therefore, $T_\epsilon^- < T_\epsilon^\delta$; otherwise, we would have $G_{T_\epsilon^-} = G_{T_\epsilon^\delta} \le \delta\epsilon$, contradicting $G_{T_\epsilon^-} > \delta\epsilon$. At T_ϵ^δ, it holds that $G_{T_\epsilon^\delta} \le \delta\epsilon$. By Lemma 12, we have $G_0 e^{-KS_{T_\epsilon^\delta}} \le \delta\epsilon$ and therefore

$$S_{T_\epsilon^\delta} \ge \frac{1}{K} \log \frac{G_0}{\delta\epsilon} = \frac{1}{K}(\ell_\epsilon + \log \frac{G_0}{\delta}) \to \infty, \quad \text{as } \epsilon \downarrow 0.$$

Consequently, it holds that $T_\epsilon^\delta \to \infty$. This implies that $T_\epsilon^\delta > T_*$ when ϵ is sufficiently small. Concretely, we set $\epsilon_{\text{base}}(b,\delta) := \min\{1, \frac{g_*}{2\delta}\}$ and choose $0 < \epsilon \le \epsilon_{\text{base}}$. Then, we have that $\delta\epsilon \le \frac{g_*}{2} < g_* \le G_{T_*}$ and consequently $T_\epsilon^\delta > T_*$. By minimality of T_ϵ^δ and Lemma 12, we have $G_{T_\epsilon^\delta - 1} > \delta\epsilon$ and $G_{T_\epsilon^\delta} \ge G_{T_\epsilon^\delta - 1} e^{-K\eta_{T_\epsilon^\delta - 1}} > \delta\epsilon e^{-K\eta_{T_\epsilon^\delta - 1}}$. Together with the defining upper bound, we have $\delta\epsilon e^{-K\eta_{T_\epsilon^\delta - 1}} < G_{T_\epsilon^\delta} \le \delta\epsilon$. After dividing by $\delta\epsilon$, we obtain

$$e^{-K\eta_{T_\epsilon^\delta - 1}} < \frac{G_{T_\epsilon^\delta}}{\delta\epsilon} \le 1. \tag{59}$$

Consequently, we conclude that $\frac{G_{T_\epsilon^\delta}}{\delta\epsilon} \to 1$ given $\eta_{T_\epsilon^\delta - 1} \to 0$ as $T_\epsilon^\delta \to \infty$. In particular, it holds that $G_{T_\epsilon^\delta} > \frac{\delta}{2}\epsilon$ for sufficiently small ϵ. Concretely, given $\delta\epsilon e^{-K\eta_{T_\epsilon^\delta - 1}} < G_{T_\epsilon^\delta}$, to obtain $G_{T_\epsilon^\delta} > \frac{\delta}{2}\epsilon$, it is sufficient to guarantee that $e^{-K\eta_{T_\epsilon^\delta - 1}} > \frac{1}{2}$ (or equivalently, $K\eta_{T_\epsilon^\delta - 1} < \log 2$). For the choice of N above, it holds that $e^{-K\eta_N} > \frac{1}{2}$. Recall that we have shown $G_t \ge c_N, 0 \le t \le N$ where $c_N := G_0 e^{-KS_N} > 0$. Set $\bar\epsilon_3(b,\delta) := \min\{1, \frac{c_N}{2\delta}\}$ and choose $0 < \epsilon \le \bar\epsilon_3(b,\delta)$. Then we have

$$\delta\epsilon \le \frac{c_N}{2} < c_N \le G_t, \quad \forall\, 0 \le t \le N.$$

Hence, the condition $G_t \le \delta\epsilon$ cannot occur by time N, and therefore $T_\epsilon^\delta > N$. Since $T_\epsilon^\delta > N$, it also holds that $T_\epsilon^\delta - 1 \ge N$ and consequently $\eta_{T_\epsilon^\delta - 1} \le \eta_N$. From this, we conclude that

$$G_{T_\epsilon^\delta} > \delta\epsilon e^{-K\eta_{T_\epsilon^\delta - 1}} \ge \delta\epsilon e^{-K\eta_N} > \frac{\delta}{2}\epsilon.$$

Moreover, as $\ell_\epsilon^{-b} \to 0$, we have $\ell_\epsilon^{-b} < \frac{\delta}{2}$ for sufficiently small ϵ. Concretely, we set $\epsilon_0^+(b,\delta) = \exp[-2(\frac{2}{\delta})^{1/b}]$ and choose $0 < \epsilon \le \epsilon_0^+(b,\delta)$. Combining them together, we have $G_{T_\epsilon^\delta} > \frac{\delta}{2}\epsilon > \epsilon\ell_\epsilon^{-b}$. Therefore, the threshold for T_ϵ^+ has not yet been reached at T_ϵ^δ. Since $T_\epsilon^\delta \le T_\epsilon^+$, we conclude that $T_\epsilon^\delta < T_\epsilon^+$. Otherwise, we would have $G_{T_\epsilon^\delta} = G_{T_\epsilon^+} \le \epsilon\ell_\epsilon^{-b}$, contradicting $G_{T_\epsilon^\delta} > \epsilon l_\epsilon^{-b}$. This finishes the proof of $T_\epsilon^- < T_\epsilon^\delta < T_\epsilon^+$.

We already proved $T_\epsilon^\delta \to \infty$. Since $T_\epsilon^+ > T_\epsilon^\delta$, it follows that $T_\epsilon^+ \to \infty$. By the definition of T_ϵ^+, we have $G_{T_\epsilon^+} \le \epsilon\ell_\epsilon^{-b}$. Therefore, it holds that $0 < \lambda_{T_\epsilon^+} = \frac{G_{T_\epsilon^+}}{\epsilon} \le \ell_\epsilon^{-b}$. Since $\ell_\epsilon^{-b} \to 0$, we conclude that $\lambda_{T_\epsilon^+} \to 0$. So far, we have shown that $T_1^-(\epsilon;b) \to \infty$, $\lambda_{T_1^-(\epsilon;b)} \to \infty$, $T_1^+(\epsilon;b) \to \infty$, and $\lambda_{T_1^+(\epsilon;b)} \to 0$ as $\epsilon \downarrow 0$. By Proposition C.1, we have

$$\left(\mathfrak{r}^{\text{sgn}}_{T_1^-(\epsilon;b)}, \mathfrak{r}^{\text{grad}}_{T_1^-(\epsilon;b)}\right) \to (0,1), \quad \text{and} \quad \left(\mathfrak{r}^{\text{sgn}}_{T_1^+(\epsilon;b)}, \mathfrak{r}^{\text{grad}}_{T_1^+(\epsilon;b)}\right) \to (1,0) \qquad \text{as } \epsilon \downarrow 0.$$

Since $T_\epsilon^\delta \to \infty$ as $\epsilon \downarrow 0$, we have $\eta_{T_\epsilon^\delta - 1} \to 0$. Therefore, we obtain by (59) that

$$\frac{G_{T_\epsilon^\delta}}{\delta\epsilon} \to 1, \qquad \lambda_{T_\epsilon^\delta} = \frac{G_{T_\epsilon^\delta}}{\epsilon} \to \delta.$$

Consequently, applying Proposition C.1 gives

$$\frac{1}{1+R_1\delta} \le \liminf_{\epsilon\downarrow 0} r^{\text{sgn}}_{T_\epsilon^\delta} \le \limsup_{\epsilon\downarrow 0} r^{\text{sgn}}_{T_\epsilon^\delta} \le \min\left\{1, \frac{d}{\gamma_\infty\delta}\right\},$$

and

$$\left[1 - \frac{\sqrt{d}}{\gamma_2\delta}\right]_+ \le \liminf_{\epsilon\downarrow 0} r^{\text{grad}}_{T_\epsilon^\delta} \le \limsup_{\epsilon\downarrow 0} r^{\text{grad}}_{T_\epsilon^\delta} \le \frac{R_1\delta}{1+R_1\delta}.$$

Define $A_\delta := \frac{R_1\delta}{1+R_1\delta}$. Since $\frac{1}{1+R_1\delta} = 1 - A_\delta$, we have

$$1 - A_\delta - o_\epsilon(1) \le r^{\text{sgn}}_{T_\epsilon^\delta} \le 1 + o_\epsilon(1), \qquad 0 \le r^{\text{grad}}_{T_\epsilon^\delta} \le A_\delta + o_\epsilon(1).$$

Equivalently, it holds that

$$\limsup_{\epsilon\downarrow 0} \max\left\{\left|r^{\text{sgn}}_{T_\epsilon^\delta} - 1\right|, r^{\text{grad}}_{T_\epsilon^\delta}\right\} \le \frac{R_1\delta}{1+R_1\delta}.$$

Because $A_\delta = O(\delta)$, the residual pair at $T_1^\delta(\epsilon)$ lies within $O(\delta)$ of the gradient-like residual pair $(1, 0)$.

Finally, we summarize the requirements on ϵ. We set

$$\epsilon_{\mathrm{all}}(b,\delta) := \min\{\bar{\epsilon}_0(b,\delta), \epsilon_{\mathrm{burn}}(b,\delta), e^{-x_0(b,\delta)}, \epsilon_{\mathrm{base}}(b,\delta), \bar{\epsilon}_3(b,\delta), \epsilon_0^+(b,\delta)\},$$

where $\bar{\epsilon}_0(b,\delta) = \exp(-2\delta^{-1/b})$, $\epsilon_{\mathrm{burn}}(b,\delta) = \min\{e^{-2}, \frac{g_*^2}{4}\}$, $x_0(b,\delta) = \max\{2, (2\delta)^{1/b}, 2\log\frac{2}{c_N}\}$, $\epsilon_{\mathrm{base}}(b,\delta) = \min\{1, \frac{g_*}{2\delta}\}$, $\bar{\epsilon}_3(b,\delta) = \min\{1, \frac{c_N}{2\delta}\}$ ($c_N = G_0 e^{-R_1 C_D S_N}$), and $\epsilon_0^+(b,\delta) = \exp[-2(\frac{2}{\delta})^{1/b}]$. □

Proof of Theorem 2. For brevity, we write $T^- := T_1^-(\epsilon; b)$, $T^\delta := T_1^\delta(\epsilon)$, and $T^+ := T_1^+(\epsilon; b)$. Recall that $C_D := \sqrt{\frac{1-\beta_1}{1-\beta_2}}$, $K_0 := R_1 C_D$, and $C_{\mathrm{loc}} := C_{\mathrm{ema}} \max\left\{\frac{1}{\gamma_\infty}, \frac{1}{\gamma_2}\right\}$. Let $\epsilon_{\mathrm{all}}(b,\delta)$ be the threshold defined in the proof of Proposition C.2. Define $L_{\mathrm{clk}} := \max\left\{16,\ 4\left[\log\frac{1}{G_0}\right]_+\right\}$ and set $\tilde{\epsilon}_{\mathrm{res}}(b,\delta) := \min\left\{\epsilon_{\mathrm{all}}(b,\delta), e^{-L_{\mathrm{clk}}}\right\}$. We prove the result holds for every $0 < \epsilon \le \tilde{\epsilon}_{\mathrm{res}}(b,\delta)$.

By the construction in the proof of Proposition 2.1, we have $T_* > t_{\mathrm{ema}}$. Consequently, we also have $T^-, T^\delta, T^+ \ge T_* > t_{\mathrm{ema}}$. Therefore, Proposition C.1 applies at all three stopping times. In particular, for every $t \ge t_{\mathrm{ema}}$, we have

$$r_t^{\mathrm{sgn}} \le \min\left\{1, \frac{d}{\gamma_\infty \lambda_t}\right\} + C_{\mathrm{loc}}\eta_t, \qquad r_t^{\mathrm{grad}} \le \frac{R_1\lambda_t}{1 + R_1\lambda_t} + C_{\mathrm{loc}}\eta_t. \tag{60}$$

We first control λ_t at the three stopping times. From (58), we have

$$\lambda_{T^-} = \frac{G_{T^-}}{\epsilon} > \frac{1}{2}\ell_\epsilon^b. \tag{61}$$

At the other two stopping times, their definitions directly give

$$\lambda_{T^\delta} = \frac{G_{T^\delta}}{\epsilon} \le \delta, \qquad \lambda_{T^+} = \frac{G_{T^+}}{\epsilon} \le \ell_\epsilon^{-b}. \tag{62}$$

Substituting (61) into (60), we obtain

$$r_{T^-}^{\mathrm{sgn}} \le \frac{2d}{\gamma_\infty}\ell_\epsilon^{-b} + C_{\mathrm{loc}}\eta_{T^-}. \tag{63}$$

Similarly, using (62) in (60), together with the monotonicity of $x \mapsto R_1 x/(1 + R_1 x)$ over $x \ge 0$, we obtain

$$r_{T^+}^{\mathrm{grad}} \le \frac{R_1\ell_\epsilon^{-b}}{1 + R_1\ell_\epsilon^{-b}} + C_{\mathrm{loc}}\eta_{T^+} \le R_1\ell_\epsilon^{-b} + C_{\mathrm{loc}}\eta_{T^+}, \tag{64}$$

$$r_{T^\delta}^{\mathrm{grad}} \le \frac{R_1\delta}{1 + R_1\delta} + C_{\mathrm{loc}}\eta_{T^\delta}. \tag{65}$$

Recall that $S_t := \sum_{s=0}^{t-1}\eta_s$. The definition of T^- together with Lemma 12 imply $G_0 e^{-K_0 S_{T^-}} \le G_{T^-} \le \epsilon\ell_\epsilon^b$. Taking logarithms and rearranging gives

$$S_{T^-} \ge \frac{1}{K_0}\log\frac{G_0}{\epsilon\ell_\epsilon^b} = \frac{1}{K_0}\left(\ell_\epsilon - b\log\ell_\epsilon + \log G_0\right). \tag{66}$$

Because $\epsilon \le e^{-L_{\mathrm{clk}}}$, we have $\ell_\epsilon \ge L_{\mathrm{clk}} \ge 16$. Therefore, it holds that $b\log\ell_\epsilon \le \log\ell_\epsilon \le \frac{1}{4}\ell_\epsilon$. Moreover, by the definition of L_{clk}, we have $\log G_0 \ge -\frac{1}{4}\ell_\epsilon$. Consequently, we obtain $\ell_\epsilon - b\log\ell_\epsilon + \log G_0 \ge \frac{1}{2}\ell_\epsilon$. Substituting this into (66) gives $S_{T^-} \ge \frac{\ell_\epsilon}{2K_0}$.

Suppose first that $a \in (1/3, 1)$. Since the function x^{-a} is positive and decreasing, it holds that $\sum_{k=2}^t k^{-a} \le \int_1^x x^{-a}dx = \frac{t^{1-a}-1}{1-a}$. This implies that $\sum_{k=1}^t k^{-a} \le 1 + \frac{t^{1-a}-1}{1-a} = \frac{t^{1-a}-a}{1-a} \le \frac{t^{1-a}}{1-a}$.

Consequently, for every integer $t \geq 1$, we have $S_t = \eta_0 \sum_{k=1}^{t} k^{-a} \leq \frac{\eta_0}{1-a} t^{1-a}$. Combining this with $S_{T^-} \geq \frac{\ell_\epsilon}{2K_0}$ gives $T^- \geq \left(\frac{1-a}{2K_0\eta_0}\ell_\epsilon\right)^{\frac{1}{1-a}}$. Consequently, we obtain

$$\eta_{T^-} = \eta_0 (T^- + 1)^{-a} \leq \eta_0 \left(\frac{2K_0\eta_0}{1-a}\right)^{\frac{a}{1-a}} \ell_\epsilon^{-\frac{a}{1-a}} = O\left(\ell_\epsilon^{-\frac{a}{1-a}}\right). \tag{67}$$

Since the step sizes are non-increasing and $T^- < T^\delta < T^+$, we have

$$\max\{\eta_{T^-}, \eta_{T^\delta}, \eta_{T^+}\} \leq \eta_0 \left(\frac{2K_0\eta_0}{1-a}\right)^{\frac{a}{1-a}} \ell_\epsilon^{-\frac{a}{1-a}} = O\left(\ell_\epsilon^{-\frac{a}{1-a}}\right).$$

Combining this estimate with (63) and (64) gives $\max\left\{r_{T^-}^{\mathrm{sgn}}, r_{T^+}^{\mathrm{grad}}\right\} = O\left(\ell_\epsilon^{-b} + \ell_\epsilon^{-\frac{a}{1-a}}\right)$. Similarly, (65) gives $r_{T^\delta}^{\mathrm{grad}} \leq \frac{R_1\delta}{1+R_1\delta} + O\left(\ell_\epsilon^{-\frac{a}{1-a}}\right)$.

In the case of $a = 1$, we have $S_t = \eta_0 \sum_{k=1}^{t} \frac{1}{k} \leq \eta_0(1 + \log t)$. Together with $S_{T^-} \geq \frac{\ell_\epsilon}{2K_0}$, we obtain $\eta_0(1 + \log T^-) \geq \frac{\ell_\epsilon}{2K_0}$, and hence

$$T^- \geq \exp\left(\frac{\ell_\epsilon}{2K_0\eta_0} - 1\right) = e^{-1}\epsilon^{-\frac{1}{2K_0\eta_0}}.$$

It follows that

$$\eta_{T^-} = \frac{\eta_0}{T^- + 1} \leq e\eta_0 \epsilon^{\frac{1}{2K_0\eta_0}} = O\left(\epsilon^{\frac{1}{2K_0\eta_0}}\right).$$

Once again, the ordering of the stopping times and the monotonicity of η_t imply

$$\max\{\eta_{T^-}, \eta_{T^\delta}, \eta_{T^+}\} \leq e\eta_0 \epsilon^{\frac{1}{2K_0\eta_0}} = O\left(\epsilon^{\frac{1}{2K_0\eta_0}}\right). \tag{68}$$

Therefore, (65) yields $r_{T^\delta}^{\mathrm{grad}} \leq \frac{R_1\delta}{1+R_1\delta} + O\left(\epsilon^{\frac{1}{2K_0\eta_0}}\right)$.

Assume that $q > 0$ and $b > 0$. Define $f(x) := e^{-qx}x^b,\ x > 0$ and $\phi(x) := \log f(x) = b\log x - qx$. Then, we have

$$\phi'(x) = \frac{b}{x} - q, \qquad \phi''(x) = -\frac{b}{x^2} < 0.$$

Thus, ϕ is strictly concave, and its unique critical point $x = \frac{b}{q}$ is the global maximizer. Given the logarithm preserves the maximizing point, we have

$$\sup_{x>0} e^{-qx}x^b = e^{-q(b/q)}\left(\frac{b}{q}\right)^b = e^{-b}\left(\frac{b}{q}\right)^b = \left(\frac{b}{eq}\right)^b.$$

In particular, taking $x = \ell_\epsilon$ gives $\epsilon^q \ell_\epsilon^b = e^{-q\ell_\epsilon}\ell_\epsilon^b \leq \left(\frac{b}{eq}\right)^b$, or equivalently, $\epsilon^q \leq \left(\frac{b}{eq}\right)^b \ell_\epsilon^{-b}$. Using this in (68) with $q = \frac{1}{2K_0\eta_0}$, we obtain

$$\max\{\eta_{T^-}, \eta_{T^+}\} \leq e\eta_0 (\frac{2bK_0\eta_0}{e})^b \ell_\epsilon^{-b} = O(\ell_\epsilon^{-b}).$$

Substituting this into (63) and (64) finishes the proof. □

D Proof of Lemma 1

Lemma 1 is essential for the proofs of Theorem 4 and Theorem 29. Before presenting its full proof, we first provide a proof sketch to illustrate the core steps.

Proof Sketch The main difficulty is to retain the sharp coefficient 1 in front of $S_{r,T} := \sum_{t=r}^{T-1} \eta_t$; a uniform pointwise bound on D_t (see Lemma 11) would only yield a constant multiple of $S_{r,T}$. The case $\beta_2 = 0$ is immediate. In this case, we have $\beta_1 = 0$ and the magnitude of every coordinate of D_t is at most one. We therefore assume $\beta_2 > 0$ and work coordinatewise.

Fixing any $j \in [d]$, we write

$$x_t := g_t[j], \qquad \mu_t := \bar{m}_t[j], \qquad \nu_t := \bar{v}_t[j], \qquad d_t := D_t[j].$$

Let $c_t := 1 - \beta_1^{t+1}$ and $s_t := 1 - \beta_2^{t+1}$. We separate the bias correction by writing

$$d_t = \chi_t R_t, \qquad \chi_t := \frac{\sqrt{s_t}}{c_t}, \qquad R_t := \frac{\mu_t}{\sqrt{\nu_t} + \epsilon\sqrt{s_t}}.$$

By decomposing d_t as $d_t = R_t + (\chi_t - 1)R_t$, it holds that

$$|\sum_{t=r}^{T-1} \eta_t d_t| \le |\sum_{t=r}^{T-1} \eta_t R_t| + \sum_{t=r}^{T-1} |\chi_{t-1}||R_t|$$

A weighted Cauchy–Schwarz inequality applied to the EMA recursions of μ_t and ν_t gives $|\mu_t|^2 \le K_\beta^2 \nu_t$,. This implies that $|R_t| \le K_\beta$, from which we show the second term is bounded by $O(\eta_r)$. The difficulty is in controlling the first term. We split the first-moment at time r:

$$\mu_t = o_t + n_t, \qquad o_t := \beta_1^{t-r+1}\mu_{r-1}, \qquad n_t := (1-\beta_1)\sum_{s=r}^{t} \beta_1^{t-s} x_s.$$

Consequently, the first term is bounded as

$$|\sum_{t=r}^{T-1} \eta_t R_t| \le \underbrace{|\sum_{t=r}^{T-1} \eta_t \frac{n_t}{\sqrt{\nu_t} + \epsilon\sqrt{s_t}}|}_{\text{post-r contribution}} + \underbrace{\sum_{t=r}^{T-1} \eta_t \frac{|o_t|}{\sqrt{\nu_t} + \epsilon\sqrt{s_t}}}_{\text{pre-r contribution}}.$$

The pre-r contribution is controlled by a geometric series: its numerator scales as β_1^{t-r+1}, whereas the corresponding part of $\sqrt{\nu_t}$ scales as $\beta_2^{(t-r+1)/2}$. Since $\beta_1/\sqrt{\beta_2} < 1$, its cumulative contribution is $O(\eta_r)$. It remains to bound the post-r contribution. We let $N := T - r$ and define

$$\alpha_k = \eta_{r+k}, \qquad M_k := n_{r+k} = (1-\beta_1)\sum_{\ell=0}^{k} \beta_1^\ell x_{r+k-\ell}, \qquad \Delta_k := \sqrt{\nu_{r+k}} + \epsilon\sqrt{s_{r+k}}.$$

Then, it holds that $\sum_{k=0}^{N-1} \alpha_k \frac{M_k}{\Delta_k} = \sum_{t=r}^{T-1} \eta_t \frac{n_t}{\sqrt{\nu_t}+\epsilon\sqrt{s_t}}$. By Cauchy-Schwarz, we have $|\sum_{k=0}^{N-1} \alpha_k \frac{M_k}{\triangle_k}|^2 \le S_{r,T}(\sum_{k=0}^{N-1} \alpha_k \frac{M_k^2}{\Delta_k^2})$. It remains to control the ratio $\frac{M_k^2}{\Delta_k^2}$. A natural lower bound on Δ_k^2 is $\nu_{r+k} + \epsilon^2 s_{r+k}$. However, this quantity contains all history before time r, whereas M_k contains only the first-moment history after time r. Therefore, they have mismatched initial conditions. To avoid this, we introduce an alternative lower bound Q_k whose index starts from r. We show

$$\nu_{r+k} \ge (1-\beta_2)\sum_{\ell=0}^{k} \beta_2^\ell x_{r+k-\ell}^2, \qquad s_{r+k} \ge (1-\beta_2)\sum_{l=0}^{k} \beta_2^l$$

Combining them together, we have

$$\nu_{r+k} + \epsilon^2 s_{r+k} \ge (1-\beta_2)\sum_{\ell=0}^{k} \beta_2^\ell \left(x_{r+k-\ell}^2 + \epsilon^2\right).$$

The right side of this inequality is precisely the definition of Q_k, i.e. $Q_k := (1-\beta_2)\sum_{\ell=0}^{k} \beta_2^\ell \left(x_{r+k-\ell}^2 + \epsilon^2\right)$. We can write M_k and Q_k in one-step recursion form:

$$M_k = \beta_1 M_{k-1} + (1-\beta_1)x_{r+k}, \qquad Q_k = \beta_2 Q_{k-1} + (1-\beta_2)\left(x_{r+k}^2 + \epsilon^2\right),$$

with $M_{-1} = Q_{-1} = 0$. Therefore, M_k and Q_k share the same initial conditions. This makes the recurrence and logarithmic cancellation clean. Using the lower bound Q_k (on Δ_k^2), we next control the ratio $\frac{M_k^2}{Q_k}$ (note that $\frac{M_k^2}{\Delta_k^2} \le \frac{M_k^2}{Q_k}$).

Let $h_t^2 := x_t^2 + \epsilon^2$. Using the upper bound $M_k \le (1-\beta_1)\sum_{l=0}^{k} \beta_1^l h_{r+k-l}^2$, the relationship $h_{r+k-l}^2 = \frac{Q_{k-l} - \beta_2 Q_{k-l-1}}{1-\beta_2}$ (obtained from the Q_k recursion), and the inequality $1 - z \le -\log z$ (applied with $z = \frac{Q_{k-l}}{Q_k} > 0$), we obtain

$$\frac{M_k^2}{Q_k} \le \underbrace{1}_{\text{main term}} + \underbrace{c_\beta \beta_1^k}_{\text{geometric transient}} \underbrace{+ c_\beta(1-\beta_1)\sum_{\ell=1}^{k} \beta_1^{\ell-1}\big(\log Q_k - \log Q_{k-\ell}\big)}_{\text{weighted logarithmic increments}},$$

where c_β depends only on (β_1, β_2). After multiplying by $\alpha_k = \eta_{r+k}$ and summing over k, the main term contributes exactly $S_{r,T}$, the geometric transient contributes $O(\eta_r)$, and the logarithmic increments contribute $O(\eta_r H_\epsilon)$. This is what preserves the sharp leading coefficient 1.

Proof of Lemma 1. Define $H_\epsilon := \log \frac{B_g^2 + \epsilon^2}{(1-\beta_2)\epsilon^2}$. Fix one coordinate j, and let $x_t := g_t[j]$, $\mu_t := \bar{m}_t[j]$, $\nu_t := \bar{v}_t[j]$, and $d_t := D_t[j]$. Then, we have

$$\mu_t = \beta_1 \mu_{t-1} + (1-\beta_1)x_t, \qquad \nu_t = \beta_2 \nu_{t-1} + (1-\beta_2)x_t^2, \tag{69}$$

with $\mu_{-1} = \nu_{-1} = 0$. Moreover, it also holds that $d_t = \frac{\mu_t / (1-\beta_1^{t+1})}{\sqrt{\nu_t/(1-\beta_2^{t+1})} + \epsilon}$. We prove that

$$\big|\sum_{t=r}^{T-1} \eta_t d_t\big| \le \sum_{t=r}^{T-1} \eta_t + C_{\mathrm{rad}} \eta_r (1 + H_\epsilon), \tag{70}$$

for some constant C_{rad} that works for all coordinates j. In the special case where $\beta_2 = 0$, it holds that $\beta_1 = 0$ as $0 \le \beta_1 \le \beta_2$. Thus, we have $\mu_t = x_t$, $\nu_t = x_t^2$, and $d_t = \frac{x_t}{|x_t| + \epsilon}$. Therefore, it holds that $|d_t| \le 1$ and consequently $|\sum_{t=r}^{T-1} \eta_t d_t| \le \sum_{t=r}^{T-1} \eta_t$. In this case, the result holds with $C_{\mathrm{rad}} = 0$. From now on, we assume that $\beta_2 > 0$.

Part 1 By defining the bias-correction factors: $c_t := 1 - \beta_1^{t+1}$ and $s_t := 1 - \beta_2^{t+1}$, we have $d_t = \frac{\mu_t / c_t}{\sqrt{\nu_t / s_t} + \epsilon} = \frac{\sqrt{v_t}}{c_t} \frac{\mu_t}{\sqrt{\nu_t} + \epsilon\sqrt{s_t}}$. We define

$$\chi_t := \frac{\sqrt{s_t}}{c_t}, \qquad R_t := \frac{\mu_t}{\sqrt{\nu_t} + \epsilon\sqrt{s_t}}.$$

Then, it holds that $d_t = \chi_t R_t$. The proof first bounds R_t, and then controls the difference $d_t - R_t = (\chi_t - 1)R_t$. Next, we show $|\mu_t|^2 \le K_\beta^2 := \frac{\beta_2(1-\beta_1)^2}{(1-\beta_2)(\beta_2 - \beta_1^2)}$. This constant is finite as $\beta_1^2 \le \beta_2^2 \le \beta_2$ (implied by $\beta_1 \le \beta_2 < 1$). Unrolling (69) gives $\mu_t = (1-\beta_1)\sum_{k=0}^{t} \beta_1^k x_{t-k}$ and $\nu_t = (1-\beta_2)\sum_{k=0}^{t} \beta_2^k x_{t-k}^2$. By Cauchy-Schwarz, we obtain

$$\begin{aligned} |\mu_t|^2 &= (1-\beta_1)^2 \big|\sum_{k=0}^{t} (\frac{\beta_1}{\sqrt{\beta_2}})^k \beta_2^{k/2} x_{t-k}\big|^2 \\ &\le (1-\beta_1)^2 \big(\sum_{k=0}^{t} (\frac{\beta_1^2}{\beta_2})^k\big)\big(\sum_{k=0}^{t} \beta_2^k x_{t-k}^2\big). \end{aligned}$$

From $\sum_{k=0}^{t} (\frac{\beta_1^2}{\beta_2})^k \le \frac{1}{1 - \beta_1^2/\beta_2} = \frac{\beta_2}{\beta_2 - \beta_1^2}$ (as $\frac{\beta_1^2}{\beta_2} < 1$) and $\sum_{k=0}^{t} \beta_2^k x_{t-k}^2 = \frac{\nu_t}{1-\beta_2}$, we further obtain

$$|\mu_t|^2 \le \frac{\beta_2(1-\beta_1)^2}{(1-\beta_2)(\beta_2 - \beta_1^2)} \nu_t = K_\beta^2 \nu_t, \qquad \forall t.$$

If $\nu_t = 0$, then all $x_0, ..., x_t$ are zero and $\mu_t = 0$ holds as well. In this case, the inequality trivially holds. Consequently, it holds that

$$|R_t| = \frac{|\mu_t|}{\sqrt{\nu_t} + \epsilon\sqrt{s_t}} \leq \frac{|\mu_t|}{\sqrt{\nu_t}} \leq K_\beta. \tag{71}$$

For $t \geq r$, we split the first moment into two parts

$$\mu_t = o_t + n_t, \qquad o_t := \beta_1^{t-r+1}\mu_{r-1}, \qquad n_t := (1-\beta_1)\sum_{s=r}^{t}\beta_1^{t-s}x_s.$$

For the second moment, we have $\nu_t = \beta_2^{t-r+1}\nu_{r-1} + (1-\beta_2)\sum_{s=r}^{t}\beta_2^{t-s}x_s^2 \geq \beta_2^{t-r+1}\nu_{r-1}$. Using $|\mu_{r-1}| \leq K_\beta\sqrt{\nu_{r-1}}$, we obtain

$$\frac{|o_t|}{\sqrt{\nu_t} + \epsilon\sqrt{s_t}} \leq \frac{\beta_1^{t-r+1}|\mu_{r-1}|}{\sqrt{\nu_t}} \leq \frac{\beta_1^{t-r+1}K_\beta\sqrt{\nu_{r-1}}}{\beta_2^{(t-r+1)/2}\sqrt{\nu_{r-1}}} = K_\beta(\frac{\beta_1}{\sqrt{\beta_2}})^{t-r+1}.$$

Define $\rho_\beta := \frac{\beta_1}{\sqrt{\beta_2}}$. We have $\beta_1^2 \leq \beta_2^2 < \beta_2$. Since $\eta_t \leq \eta_r, t \geq r$, we obtain

$$\sum_{t=r}^{T-1}\eta_t\frac{|o_t|}{\sqrt{\nu_t} + \epsilon\sqrt{s_t}} \leq K_\beta\eta_r\sum_{k=1}^{\infty}\rho_\beta^k = K_\beta\eta_r\frac{\rho_\beta}{1-\rho_\beta}. \tag{72}$$

Part 2 Set $N := T - r$. For $k = 0, ..., N-1$, we define

$$M_k := n_{r+k} = (1-\beta_1)\sum_{l=0}^{k}\beta_1^l x_{r+k-l}, \qquad Q_k := (1-\beta_2)\sum_{l=0}^{k}\beta_2^l(x_{r+k-l}^2 + \epsilon^2) \tag{73}$$

From $\nu_{r+k} \geq (1-\beta_2)\sum_{l=0}^{k}\beta_2^l x_{r+k-l}^2$ and $s_{r+k} = 1 - \beta_2^{r+k+1} \geq 1 - \beta_2^{k+1} = (1-\beta_2)\sum_{l=0}^{k}\beta_2^l$, we obtain using the definition of Q_k that

$$\nu_{r+k} + \epsilon^2 s_{r+k} \geq Q_k.$$

Moreover, we have

$$(\sqrt{\nu_{r+k}} + \epsilon\sqrt{s_{r+k}})^2 = \nu_{r+k} + \epsilon^2 s_{r+k} + 2\epsilon\sqrt{\nu_{r+k}s_{r+k}} \geq \nu_{r+k} + \epsilon^2 s_{r+k}.$$

Hence, we obtain $(\sqrt{\nu_{r+k}} + \epsilon\sqrt{s_{r+k}})^2 \geq Q_k$ or equivalently $\sqrt{\nu_{r+k}} + \epsilon\sqrt{s_{r+k}} \geq \sqrt{Q_k}$. We note that $Q_k \geq (1-\beta_2)\epsilon^2$ because the $l = 0$ term contains ϵ^2. Since $|x_t| \leq B_g$, it holds that $x_t^2 + \epsilon^2 \leq B_g^2 + \epsilon^2$. Moreover, we have $(1-\beta_2)\sum_{l=0}^{k}\beta_2^l = 1 - \beta_2^{k+1} \leq 1$. Therefore, we obtain

$$(1-\beta_2)\epsilon^2 \leq Q_k \leq (1-\beta_2)\sum_{l=0}^{k}\beta_2^l(B_g^2 + \epsilon^2) \leq B_g^2 + \epsilon^2. \tag{74}$$

Let $L_\epsilon := (1-\beta_2)\epsilon^2$ and $U_\epsilon := B_g^2 + \epsilon^2$. Thus, we have shown that $\log Q_i \in [\log L_\epsilon, \log U_\epsilon]$ for any $i \in \{0, ..., k\}$. Consequently, we obtain

$$\max_{u,v\in\{0,...,k\}} |\log Q_u - \log Q_v| \leq \log U_\epsilon - \log L_\epsilon = \log\frac{U_\epsilon}{L_\epsilon} =: H_\epsilon.$$

Define $h_t^2 := x_t^2 + \epsilon^2, \forall t \in \{r, ..., T-1\}$ and $w_l := (1-\beta_1)\beta_1^l$. By Cauchy-Schwarz, we have

$$\begin{aligned}
M_k^2 &= [(1-\beta_1)\sum_{l=0}^{k}\beta_1^l x_{r+k-l}]^2 = [(1-\beta_1)]^2[\sum_{l=0}^{k}\sqrt{\beta_1^l}\sqrt{\beta_1^l}x_{r+k-l}]^2 \\
&\leq [(1-\beta_1)\sum_{l=0}^{k}\beta_1^l][(1-\beta_1)\sum_{l=0}^{k}\beta_1^l x_{r+k-l}^2] \\
&\leq (1-\beta_1)\sum_{l=0}^{k}\beta_1^l x_{r+k-l}^2 \\
&\leq (1-\beta_1)\sum_{l=0}^{k}\beta_1^l h_{r+k-l}^2.
\end{aligned}$$

where in the second inequality we used $\sum_{l=0}^{k} w_l = 1 - \beta_1^{k+1} \leq 1$.

By the definition of Q_k, we have

$$Q_k = \beta_2 Q_{k-1} + (1-\beta_2) h_{r+k}^2, \qquad Q_{-1} := 0. \tag{75}$$

Thus, it holds that $h_{r+k-l}^2 = \frac{Q_{k-l} - \beta_2 Q_{k-l-1}}{1-\beta_2}, \forall l \in \{0, ..., k\}$. Consequently, we obtain

$$\begin{aligned} M_k^2 &\leq \frac{1-\beta_1}{1-\beta_2} \sum_{l=0}^{k} \beta_1^l (Q_{k-l} - \beta_2 Q_{k-l-1}) \\ &= \frac{1-\beta_1}{1-\beta_2} [Q_k - (\beta_2 - \beta_1) \sum_{l=1}^{k} \beta_1^{l-1} Q_{k-l}]. \end{aligned} \tag{76}$$

Define $c_\beta := \frac{\beta_2 - \beta_1}{1-\beta_2} \geq 0$. Then we have $\frac{1-\beta_1}{1-\beta_2} = 1 + c_\beta$. Dividing (76) by Q_k and using the definition of c_β, we obtain

$$\frac{M_k^2}{Q_k} \leq 1 + c_\beta [1 - (1-\beta_1) \sum_{l=1}^{k} \beta_1^{l-1} \frac{Q_{k-l}}{Q_k}].$$

From $(1-\beta_1) \sum_{l=1}^{k} \beta_1^{l-1} = 1 - \beta_1^k$, we have

$$1 - (1-\beta_1) \sum_{l=1}^{k} \beta_1^{l-1} \frac{Q_{k-l}}{Q_k} = \beta_1^k + (1-\beta_1) \sum_{l=1}^{k} \beta_1^{l-1} (1 - \frac{Q_{k-l}}{Q_k}).$$

For every $z > 0$, it holds that $1 - z \leq -\log z$. Applying this with $z = \frac{Q_{-l}}{Q_k}$, we have $1 - \frac{Q_{k-l}}{Q_k} \leq \log Q_k - \log Q_{k-l}$. Putting things together, we have

$$\frac{M_k^2}{Q_k} \leq 1 + c_\beta [\beta_1^k + (1-\beta_1) \sum_{l=1}^{k} \beta_1^{l-1} (\log Q_k - \log Q_{k-l})]. \tag{77}$$

Next, we verify (77) still holds in the $\beta_1 = 0$ case. In this case, we have $M_k = x_{r+k}$, $c_\beta = \frac{\beta_2}{1-\beta_2}$, and the recursion in (75) still holds. Note that all logarithms are well-defined as $Q_k \geq (1-\beta_2)\epsilon^2 > 0$ as $\epsilon > 0$. When $k = 0$, we have $M_0 = x_r$, $Q_0 = (1-\beta_2)(x_r^2 + \epsilon^2)$, and therefore $\frac{M_0^2}{Q_0} = \frac{x_r^2}{(1-\beta_2)(x_r^2+\epsilon^2)} \leq \frac{1}{1-\beta_2}$. Thus, (77) is valid at $k = 0$. For $k \geq 1$, we have $M_k^2 = x_{r+k}^2 \leq x_{r+k}^2 + \epsilon^2 = \frac{Q_k - \beta_2 Q_{k-1}}{1-\beta_2}$ where the last equality is by the recursion $Q_k - \beta_2 Q_{k-1} = (1-\beta_2)(x_{r+k}^2 + \epsilon^2)$. Dividing by Q_k gives

$$\begin{aligned} \frac{M_k^2}{Q_k} \leq \frac{Q_k - \beta_2 Q_{k-1}}{(1-\beta_2) Q_k} &= \frac{1 - \beta_2 (Q_{k-1}/Q_k)}{1-\beta_2} = 1 + \frac{\beta_2}{1-\beta_2} (1 - \frac{Q_{k-1}}{Q_k}) \\ &\leq 1 + \frac{\beta_2}{1-\beta_2} (\log Q_k - \log Q_{k-1}). \end{aligned}$$

This is the inequality (77) given $\beta_1^k = 0$ and only the $l = 1$ term survives with the convention that $0^0 = 1$.

Part 3 Define $\alpha_k := \eta_{r+k}, k \in \{0, ..., N-1\}$. After multiplying (77) by α_k and summing, we obtain

$$\sum_{k=0}^{N-1} \alpha_k \frac{M_k^2}{Q_k} \leq \sum_{k=0}^{N-1} \alpha_k + c_\beta \sum_{k=0}^{N-1} \alpha_k \beta_1^k + c_\beta \Omega, \tag{78}$$

where

$$\Omega := (1-\beta_1) \sum_{k=0}^{N-1} \alpha_k \sum_{l=1}^{k} \beta_1^{l-1} (\log Q_k - \log Q_{k-l}).$$

We now bound the term Ω. We reorganize the terms in Ω and write it as $\Omega = \sum_{u=0}^{N-1} A_u \log Q_u$. The positive appearances of $\log Q_u$ come from $k = u$ and their total coefficient is $\alpha_u(1-\beta_1)\sum_{l=1}^{u}\beta_1^{l-1} = \alpha_u(1-\beta_1^u)$. The negative appearances occur when $k-l=u$, or $k=u+l$. Thus, we have

$$A_u = \alpha_u(1-\beta_1^u) - (1-\beta_1)\sum_{l=1}^{N-1-u}\alpha_{u+l}\beta_1^{l-1}. \tag{79}$$

Next, we show two important properties of A_u. After summing (79), we have

$$\begin{aligned}\sum_{u=0}^{N-1} A_u &= \sum_{u=0}^{N-1}\alpha_u(1-\beta_1^u) - (1-\beta_1)\sum_{u=0}^{N-1}\sum_{l=1}^{N-1-u}\alpha_{u+l}\beta_1^{l-1}\\ &= \sum_{u=0}^{N-1}\alpha_u(1-\beta_1^u) - (1-\beta_1)\sum_{k=1}^{N-1}\alpha_k\sum_{l=1}^{k}\beta_1^{l-1}\\ &= \sum_{u=0}^{N-1}\alpha_u(1-\beta_1^u) - \sum_{k=1}^{N-1}\alpha_k(1-\beta_1^k) = 0,\end{aligned}$$

where in the second equality we set $k = u+l$. Then $k = 1, ..., N-1$, and for each k, we have $l = 1, ..., k$. For the last equality, we used $\alpha_0(1-\beta_1^0) = 0$. Consequently, we conclude that $\sum_{u=0}^{N-1} A_u = 0$. For the second property, as $\alpha_{u+l} \le \alpha_u$, we have

$$A_u \ge \alpha_u(1-\beta_1^u) - (1-\beta_1)\alpha_u\sum_{l=1}^{N-1-u}\beta_1^{l-1} = \alpha_u(1-\beta_1^u) - \alpha_u(1-\beta_1^{N-1-u}) = \alpha_u(\beta_1^{N-1-u} - \beta_1^u).$$

In particular, we have $A_u \ge -\alpha_u\beta_1^u$. Consequently, whenever $A_u < 0$, it holds that $-A_u \le \alpha_u\beta_1^u$. From (74), we have $\log L_\epsilon \le \log Q_u \le \log U_\epsilon$ for any $u \in \{0, ..., N-1\}$. Define $l_u := \log Q_u - \log U_\epsilon$. From the definition of H_ϵ, we have $-H_\epsilon \le l_u \le 0, \forall u \in \{0, ..., N-1\}$. Since $\sum_{u=0}^{N-1} A_u = 0$, we have $\sum_{u=0}^{N-1} A_u \log Q_u = \sum_{u=0}^{N-1} A_u l_u$. If $A_u \ge 0$, then we have $A_u l_u \le 0$ because $l_u \le 0$. If $A_u < 0$, from $-A_u \le \alpha_u\beta_1^u$ and $-l_u \le H_\epsilon$, we have $A_u l_u = (-A_u)(-l_u) \le \alpha_u\beta_1^u H_\epsilon$. Therefore, we obtain $\Omega \le H_\epsilon\sum_{u=0}^{N-1}\alpha_u\beta_1^u$. Since $\alpha_u \le \alpha_0 = \eta_r$, it holds that $\sum_{u=0}^{N-1}\alpha_u\beta_1^u \le \eta_r\sum_{u=0}^{\infty}\beta_1^u = \frac{\eta_r}{1-\beta_1}$. Consequently, we have $\Omega \le \frac{\eta_r H_\epsilon}{1-\beta_1}$. Substituting these bounds into (78), we obtain

$$\sum_{k=0}^{N-1}\alpha_k\frac{M_k^2}{Q_k} \le \sum_{k=0}^{N-1}\alpha_k + \frac{c_\beta}{1-\beta_1}\eta_r(1+H_\epsilon).$$

Define $S_{r,T} := \sum_{t=r}^{T-1}\eta_t = \sum_{k=0}^{N-1}\alpha_k$ and $C_E := \frac{c_\beta}{1-\beta_1} = \frac{\beta_2-\beta_1}{(1-\beta_2)(1-\beta_1)}$.Then we can further write

$$\sum_{k=0}^{N-1}\alpha_k\frac{M_k^2}{Q_k} \le S_{r,T} + C_E\eta_r(1+H_\epsilon). \tag{80}$$

Part 4 Define $\triangle_k := \sqrt{\nu_{r+k}} + \epsilon\sqrt{s_{r+k}}$. Previously, we have shown that $\triangle_k^2 \ge Q_k$. By Cauchy-Schwarz, we have

$$\begin{aligned}|\sum_{k=0}^{N-1}\alpha_k\frac{M_k}{\triangle_k}|^2 &\le (\sum_{k=0}^{N-1}\alpha_k)(\sum_{k=0}^{N-1}\alpha_k\frac{M_k^2}{\triangle_k^2}) = S_{r,T}(\sum_{k=0}^{N-1}\alpha_k\frac{M_k^2}{\triangle_k^2}) \le S_{r,T}(\sum_{k=0}^{N-1}\alpha_k\frac{M_k^2}{Q_k})\\ &\le S_{r,T}[S_{r,T} + C_E\eta_r(1+H_\epsilon)],\end{aligned}$$

where we used (80) in the last inequality. For any $A, B \ge 0$, we have $\sqrt{A(A+B)} \le A + \frac{B}{2}$. Applying this fact, we further obtain that

$$|\sum_{k=0}^{N-1}\alpha_k\frac{M_k}{\triangle_k}| \le S_{r,T} + \frac{C_E}{2}\eta_r(1+H_\epsilon). \tag{81}$$

Recall that $\alpha_k = \eta_{r+k}$, $M_k = n_{r+k}$, and $\triangle_k = \sqrt{\nu_{r+k}} + \epsilon\sqrt{s_{r+k}}$. Consequently, it holds that

$$\sum_{k=0}^{N-1} \alpha_k \frac{M_k}{\triangle_k} = \sum_{k=0}^{N-1} \eta_{r+k} \frac{n_{r+k}}{\sqrt{\nu_{r+k}} + \epsilon\sqrt{s_{r+k}}} = \sum_{t=r}^{T-1} \eta_t \frac{n_t}{\sqrt{\nu_t} + \epsilon\sqrt{s_t}}.$$

From this and (81), we have

$$|\sum_{t=r}^{T-1} \eta_t \frac{n_t}{\sqrt{\nu_t} + \epsilon\sqrt{s_t}}| \le S_{r,T} + \frac{C_E}{2}\eta_r(1 + H_\epsilon). \tag{82}$$

Part 5 Recall that $R_t = \frac{o_t + n_t}{\sqrt{\nu_t} + \epsilon\sqrt{s_t}}$. Combining (72) and (82), we obtain

$$\begin{aligned} |\sum_{t=r}^{T-1} \eta_t R_t| &\le |\sum_{t=r}^{T-1} \eta_t \frac{n_t}{\sqrt{\nu_t} + \epsilon\sqrt{s_t}}| + \sum_{t=r}^{T-1} \eta_t \frac{|o_t|}{\sqrt{\nu_t} + \epsilon\sqrt{s_t}} \\ &\le S_{r,T} + \frac{C_E}{2}\eta_r(1 + H_\epsilon) + C_o\eta_r, \end{aligned}$$

where $C_o := K_\beta \frac{\rho_\beta}{1-\rho_\beta}$. Since $H_\epsilon \ge 0$, it holds that $C_o\eta_r \le C_o\eta_r(1 + H_\epsilon)$. Hence, we obtain

$$|\sum_{t=r}^{T-1} \eta_t R_t| \le S_{r,T} + (\frac{C_E}{2} + C_o)\eta_r(1 + H_\epsilon).$$

Recall that $\chi_t = \frac{\sqrt{s_t}}{c_t}$, $c_t = 1 - \beta_1^{t+1}$, and $s_t = 1 - \beta_2^{t+1}$. Since $c_t \ge 1 - \beta_1$, we have

$$|\chi_t - 1| = \frac{|\sqrt{s_t} - c_t|}{c_t} \le \frac{|1 - \sqrt{s_t}| + |1 - c_t|}{1 - \beta_1}.$$

Note that $1 - c_t = \beta_1^{t+1}$. For any $z \in [0, 1]$, $1 - \sqrt{1 - z} = \frac{z}{1+\sqrt{1-z}} \le z$. Taking $z = \beta_2^{t+1}$, we have $1 - \sqrt{s_t} = 1 - \sqrt{1 - \beta_2^{t+1}} \le \beta_2^{t+1}$. Hence, we have

$$|\chi_t - 1| \le \frac{\beta_1^{t+1} + \beta_2^{t+1}}{1 - \beta_1}. \tag{83}$$

From (71) and (83), we have

$$\begin{aligned} \sum_{t=r}^{T-1} \eta_t |\chi_t - 1||R_t| &\le \frac{K_\beta}{1 - \beta_1} \sum_{t=r}^{T-1} \eta_t(\beta_1^{t+1} + \beta_2^{t+1}) \\ &\le \frac{K_\beta \eta_r}{1 - \beta_1}[\sum_{t=r}^{\infty} \beta_1^{t+1} + \sum_{t=r}^{\infty} \beta_2^{t+1}] \\ &\le \frac{K_\beta \eta_r}{1 - \beta_1}[\frac{\beta_1}{1 - \beta_1} + \frac{\beta_2}{1 - \beta_2}]. \end{aligned}$$

Define $C_{\mathrm{bc}} := \frac{K_\beta}{1-\beta_1}[\frac{\beta_1}{1-\beta_1} + \frac{\beta_2}{1-\beta_2}]$. Then we have $\sum_{t=r}^{T-1} \eta_t |\chi_t - 1||R_t| \le C_{\mathrm{bc}}\eta_r$. Since $d_t = \chi_t R_t = R_t + (\chi_t - 1)R_t$, we have

$$\begin{aligned} |\sum_{t=r}^{T-1} \eta_t d_t| &\le |\sum_{t=r}^{T-1} \eta_t R_t| + \sum_{t=r}^{T-1} \eta_t |\chi_t - 1||R_t| \\ &\le S_{r,T} + (\frac{C_E}{2} + C_o + C_{\mathrm{bc}})\eta_r(1 + H_\epsilon). \end{aligned}$$

This proves (70) with $C_{\mathrm{rad}} = \frac{C_E}{2} + C_o + C_{\mathrm{bc}}$.

Part 6 Recall that $d_t = D_t[j]$. Substituting this into (70) and using the definition of H_ϵ, we have

$$|\sum_{t=r}^{T-1} \eta_t D_t[j]| \le \sum_{t=r}^{T-1} \eta_t + C_{\mathrm{rad}}\eta_r[1 + \log \frac{B_g^2 + \epsilon^2}{(1 - \beta_2)\epsilon^2}]$$

Taking the maximum over the coordinates j completes the proof. □

E ADAM'S LOCAL UPDATE TRANSITION

Throughout this section, we assume data separability, $0 \le \beta_1 \le \beta_2 < 1$, and $\epsilon \in (0, 1]$. Recall that $R_1 := \max_{i\in[n]} \|x_i\|_1$ and $R_2 := \max_{i\in[n]} \|x_i\|_2$. The step size schedule is $\eta_t = \eta_0(t+1)^{-a}$ with $\eta_0 > 0$ and $a \in (1/3, 1]$.

E.1 PROOF OF THEOREM 4

Recall the three stopping times:

$$T_1^-(\epsilon; b) = \inf\left\{t \ge T_* : G_t \le \epsilon \ell_\epsilon^b\right\},$$
$$T_1^\delta(\epsilon) = \inf\left\{t \ge T_* : G_t \le \delta\epsilon\right\},$$
$$T_1^+(\epsilon; b) = \inf\left\{t \ge T_* : G_t \le \epsilon \ell_\epsilon^{-b}\right\},$$

where $\ell_\epsilon = \log\frac{1}{\epsilon}$ and $b \in (0, 1)$. For simplicity, we let $T^- := T_1^-(\epsilon; b)$, $T^\delta := T_1^\delta(\epsilon)$, and $T^+ := T_1^+(\epsilon; b)$.

We first show a lemma that follows directly from Proposition 2.1.

Lemma 20. *For every $\zeta > 0$, there exists $t_\zeta < \infty$ (independent of ϵ) such that for every $\epsilon \in (0, 1]$*

$$L_t \le \zeta, \qquad \forall t \ge t_\zeta. \tag{84}$$

Proof. Fix $\zeta > 0$. We use the constants c, K from (45) and the notations $t_0 := t_{\mathrm{ema}}$, $R_t := \sum_{s=t}^\infty \eta_s^3$, and $\Phi_t := L_t + KR_t$ introduced in the proof of Proposition 2.1. By (46), we have

$$0 \le \Phi_{t+1} \le \Phi_t - c\eta_t G_t^2, \qquad t \ge t_0.$$

Moreover, the bound in (48) gives $\Phi_{t_0} \le B$ where $B < \infty$ is a constant independent of ϵ. Therefore, we obtain

$$0 \le L_t \le \Phi_t \le B, \qquad t \ge t_0, \tag{85}$$

and Φ_t is non-increasing on this range. Since $R_t \downarrow 0$, we can choose an integer $r_\zeta \ge t_0$ (independent of ϵ) such that $KR_{r_\zeta} \le \frac{\zeta}{2}$. Using the estimate established in the proof of Proposition 2.1, one possible choice is

$$r_\zeta := \max\left\{t_0,\, 1,\, \left\lceil \left(\frac{2K\eta_0^3}{(3a-1)\zeta}\right)^{1/(3a-1)} \right\rceil\right\}.$$

This is the choice of t_1 in that proof with h replaced by ζ. The argument leading to (50) uses only $h > 0$, $KR_{t_1} \le h/2$, (46), and (49). Therefore, replacing h by ζ, t_1 by r_ζ, and $\bar{g}$ by $\bar{g}_\zeta := \frac{1}{n}(1 - e^{-\zeta/2}) > 0$, we obtain the following

$$t \ge r_\zeta, \quad \Phi_t > \zeta \quad \Longrightarrow \quad \Phi_{t+1} \le \Phi_t - c\bar{g}_\zeta^2 \eta_t.$$

Because $\sum_{t=0}^\infty \eta_t = \infty$, we can choose an integer $t_\zeta > r_\zeta$ such that $c\bar{g}_\zeta^2 \sum_{t=r_\zeta}^{t_\zeta - 1} \eta_t > B$. Concretely, we define

$$t_\zeta := \min\left\{T \in \mathbb{N} : T > r_\zeta, \quad c\bar{g}_\zeta^2 \sum_{t=r_\zeta}^{T-1} \eta_t > B\right\}.$$

The set is nonempty and t_ζ is independent of ϵ. For every $\epsilon \in (0, 1]$, the arguments in Proposition 2.1 give a time $\tau = \tau(\epsilon, \zeta) \in [r_\zeta, t_\zeta]$ such that $\Phi_\tau \le \zeta$ (note that τ may depend on ϵ but t_ζ does not). By the monotonicity of Φ_t following from (46), we obtain

$$L_t \le \Phi_t \le \Phi_\tau \le \zeta, \qquad t \ge t_\zeta.$$

This holds for every $\epsilon \in (0, 1]$. □

Lemma 21. *Let $h_\epsilon > 0$ satisfy $h_\epsilon \to 0$ as $\epsilon \downarrow 0$ and define $\tau_\epsilon := \inf\{t \ge T_* : G_t \le h_\epsilon\}$. The following holds:*

$$\tau_\epsilon \to \infty, \qquad \frac{G_{\tau_\epsilon}}{h_\epsilon} \to 1, \qquad \frac{L_{\tau_\epsilon}}{h_\epsilon} \to 1, \qquad \text{as } \epsilon \downarrow 0.$$

Proof. Denote $K_0 := R_1 C_D$ where $C_D = \sqrt{\frac{1-\beta_1}{1-\beta_2}}$. By Lemma 12, it holds that $G_t \geq G_0 e^{-K_0 S_t}$. Fix any deterministic integer N. Since $S_t \leq S_N$ for $t \leq N$, we have

$$G_t \geq G_0 e^{-K_0 S_N} =: c_N > 0, \qquad 0 \leq t \leq N.$$

For sufficiently small ϵ, we have $h_\epsilon < c_N$. Therefore, the event $G_t \leq h_\epsilon$ cannot occur by time N. Therefore, it holds that $\tau_\epsilon > N$. Since N was arbitrary, we conclude that $\tau_\epsilon \to \infty$.

By minimality of τ_ϵ, we have $G_{\tau_\epsilon - 1} > h_\epsilon$ and $G_{\tau_\epsilon} \leq h_\epsilon$. Applying Lemma 12 over the single step from $\tau_\epsilon - 1$ to τ_ϵ, we obtain

$$G_{\tau_\epsilon} \geq e^{-K_0 \eta_{\tau_\epsilon - 1}} G_{\tau_\epsilon - 1} > e^{-K_0 \eta_{\tau_\epsilon - 1}} h_\epsilon.$$

Therefore, it holds that $e^{-K_0 \eta_{\tau_\epsilon - 1}} < \frac{G_{\tau_\epsilon}}{h_\epsilon} \leq 1$. Because $\tau_\epsilon \to \infty$, it holds that $e^{-K_0 \eta_{\tau_\epsilon - 1}} \to 1$. Therefore, we conclude that $\frac{G_{\tau_\epsilon}}{h_\epsilon} \to 1$. For $t \geq T_*$, Proposition 2.1 gives $G_t \leq L_t \leq 2G_t$. Therefore, it holds that $L_{\tau_\epsilon} \leq 2G_{\tau_\epsilon} \leq 2h_\epsilon \to 0$. By Lemma 8, $1 \geq \frac{G_t}{L_t} \geq 1 - \frac{nL_t}{2}$. Since $L_{\tau_\epsilon} \to 0$, we have $\frac{G_{\tau_\epsilon}}{L_{\tau_\epsilon}} \to 1$ and consequently $\frac{L_{\tau_\epsilon}}{G_{\tau_\epsilon}} \to 1$. Combining this with $\frac{G_{\tau_\epsilon}}{h_\epsilon} \to 1$, we obtain $\frac{L_{\tau_\epsilon}}{h_\epsilon} = \frac{L_{\tau_\epsilon}}{G_{\tau_\epsilon}} \frac{G_{\tau_\epsilon}}{h_\epsilon} \to 1$. □

Applying Lemma 21 to the three stopping times, we obtain

$$\begin{aligned}
G_{T^-} &= \epsilon \ell_\epsilon^b (1 + o(1)), \qquad & L_{T^-} &= \epsilon \ell_\epsilon^b (1 + o(1)), \\
G_{T^\delta} &= \delta\epsilon (1 + o(1)), \qquad & L_{T^\delta} &= \delta\epsilon (1 + o(1)), \\
G_{T^+} &= \epsilon \ell_\epsilon^{-b} (1 + o(1)), \qquad & L_{T^+} &= \epsilon \ell_\epsilon^{-b} (1 + o(1)).
\end{aligned} \tag{86}$$

Moreover, all three stopping times diverge as $\epsilon \downarrow 0$, and hence the corresponding learning rates converge to zero.

Proof Sketch of Theorem 4 We first provide a proof sketch of Theorem 4. The main idea is that the loss satisfies two qualitatively different recursions:

$$L_{t+1} - L_t \approx -\gamma_\infty \eta_t L_t \quad \text{before the crossover,}$$

whereas

$$L_{t+1} - L_t \asymp -\frac{\eta_t}{\epsilon} L_t^2 \quad \text{after the crossover.}$$

Therefore, the loss decays exponentially in the S-clock before the crossover, whereas after the crossover, the reciprocal loss grows linearly in S_t/ϵ.

Part 1: Characterization of S_{T^-}**.** Fix $0 < \zeta < 2/n$. By Lemma 20, there exists an ϵ-independent time t_ζ such that $L_t \leq \zeta$ when $t \geq t_\zeta$. Choose an integer r such that $r \geq t_\zeta$, $r \geq T_*$, $C_{\mathrm{md}} \eta_r < \gamma_\infty$, and $\gamma_\infty \eta_r \leq \frac{1}{2}$. We highlight the important roles played by r in the proof:

1. To make G_t arbitrarily close to L_t. By choosing $r \geq t_\zeta$, Lemma 8 gives

   $$\frac{G_t}{L_t} \geq 1 - \frac{nL_t}{2} \geq 1 - \frac{n\zeta}{2} =: \alpha_\zeta, \qquad t \geq r.$$

   Therefore, it holds that $G_t \geq \alpha_\zeta L_t$ where $\alpha_\zeta \longrightarrow 1$ as $\zeta \downarrow 0$. This estimate is needed to recover the sharp leading coefficient $1/\gamma_\infty$. If we started directly from T_*, we would only know that $G_t \geq \frac{1}{2} L_t$. This would lose a factor of 2 in the resulting upper bound (see (93)).

2. To obtain a strict contraction. Before T^-, the proof obtains

   $$L_{t+1} \leq (1 - \alpha_\zeta q_{\epsilon,r} \eta_t) L_t, \qquad q_{\epsilon,r} := \gamma_\infty - C_{\mathrm{md}} \eta_r - d\ell_\epsilon^{-b}. \tag{87}$$

   For fixed r, we have $q_{\epsilon,r} \to \gamma_\infty - C_{\mathrm{md}} \eta_r > 0$ as $\epsilon \downarrow 0$. Therefore, it holds that $0 < q_{\epsilon,r} \leq \gamma_\infty$ when ϵ is sufficiently small. Consequently, the recursion in (87) is a strict contraction since $0 < \alpha_\zeta q_{\epsilon,r} \eta_t \leq \gamma_\infty \eta_r \leq \frac{1}{2}$ and it holds for every $t \geq r$.

3. To separate the finite burn-in from the ϵ-dependent time scale. We know $T^- \to \infty$ as $\epsilon \downarrow 0$. Therefore, it holds that $T^> r + 1$ when ϵ is sufficiently small, and the recursion in (87) holds throughout the interval $r \le t < T^-$. Moreover, because r is fixed, we have $S_r - S_{T_*} = O(1) = o(\ell_\epsilon)$. Thus, removing the finite initial segment $[T_*, r]$ does not affect the leading-order asymptotics on the ℓ_ϵ scale. The limits are taken in the order $\epsilon \downarrow 0$, $r \to \infty$, and $\zeta \downarrow 0$. For fixed r and ζ, the proof first obtains

$$\limsup_{\epsilon \downarrow 0} \frac{S_{T^-} - S_{T_*}}{\ell_\epsilon} \le \frac{1}{\alpha_\zeta \left(\gamma_\infty - C_{\mathrm{md}} \eta_r\right)}.$$

Since $\eta_r \to 0$ (as $r \to \infty$) and $\alpha_\zeta \to 1$ (as $\zeta \downarrow 0$), we obtain $\limsup_{\epsilon \downarrow 0} \frac{S_{T^-} - S_{T_*}}{\ell_\epsilon} \le \frac{1}{\gamma_\infty}$.

For the reverse inequality, Proposition 2.1 gives $L_{T^-} \le 2G_{T^-} \le 2\epsilon \ell_\epsilon^b$. The loss-margin inequality in Lemma 8 therefore yields

$$\rho(w_{T^-}) \ge \log \frac{c_{\log}}{L_{T^-}} \ge \ell_\epsilon - b \log \ell_\epsilon + \log \frac{c_{\log}}{2},$$

where $c_{\log} := (\log 2)/n$. This implies that $\liminf_{\epsilon \downarrow 0} \frac{\rho(w_{T^-})}{\ell_\epsilon} \ge 1$. By the definition of γ_∞, we have $\rho(w_{T^-}) \le \gamma_\infty \|w_{T^-}\|_\infty$. On the other hand, writing $w_{T^-} = w_r - \sum_{t=r}^{T^- - 1} \eta_t D_t$ and applying Lemma 1 gives

$$\|w_{T^-}\|_\infty \le M_r + (S_{T^-} - S_r) + C_{\mathrm{rad}} \eta_r \left[1 + \log \frac{R_1^2 + \epsilon^2}{(1-\beta_2)\epsilon^2}\right],$$

where M_r is a bound on $\|w_r\|_\infty$ independent of ϵ. Chaining these two inequalities, rearranging, dividing by ℓ_ϵ, sending $\epsilon \downarrow 0$, and then letting $r \to \infty$, we obtain $\liminf_{\epsilon \downarrow 0} \frac{S_{T^-} - S_{T_*}}{\ell_\epsilon} \ge \frac{1}{\gamma_\infty}$. Combining the upper and lower bounds gives

$$S_{T^-} - S_{T_*} = \frac{1}{\gamma_\infty} \ell_\epsilon + o(\ell_\epsilon).$$

Part 2: Characterization of $S_{T^\delta} - S_{T^-}$**.** For $T^- \le t < T^\delta$, the definition of T^δ gives $G_t > \delta\epsilon$. Using the soft-sign bound from Lemma 18 and the gradient bound in Lemma 7, we obtain

$$\langle g_t, \sigma_\epsilon(g_t) \rangle \ge \frac{\|g_t\|_1^2}{\|g_t\|_1 + d\epsilon} \ge \frac{\gamma_\infty^2 G_t^2}{R_1 G_t + d\epsilon} \ge \frac{\gamma_\infty^2}{R_1 + d/\delta} G_t =: \kappa_\delta G_t.$$

Because $T^- \to \infty$ as $\epsilon \downarrow 0$, the term $C_{\mathrm{md}} \eta_t^2 G_t$ in the loss recursion can be absorbed for sufficiently small ϵ. Using $G_t \ge L_t/2$, we obtain $L_{t+1} \le \left(1 - \frac{\kappa_\delta}{4} \eta_t\right) L_t$ when $T^- \le t < T^\delta$. Since $L_{T^-} \le 2\epsilon \ell_\epsilon^b$, it follows after unrolling that

$$L_t \le 2\epsilon \ell_\epsilon^b \exp\left(-\frac{\kappa_\delta}{4}(S_t - S_{T^-})\right), \qquad T^- \le t \le T^\delta.$$

The right-hand side is at most $\delta\epsilon$ once

$$S_t - S_{T^-} \ge \frac{4}{\kappa_\delta} \log \frac{2\ell_\epsilon^b}{\delta} =: A_\epsilon > 0,$$

where the last inequality holds for sufficiently small ϵ. Consequently, if we denote $\bar{T} = \inf\{t \ge T^- : S_t - S_{T^-} \ge A_\epsilon\}$, it must hold that $\bar{T} \ge T^- + 1$ as $S_{\bar{T}} - S_{T^-} > 0$. Moreover, it also holds that $\bar{T} \ge T^\delta$. Otherwise, we would have $G_{\bar{T}} \le L_{\bar{T}} \le \delta\epsilon$, contradicting the definition of T^δ. Based on these two inequalities, we have by the minimality of $\bar{T}$ that

$$S_{T^\delta} - S_{T^-} \le S_{\bar{T}} - S_{T^-} = S_{\bar{T}-1} - S_{T^-} + \eta_{\bar{T}-1} < A_\epsilon + \eta_{\bar{T}-1} \le A_\epsilon + \eta_{T^-},$$

where we used $\eta_{\bar{T}-1} \le \eta_{T^-}$ as $\bar{T} - 1 \ge T^-$. Based on this, it follows that $S_{T^\delta} - S_{T^-} = O(\log \ell_\epsilon)$. Since $\log \ell_\epsilon = o(\ell_\epsilon)$, using the results of Part 1, we obtain

$$S_{T^\delta} - S_{T_*} = S_{T^\delta} - S_{T^-} + S_{T^-} - S_{T_*} = \frac{1}{\gamma_\infty} \ell_\epsilon + o(\ell_\epsilon).$$

Part 3: Characterization of $S_{T^+} - S_{T^\delta}$**.** Set $\tau := T^\delta$. We first prove inductively that $G_t \le L_t \le 2\delta\epsilon$, $t \ge \tau$. At $t = \tau$, this follows from $G_\tau \le \delta\epsilon$ and $L_\tau \le 2G_\tau$. Suppose that $L_t \le 2\delta\epsilon$. Then it holds that $\frac{G_t}{\epsilon} \le \frac{L_t}{\epsilon} \le 2\delta$. Lemma 17 gives $\epsilon \|D_t - \sigma_\epsilon(g_t)\|_2 \le C_{\mathrm{ema}}\eta_t G_t$. Moreover, we can show $\epsilon\|\sigma_\epsilon(g_t)\|_2 \le \|g_t\|_2 \le R_2 G_t$. Combining them together, we obtain $\|D_t\|_2 \le M_D \frac{G_t}{\epsilon}$ for a constant M_D independent of t and ϵ. Using $|g_t[j]| \le \|g_t\|_1 \le R_1 G_t \le 2R_1\delta\epsilon$ and Lemma 7, we obtain

$$a_\delta \frac{G_t^2}{\epsilon} \le \langle g_t, \sigma_\epsilon(g_t)\rangle \le R_2^2 \frac{G_t^2}{\epsilon}, \qquad a_\delta := \frac{\gamma_2^2}{1 + 2R_1\delta}.$$

The bound in Lemma 17 further gives $|\langle g_t, D_t - \sigma_\epsilon(g_t)\rangle| \le C_{\mathrm{ema}} R_2 \eta_t \frac{G_t^2}{\epsilon}$. Since $\eta_\tau \to 0$ as $\tau \to \infty$, the difference caused by using D_t instead of $\sigma_\epsilon(g_t)$ can be absorbed for sufficiently small ϵ. Therefore, it holds that

$$\frac{a_\delta}{2}\frac{G_t^2}{\epsilon} \le \langle g_t, D_t\rangle \le A_\delta \frac{G_t^2}{\epsilon}$$

for a constant $A_\delta > 0$. The Taylor expansion of the loss is $L_{t+1} = L_t - \eta_t\langle g_t, D_t\rangle + R_t$. Since $\frac{G_t}{\epsilon} \le 2\delta$ and $\|D_t\|_2 \le M_D \frac{G_t}{\epsilon}$, we can show the Taylor remainder R_t satisfies $0 \le R_t \le B_\delta \eta_t^2 \frac{G_t^3}{\epsilon^2}$ for a constant $B_\delta > 0$. Because $G_t/\epsilon \le 2\delta$ and $\eta_t \le \eta_\tau \to 0$, the Taylor remainder is lower order than the first-order descent term. Consequently, we have

$$\frac{a_\delta}{4}\frac{\eta_t}{\epsilon} G_t^2 \le L_t - L_{t+1} \le A_\delta \frac{\eta_t}{\epsilon} G_t^2.$$

In particular, it holds that $L_{t+1} \le L_t \le 2\delta\epsilon$, which closes the induction.

Since $t \ge \tau \ge T_*$, Proposition 2.1 gives $\frac{1}{2}L_t \le G_t \le L_t$. It follows that

$$c_\delta \frac{\eta_t}{\epsilon} L_t^2 \le L_t - L_{t+1} \le A_\delta \frac{\eta_t}{\epsilon} L_t^2, \tag{88}$$

for some $c_\delta > 0$. We also derive that $L_{t+1} \ge L_t/2$ for sufficiently small ϵ. Since $\frac{1}{L_{t+1}} - \frac{1}{L_t} = \frac{L_t - L_{t+1}}{L_t L_{t+1}}$, $L_t L_{t+1} \le L_t^2$, and $L_t L_{t+1} \ge \frac{1}{2}L_t^2$, we further obtain via (88) that

$$c_\delta \frac{\eta_t}{\epsilon} \le \frac{1}{L_{t+1}} - \frac{1}{L_t} \le C_\delta \frac{\eta_t}{\epsilon}.$$

Summing from $t = T^\delta$ to $T^+ - 1$ gives

$$c_\delta \frac{S_{T^+} - S_{T^\delta}}{\epsilon} \le \frac{1}{L_{T^+}} - \frac{1}{L_{T^\delta}} \le C_\delta \frac{S_{T^+} - S_{T^\delta}}{\epsilon}.$$

Finally, the stopping-time estimates from Lemma 21 imply

$$\epsilon\left(\frac{1}{L_{T^+}} - \frac{1}{L_{T^\delta}}\right) = \ell_\epsilon^b(1 + o(1)) - \frac{1}{\delta}(1 + o(1)) = \Theta(\ell_\epsilon^b),$$

from which we conclude that $S_{T^+} - S_{T^\delta} = \Theta(\ell_\epsilon^b)$.

Proof of Theorem 4. Fix $0 < \zeta < \frac{2}{n}$ and define $\alpha_\zeta := 1 - \frac{n\zeta}{2} > 0$. By Lemma 20, there is a time t_ζ (independent of ϵ) such that $L_t \le \zeta, t \ge t_\zeta$. Choose an integer r such that $r \ge t_\zeta$, $r \ge T_*$, $C_{\mathrm{md}}\eta_r < \gamma_\infty$, and $\gamma_\infty \eta_r \le \frac{1}{2}$. Recall that C_{md} is the constant in (44), i.e.

$$L_{t+1} \le L_t - \eta_t\langle g_t, \sigma_\epsilon(g_t)\rangle + C_{\mathrm{md}}\eta_t^2 G_t. \tag{89}$$

Importantly, the integer r is fixed before sending $\epsilon \downarrow 0$. Since $T^- \to \infty$ by Lemma 21, we have $T^- > r + 1$ for all sufficiently small ϵ. For $r \le t < T^-$, the definition of T^- gives $G_t > h_\epsilon = \epsilon\ell_\epsilon^b$. From $\langle g_t, \sigma_\epsilon(g_t)\rangle \ge \|g_t\|_1 - d\epsilon$ (Lemma 18) and $\|g_t\|_1 \ge \gamma_\infty G_t$ (Lemma 7), we have $\langle g_t, \sigma_\epsilon(g_t)\rangle \ge \gamma_\infty G_t - d\epsilon$. Substituting this into (89), we obtain

$$\begin{aligned} L_{t+1} &\le L_t - \eta_t(\gamma_\infty G_t - d\epsilon) + C_{\mathrm{md}}\eta_t^2 G_t \\ &= L_t - \eta_t(\gamma_\infty - C_{\mathrm{md}}\eta_t - \frac{d\epsilon}{G_t})G_t. \end{aligned} \tag{90}$$

From $G_t > \epsilon\ell_\epsilon^b$, we have $\frac{\epsilon}{G_t} < \frac{1}{\ell_\epsilon^b}$. This implies that $\gamma_\infty - C_{\mathrm{md}}\eta_t - \frac{d\epsilon}{G_t} \ge \gamma_\infty - C_{\mathrm{md}}\eta_r - \frac{d}{l_\epsilon^b}$. Define $q_{\epsilon,r} := \gamma_\infty - C_{\mathrm{md}}\eta_r - \frac{d}{\ell_\epsilon^b}$. For fixed r, since $\ell_\epsilon^b \to \infty$, we have $q_{\epsilon,r} \to \gamma_\infty - C_{\mathrm{md}}\eta_r > 0$. Therefore, we have $q_{\epsilon,r} > 0$ for all sufficiently small ϵ (for example, choosing ϵ such that $\frac{d}{\ell_\epsilon^b} \le \frac{\gamma_\infty - C_{\mathrm{md}}\eta_r}{2}$) and

$$L_{t+1} \le L_t - q_{\epsilon,r}\eta_t G_t. \tag{91}$$

Part 1 From Lemma 8, we know $\frac{G_t}{L_t} \geq 1 - \frac{nL_t}{2}$. Since $t \geq r \geq t_\zeta$, from (84), we have $G_t \geq (1 - \frac{n\zeta}{2})L_t = \alpha_\zeta L_t$. Substituting this into (91), we have $L_{t+1} \leq (1 - \alpha_\zeta q_{\epsilon,r}\eta_t)L_t$. Given $0 < q_{\epsilon,r} \leq \gamma_\infty$ for sufficiently small ϵ, we have $0 < \alpha_\zeta q_{\epsilon,r}\eta_t \leq \gamma_\infty \eta_r \leq \frac{1}{2}$. Iterating from r to $t-1$ and using $1 - x \leq e^{-x}$, we obtain

$$L_t \leq L_r \prod_{s=r}^{t-1}(1 - \alpha_\zeta q_{\epsilon,r}\eta_s) \leq L_r \exp(-\alpha_\zeta q_{\epsilon,r} \sum_{s=r}^{t-1} \eta_s) = L_r \exp(-\alpha_\zeta q_{\epsilon,r}(S_t - S_r)). \tag{92}$$

This holds for every $r \leq t < T^-$. By minimality of T^- and Lemma 8, we have $L_{T^- - 1} \geq G_{T^- - 1} > h_\epsilon$. Applying (92) with $t = T^- - 1$, we obtain

$$h_\epsilon < L_{T^- - 1} \leq L_r \exp(-\alpha_\zeta q_{\epsilon,r}(S_{T^- - 1} - S_r)).$$

Rearranging this leads to $S_{T^- - 1} - S_r \leq \frac{1}{\alpha_\zeta q_{\epsilon,r}} \log \frac{L_r}{h_\epsilon}$. Since $S_{T^-} = S_{T^- - 1} + \eta_{T^- - 1}$ and $\eta_{T^- - 1} \leq \eta_r$, we further obtain

$$S_{T^-} - S_r \leq \frac{1}{\alpha_\zeta q_{\epsilon,r}} \log \frac{L_r}{h_\epsilon} + \eta_r.$$

Because $L_r \leq \zeta$, we have $\log \frac{L_r}{h_\epsilon} \leq \log \frac{\zeta}{\epsilon \ell_\epsilon^b} = \ell_\epsilon - b \log \ell_\epsilon + \log \zeta$. Given this, we obtain

$$S_{T^-} - S_r \leq \frac{\ell_\epsilon - b \log \ell_\epsilon + \log \zeta}{\alpha_\zeta(\gamma_\infty - C_{\mathrm{md}}\eta_r - d\ell_\epsilon^{-b})} + \eta_r = \ell_\epsilon \frac{1 - \frac{b \log \ell_\epsilon}{\ell_\epsilon} + \frac{\log \zeta}{\ell_\epsilon}}{\alpha_\zeta(\gamma_\infty - C_{\mathrm{md}}\eta_r - d\ell_\epsilon^{-b})} + \eta_r.$$

From $S_{T^-} - S_{T_*} = (S_{T^-} - S_r) + (S_r - S_{T_*})$, we have

$$\frac{S_{T^-} - S_{T_*}}{\ell_\epsilon} \leq \frac{1 - \frac{b \log \ell_\epsilon}{\ell_\epsilon} + \frac{\log \zeta}{\ell_\epsilon}}{\alpha_\zeta(\gamma_\infty - C_{\mathrm{md}}\eta_r - d\ell_\epsilon^{-b})} + \frac{S_r - S_{T_*} + \eta_r}{\ell_\epsilon},$$

Note that $S_r - S_{T_*} + \eta_r$ is independent of ϵ. Using $\frac{\log \ell_\epsilon}{\ell_\epsilon} \to 0$ and $\ell_\epsilon^{-b} \to 0$ as $\epsilon \downarrow 0$, we obtain

$$\limsup_{\epsilon \downarrow 0} \frac{S_{T^-} - S_{T_*}}{\ell_\epsilon} \leq \frac{1}{\alpha_\zeta(\gamma_\infty - C_{\mathrm{md}}\eta_r)}.$$

This holds for every sufficiently large fixed $r \geq t_\zeta$. Given $\eta_r \to 0$ as $r \to \infty$, we have $\limsup_{\epsilon \downarrow 0} \frac{S_{T^-} - S_{T_*}}{\ell_\epsilon} \leq \frac{1}{\alpha_\zeta \gamma_\infty}$. Finally, letting $\zeta \downarrow 0$ gives $\alpha_\zeta = 1 - \frac{n\zeta}{2} \to 1$ and we conclude that

$$\limsup_{\epsilon \downarrow 0} \frac{S_{T^-} - S_{T_*}}{\ell_\epsilon} \leq \frac{1}{\gamma_\infty}. \tag{93}$$

We now prove $\liminf_{\epsilon \downarrow 0} \frac{S_{T^-} - S_{T_*}}{\ell_\epsilon} \geq \frac{1}{\gamma_\infty}$. By definition, we have $G_{T^-} \leq h_\epsilon$. Because $T^- \geq T_*$, Proposition 2.1 gives $L_{T^-} \leq 2G_{T^-} \leq 2h_\epsilon = 2\epsilon\ell_\epsilon^b$. By Lemma 8, we have $L_t \geq \frac{\log 2}{n} e^{-\rho(w_t)}$. From this, we then have $\rho(w_{T-}) \geq \log \frac{c_{\log}}{L_{T^-}}$ where we let $c_{\log} := \frac{\log 2}{n}$. From $L_{T^-} \leq 2\epsilon\ell_\epsilon^b$, we obtain

$$\rho(w_{T^-}) \geq \log \frac{c_{\log}}{2\epsilon\ell_\epsilon^b} = \ell_\epsilon - b \log \ell_\epsilon + \log \frac{c_{\log}}{2}.$$

This implies that

$$\liminf_{\epsilon \downarrow 0} \frac{\rho(w_{T^-})}{\ell_\epsilon} \geq 1. \tag{94}$$

By the definition of γ_∞, we have $\rho(w) \leq \gamma_\infty \|w\|_\infty$ for every $w \in \mathbb{R}^d$. Therefore, we have $\frac{\rho(w_{T^-})}{\gamma_\infty} \leq \|w_{T^-}\|_\infty$. Given $T^- > r$ for sufficiently small ϵ, we have $w_{T^-} = w_r - \sum_{t=r}^{T^- - 1} \eta_t D_t$ and consequently

$$\|w_{T^-}\|_\infty \leq \|w_r\|_\infty + \|\sum_{t=r}^{T^- - 1} \eta_t D_t\|_\infty. \tag{95}$$

The first term on the right-hand side can be bounded as: $\|w_r\|_\infty \le \|w_0\|_\infty + C_D S_r =: M_r$. Note that M_r is independent of ϵ. From Lemma 7, we know $|g_t[j]| \le \|g_t\|_\infty \le R_1 G_t \le R_1$. Therefore, by taking $B_g = R_1$, we obtain

$$\left\| \sum_{t=r}^{T^- - 1} \eta_t D_t \right\|_\infty \le S_{T^-} - S_r + C_{\mathrm{rad}} \eta_r \big[1 + \log \frac{R_1^2 + \epsilon^2}{(1-\beta_2)\epsilon^2}\big]. \tag{96}$$

Substituting (96) into (95), we obtain

$$\frac{\rho(w_{T^-})}{\gamma_\infty} \le \|w_{T^-}\|_\infty \le M_r + S_{T^-} - S_r + C_{\mathrm{rad}} \eta_r [1 + \log \frac{R_1^2 + \epsilon^2}{(1-\beta_2)\epsilon^2}].$$

Rearranging and dividing by ℓ_ϵ gives

$$\frac{S_{T^-} - S_r}{\ell_\epsilon} \ge \frac{1}{\gamma_\infty} \frac{\rho(w_{T^-})}{\ell_\epsilon} - \frac{M_r}{\ell_\epsilon} - C_{\mathrm{rad}} \eta_r \frac{1}{\ell_\epsilon} [1 + \log \frac{R_1^2 + \epsilon^2}{(1-\beta_2)\epsilon^2}].$$

By taking $\epsilon \downarrow 0$ and using (94), we obtain

$$\liminf_{\epsilon \downarrow 0} \frac{S_{T^-} - S_{T_*}}{\ell_\epsilon} \ge \frac{1}{\gamma_\infty} - 2C_{\mathrm{rad}} \eta_r. \tag{97}$$

where we used $\log \frac{R_1^2+\epsilon^2}{(1-\beta_2)\epsilon^2} = 2\log\frac{1}{\epsilon} + \log \frac{R_1^2+\epsilon^2}{1-\beta_2} = 2\ell_\epsilon + O(1)$ and consequently $\frac{1}{\ell_\epsilon}[1 + \log \frac{R_1^2+\epsilon^2}{(1-\beta_2)\epsilon^2}] \to 2$. (97) holds for every sufficiently large fixed r. We let $r \to \infty$ and use $\eta_r \to 0$ to conclude that

$$\liminf_{\epsilon \downarrow 0} \frac{S_{T^-} - S_{T_*}}{\ell_\epsilon} \ge \frac{1}{\gamma_\infty}. \tag{98}$$

Combining (93) and (98), we have

$$\frac{1}{\gamma_\infty} \le \liminf_{\epsilon \downarrow 0} \frac{S_{T^-} - S_{T_*}}{\ell_\epsilon} \le \limsup_{\epsilon \downarrow 0} \frac{S_{T^-} - S_{T_*}}{\ell_\epsilon} \le \frac{1}{\gamma_\infty}.$$

Therefore, we obtain

$$\frac{S_{T^-} - S_{T_*}}{\ell_\epsilon} \longrightarrow \frac{1}{\gamma_\infty}, \qquad \text{as } \epsilon \downarrow 0,$$

or equivalently,

$$S_{T^-} - S_{T_*} = \frac{1}{\gamma_\infty} \ell_\epsilon + o(\ell_\epsilon). \tag{99}$$

Part 2 Next, we prove $S_{T^\delta} - S_{T^-} = O(\log \ell_\epsilon)$. Using $\|g_t\|_1 \ge \gamma_\infty G_t$, $\|g_t\|_1 \le R_1 G_t$ and (43), we get

$$\langle g_t, \sigma_\epsilon(g_t) \rangle \ge \frac{\gamma_\infty^2 G_t^2}{R_1 G_t + d\epsilon} = \frac{\gamma^2 G_t}{R_1 + d\epsilon / G_t} \ge \frac{\gamma_\infty^2 G_t}{R_1 + d/\delta}.$$

Define $\kappa_\delta = \frac{\gamma_\infty^2}{R_1 + d/\delta} = \frac{\gamma_\infty^2 \delta}{R_1 \delta + d} > 0$. Then we have $\langle g_t, \sigma_\epsilon(g_t) \rangle \ge \kappa_\delta G_t$. Substituting this into (89), we obtain

$$L_{t+1} \le L_t - \eta_t (\kappa_\delta - C_{\mathrm{md}} \eta_t) G_t.$$

Since $T^- \to \infty$, we have $\eta_{T^-} \to 0$. Thus, for sufficiently small ϵ, we have $C_{\mathrm{md}} \eta_t \le C_{\mathrm{md}} \eta_{T^-} \le \frac{\kappa_\delta}{2}, t \ge T^-$. This leads to $L_{t+1} \le L_t - \frac{\kappa_\delta}{2} \eta_t G_t$. Using $G_t \ge L_t/2, \forall t \ge T_*$, we further obtain

$$L_{t+1} \le (1 - \frac{\kappa_\delta}{4} \eta_t) L_t. \tag{100}$$

At the initial time, it holds that $L_{T^-} \le 2G_{T^-} \le 2\epsilon \ell_\epsilon^b$. Iterating (100) leads to

$$L_t \le 2\epsilon \ell_\epsilon^b \exp(-\frac{\kappa_\delta}{4}(S_t - S_{T^-})) =: U_t, \qquad T^- \le t \le T^\delta.$$

Let $\bar{T}$ be the first time at which the right side is at most $\delta\epsilon$, i.e. $\bar{T} := \inf\{t \geq T^- : U_t \leq \delta\epsilon\}$. The condition $U_t \leq \delta\epsilon$ is equivalent to

$$S_t - S_{T^-} \geq \frac{4}{\kappa_\delta} \log \frac{2\ell_\epsilon^b}{\delta} =: A_\epsilon.$$

Thus, we can rewrite $\bar{T}$ as $\bar{T} = \inf\{t \geq T^- : S_t - S_{T^-} \geq A_\epsilon\}$. For sufficiently small ϵ, we have $\frac{2l_\epsilon^b}{\delta} > 1$ and consequently $A_\epsilon > 0$. Therefore, we have $\bar{T} \geq T^- + 1$ when ϵ is sufficiently small. Moreover, we must have $\bar{T} \geq T^\delta$, otherwise we would have $G_{\bar{T}} \leq L_{\bar{T}} \leq \delta\epsilon$ contradicting the definition of T^δ. Since S_t is increasing and $T^\delta \leq \bar{T}$, we have $S_{T^\delta} \leq S_{\bar{T}}$ and

$$S_{T^\delta} - S_{T^-} \leq S_{\bar{T}} - S_{T^-}. \tag{101}$$

By the minimality of $\bar{T}$, we have

$$S_{\bar{T}} - S_{T^-} = S_{\bar{T}-1} - S_{T^-} + \eta_{\bar{T}-1} < A_\epsilon + \eta_{\bar{T}-1} \leq A_\epsilon + \eta_{T^-},$$

where we used $\eta_{\bar{T}-1} \leq \eta_{T^-}$ in the last inequality as $\bar{T} - 1 \geq T^-$. Combining this with (101) and using the definition of A_ϵ, we obtain

$$S_{T^\delta} - S_{T^-} \leq \frac{4}{\kappa_\delta} \log \frac{2\ell_\epsilon^b}{\delta} + \eta_{T^-} = \frac{4b}{\kappa_\delta} \log \ell_\epsilon + \frac{4}{\kappa_\delta} \log \frac{2}{\delta} + \eta_{T^-}.$$

Since $\frac{4}{\kappa_\delta} \log \frac{2}{\delta}$ is a constant independent of ϵ and $\eta_{T^-} \to 0$ as $\epsilon \downarrow 0$, we conclude that $S_{T^\delta} - S_{T^-} = O(\log \ell_\epsilon)$. Combining this with (99), we obtain

$$S_{T^\delta} - S_{T^*} = (S_{T^-} - S_{T^*}) + (S_{T^\delta} - S_{T^-}) = \frac{1}{\gamma_\infty} \ell_\epsilon + o(\ell_\epsilon).$$

Part 3 Finally, we prove $S_{T^+} - S_{T^\delta} = \Theta(\ell_\epsilon^b)$. Set $\tau := T^\delta$. We prove inductively that $L_t \leq 2\delta\epsilon, t \geq \tau$. For the base case, we have $G_\tau \leq \delta\epsilon$ by definition. Because $\tau \geq T_*$, by Proposition 2.1, it holds that $L_\tau \leq 2G_\tau \leq 2\delta\epsilon$. Now, we assume that the statement holds at time t and try to show that it also holds at $t+1$. Since $G_t \leq L_t$, we have $\lambda_t = \frac{G_t}{\epsilon} \leq \frac{L_t}{\epsilon} \leq 2\delta$. Lemma 17 gives $\epsilon\|D_t - \sigma_\epsilon(g_t)\|_2 \leq C_{\text{ema}}\eta_t G_t$. Also, coordinatewise, we have $\epsilon|\sigma_\epsilon(g_t)[j]| = \frac{\epsilon|g_t[j]|}{|g_t[j]|+\epsilon} \leq |g_t[j]|$. Hence, we have $\epsilon\|\sigma_\epsilon(g_t)\|_2 \leq \|g_t\|_2 \leq R_2 G_t$. Putting things together, we obtain

$$\epsilon\|D_t\|_2 \leq (R_2 + C_{\text{ema}}\eta_t)G_t \leq M_D G_t,$$

where $M_D := R_2 + C_{\text{ema}}\eta_0$. Therefore, we have $\|D_t\|_2 \leq M_D \frac{G_t}{\epsilon}$.

Because $|g_t[j]| \leq \|g_t\|_1 \leq R_1 G_t$ and $\frac{G_t}{\epsilon} \leq 2\delta$, we have $|g_t[j]| + \epsilon \leq \epsilon(1 + 2R_1\delta)$. Therefore, we obtain

$$\langle g_t, \sigma_\epsilon(g_t)\rangle = \sum_j \frac{g_t[j]^2}{|g_t[j]| + \epsilon} \geq \frac{\|g_t\|_2^2}{\epsilon(1+2R_1\delta)} \geq \frac{\gamma_2^2}{1+2R_1\delta} \frac{G_t^2}{\epsilon} = a_\delta \frac{G_t^2}{\epsilon},$$

where we define $a_\delta := \frac{\gamma_2^2}{1+2R_1\delta} > 0$. Moreover, we further obtain

$$\langle g_t, \sigma_\epsilon(g_t)\rangle \leq \frac{1}{\epsilon} \sum_{j=1}^d g_t[j]^2 = \frac{\|g_t\|_2^2}{\epsilon} \leq R_2^2 \frac{G_t^2}{\epsilon}.$$

Denote $\sigma_t := \sigma_\epsilon(g_t)$. Next, we bound the term $|\langle g_t, D_t - \sigma_t\rangle|$ as

$$|\langle g_t, D_t - \sigma_t\rangle| \leq \|g_t\|_2 \|D_t - \sigma_t\|_2 \leq R_2 G_t (C_{\text{ema}}\eta_t \frac{G_t}{\epsilon}) = C_{\text{ema}} R_2 \eta_t \frac{G_t^2}{\epsilon}.$$

We can write $\langle g_t, D_t\rangle = \langle g_t, \sigma_t\rangle + \langle g_t, D_t - \sigma_t\rangle$. For the lower bound, we have

$$\langle g_t, D_t\rangle \geq \langle g_t, \sigma_t\rangle - |\langle g_t, D_t - \sigma_t\rangle| \geq a_\delta \frac{G_t^2}{\epsilon} - C_{\text{ema}} R_2 \eta_t \frac{G_t^2}{\epsilon} = (a_\delta - C_{\text{ema}} R_2 \eta_t) \frac{G_t^2}{\epsilon}.$$

For the upper bound, we have

$$\langle g_t, D_t\rangle \le \langle g_t, \sigma_t\rangle + |\langle g_t, D_t - \sigma_t\rangle| \le R_2^2 \frac{G_t^2}{\epsilon} + C_{\text{ema}} R_2 \eta_t \frac{G_t^2}{\epsilon} = (R_2^2 + C_{\text{ema}} R_2 \eta_t)\frac{G_t^2}{\epsilon}.$$

Recall that $\tau = T^\delta = T_1^\delta(\epsilon)$. It holds that $\tau \to \infty$ and $\eta_\tau \to 0$ as $\epsilon \downarrow 0$. Consequently, we have for sufficiently small ϵ that

$$C_{\text{ema}} R_2 \eta_t \le C_{\text{ema}} R_2 \eta_\tau \le \frac{a_\delta}{2}, \qquad t \ge \tau.$$

Substituting this into the lower bound, we obtain $a_\delta - C_{\text{ema}} R_2 \eta_t \ge a_\delta - \frac{a_\delta}{2} = \frac{a_\delta}{2}$. For the upper bound, we have $R_2^2 + C_{\text{ema}} R_2 \eta_t \le R_2^2 + \frac{a_\delta}{2}$. Putting things together, we conclude that

$$\frac{a_\delta}{2}\frac{G_t^2}{\epsilon} \le \langle g_t, D_t\rangle \le A_\delta \frac{G_t^2}{\epsilon}, \tag{102}$$

where $A_\delta := R_2^2 + \frac{a_\delta}{2} = R_2^2 + \frac{\gamma_2^2}{2(1+2R_1\delta)}$. Recall the Taylor remainder formula used in Lemma 13:

$$L_{t+1} = L_t - \eta_t \langle g_t, D_t\rangle + R_t,$$

where $R_t = \eta_t^2 \int_0^1 (1-s) D_t^T \nabla^2 L(w_t - s\eta_t D_t) D_t ds \ge 0$. We now exploit the bound $\|D_t\|_2 \le M_D \frac{G_t}{\epsilon}$ shown above. Let $\tilde{w}_{t,s} := w_t - s\eta_t D_t$. For every sample i, Lemma 10 gives $a_i(\tilde{w}_{t,s}) \le e^{|z_i^T(\tilde{w}_{t,s} - w_t)|} a_i(w_t)$. Since $\frac{G_t}{\epsilon} \le 2\delta$ and $\|D_t\|_2 \le M_D \frac{G_t}{\epsilon}$, defining $E_\delta := \exp(2\delta R_2 M_D \eta_0)$, we have $a_i(\tilde{w}_{t,s}) \le E_\delta a_i(w_t)$. Recall that $a_i(w) = -\ell'(z_i^T w)$. From (37), we have

$$D_t^T \nabla^2 L(\tilde{w}_{t,s}) D_t \le \frac{E_\delta}{n}\sum_i a_i(w_t) R_2^2 \|D_t\|_2^2 = E_\delta R_2^2 G_t \|D_t\|_2^2 \le E_\delta R_2^2 M_D^2 \frac{G_t^3}{\epsilon^2}.$$

Consequently, we have

$$0 \le R_t \le B_\delta \eta_t^2 \frac{G_t^3}{\epsilon^2}, \tag{103}$$

where $B_\delta = \frac{1}{2} E_\delta R_2^2 M_D^2$. Denote $\triangle_t := L_t - L_{t+1}$. From $\triangle_t = \eta_t \langle g_t, D_t\rangle - R_t$ and using (102) and (103), we have

$$\triangle_t \ge \frac{a_\delta}{2}\eta_t \frac{G_t^2}{\epsilon} - B_\delta \eta_t^2 \frac{G_t^3}{\epsilon^2} = \eta_t \frac{G_t^2}{\epsilon}\left(\frac{a_\delta}{2} - B_\delta \eta_t \frac{G_t}{\epsilon}\right).$$

Since $\frac{G_t}{\epsilon} \le 2\delta$ and $\eta_t \le \eta_\tau$ (as $t \ge \tau$), we have for sufficiently small ϵ that

$$B_\delta \eta_t \frac{G_t}{\epsilon} \le 2\delta B_\delta \eta_\tau \le \frac{a_\delta}{4}.$$

Therefore, it holds that $\triangle_t \ge \frac{a_\delta}{4}\eta_t \frac{G_t^2}{\epsilon} \ge 0$. This implies that $L_{t+1} \le L_t \le 2\delta\epsilon$. This closes the induction argument and shows that $L_t \le 2\delta\epsilon, t \ge \tau$.

Given $R_t \ge 0$, we have $\triangle_t = \eta_t \langle g_t, D_t\rangle - R_t \le \eta_t \langle g_t, D_t\rangle$. Using the upper bound in (102), we conclude that

$$\frac{a_\delta}{4}\frac{\eta_t}{\epsilon} G_t^2 \le L_t - L_{t+1} \le A_\delta \frac{\eta_t}{\epsilon} G_t^2.$$

Since $t \ge \tau \ge T_*$, Proposition 2.1 gives $\frac{1}{2}L_t \le G_t \le L_t$ and consequently $\frac{1}{4}L_t^2 \le G_t^2 \le L_t^2$. Therefore, we further obtain

$$c_\delta \frac{\eta_t}{\epsilon} L_t^2 \le L_t - L_{t+1} \le A_\delta \frac{\eta_t}{\epsilon} L_t^2, \tag{104}$$

where $c_\delta := \frac{a_\delta}{16}$. Given this, we have for $t \ge \tau$ that

$$\frac{L_t - L_{t+1}}{L_t} \le A_\delta \eta_t \frac{L_t}{\epsilon} \le 2\delta A_\delta \eta_\tau. \tag{105}$$

Since $\eta_\tau \to 0$, for sufficiently small ϵ, we have $L_{t+1} \geq \frac{1}{2} L_t$, $t \geq \tau$. Note that $\frac{1}{L_{t+1}} - \frac{1}{L_t} = \frac{L_t - L_{t+1}}{L_t L_{t+1}}$. Given $L_t L_{t+1} \leq L_t^2$ (as $L_{t+1} \leq L_t$), we obtain via the lower bound in (104) that

$$\frac{1}{L_{t+1}} - \frac{1}{L_t} \geq c_\delta \frac{\eta_t}{\epsilon}.$$

Given $L_t L_{t+1} \geq \frac{1}{2} L_t^2$, we obtain via the upper bound in (104) that

$$\frac{1}{L_{t+1}} - \frac{1}{L_t} \leq 2A_\delta \frac{\eta_t}{\epsilon}.$$

Combining them together, we conclude that $c_\delta \frac{\eta_t}{\epsilon} \leq \frac{1}{L_{t+1}} - \frac{1}{L_t} \leq 2A_\delta \frac{\eta_t}{\epsilon}$, $t \geq \tau$. Recall that $\tau = T^\delta$. Letting $C_\delta = 2A_\delta$ and summing from $t = T^\delta$ to $T^+ - 1$ gives

$$c_\delta \frac{S_{T^+} - S_{T^\delta}}{\epsilon} \leq \frac{1}{L_{T^+}} - \frac{1}{L_{T^\delta}} \leq C_\delta \frac{S_{T^+} - S_{T^\delta}}{\epsilon}. \tag{106}$$

Given $\frac{L_{T^+}}{\epsilon \ell_\epsilon^{-b}} \to 1$ and $\frac{L_{T^\delta}}{\delta \epsilon} \to 1$, we can write $L_{T^+} = \epsilon \ell_\epsilon^{-b}(1 + r_\epsilon)$ and $L_{T^\delta} = \delta\epsilon(1 + s_\epsilon)$ where $r_\epsilon \to 0$ and $s_\epsilon \to 0$ as $\epsilon \downarrow 0$. Define $\tilde{r}_\epsilon := -\frac{r_\epsilon}{1 + r_\epsilon}$ and $\tilde{s}_\epsilon := -\frac{s_\epsilon}{1 + s_\epsilon}$. Then we have

$$\frac{\epsilon}{L_{T^+}} = \ell_\epsilon^b (1 + \tilde{r}_\epsilon) = \ell_\epsilon^b (1 + o(1)), \qquad \frac{\epsilon}{L_{T^\delta}} = \frac{1}{\delta}(1 + \tilde{s}_\epsilon) = \frac{1}{\delta}(1 + o(1)).$$

Combining them together, we have

$$\epsilon \left(\frac{1}{L_{T^+}} - \frac{1}{L_{T^\delta}} \right) = \ell_\epsilon^b (1 + o(1)) - \frac{1}{\delta}(1 + o(1)) = \Theta(\ell_\epsilon^b). \tag{107}$$

Multiplying (106) by ϵ and using (107) leads to

$$S_{T^+} - S_{T^\delta} = \Theta(\ell_\epsilon^b).$$

□

E.2 Implications of Theorem 4

Corollary 22. *For any $\delta \in (0, 1)$, there exists a constant $\epsilon_\delta^{\mathrm{inv}}$ such that for every $0 < \epsilon \leq \epsilon_\delta^{\mathrm{inv}}$, it holds that*

$$G_t \leq L_t \leq 2\delta\epsilon, \qquad \forall t \geq T_1^\delta(\epsilon)$$

Proof. This is already proved in Part 3 of the proof of Theorem 4. Here, we specify the constant $\epsilon_\delta^{\mathrm{inv}}$. The two requirements appearing in Part 3 are:

$$C_{\mathrm{ema}} R_2 \eta_\tau \leq \frac{a_\delta}{2}, \qquad 2\delta B_\delta \eta_\tau \leq \frac{a_\delta}{4},$$

where $\tau := T_1^\delta(\epsilon)$, $a_\delta := \frac{\gamma_2^2}{1 + 2R_1\delta}$, $B_\delta := \frac{1}{2} E_\delta R_2^2 M_D^2$, and $M_D := R_2 + C_{\mathrm{ema}} \eta_0$. Define $\bar{\eta}_\delta := \min\{\eta_0, \frac{a_\delta}{2C_{\mathrm{ema}} R_2}, \frac{a_\delta}{8\delta B_\delta}\}$. It is sufficient to require $\eta_\tau \leq \bar{\eta}_\delta$. Choosing $N_\delta := \max\{T_*, t_{\mathrm{ema}}, \lceil (\frac{\eta_0}{\bar{\eta}_\delta})^{1/a} \rceil\}$, it holds that $\eta_{N_\delta} \leq \bar{\eta}_\delta$.

Lemma 12 gives $G_t \geq G_0 e^{-K_0 S_t}$ where $K_0 = R_1 C_D$. Consequently, for every $0 \leq t \leq N_\delta$, we have

$$G_t \geq G_0 e^{-K_0 S_{N_\delta}} =: c_{N_\delta} > 0.$$

Set $\epsilon_\delta^{\mathrm{inv}} := \min\{1, \frac{c_{N_\delta}}{2\delta}\}$. Next, we show this choice works. Recall that $\tau = T_1^\delta(\epsilon) = \inf\{t \geq T_* : G_t \leq \delta\epsilon\}$. If $0 < \epsilon \leq \epsilon_\delta^{\mathrm{inv}}$, then it holds that

$$\delta\epsilon \leq \frac{c_{N_\delta}}{2} < c_{N_\delta} \leq G_t, \qquad 0 \leq t \leq N_\delta.$$

Therefore, the event $G_t \leq \delta\epsilon$ cannot occur before or at N_δ, which implies that $\tau > N_\delta$. Since the learning rate is decreasing, it holds that $\eta_\tau \leq \eta_{N_\delta} \leq \bar{\eta}_\delta$ as desired. □

Corollary 23. *Suppose that $p, q \in (0,1)$ are two constants such that $p+q=1$. For any $0<\delta<\min\{1, \frac{p}{2R_1 q}\}$, there exists a constant $\hat{\epsilon}_1(\delta)$ such that for every $0<\epsilon\le\hat{\epsilon}_1(\delta)$, we have $\mathfrak{r}_t^{\mathrm{grad}} \le p$, $\mathfrak{r}_t^{\mathrm{sgn}} \ge q$ for every $t \ge T_1^\delta(\epsilon)$.*

Proof of Corollary 23. Define $A'_\delta := \frac{2R_1\delta}{1+2R_1\delta}$. We first choose δ such that $A'_\delta < p$ (equivalently, $2R_1(1-p)\delta < p$). Therefore, a feasible δ satisfies

$$0<\delta<\min\{1, \frac{p}{2R_1(1-p)}\} = \min\{1, \frac{p}{2R_1 q}\}.$$

For any fixed δ in this range, we define a slack variable $\omega_{\delta,p} := p - \frac{2R_1\delta}{1+2R_1\delta} = \frac{p-2R_1 q\delta}{1+2R_1\delta} > 0$. Recall the constants used in Corollary 22: $a_\delta := \frac{\gamma_2^2}{1+2R_1\delta}$, $M_D := R_2 + C_{\mathrm{ema}}\eta_0$, $E_\delta := \exp(2\delta R_2 M_D \eta_0)$, and $B_\delta := \frac{1}{2}E_\delta R_2^2 M_D^2$. We set

$$\bar{\eta}_{\delta,p} := \min\{\eta_0, \frac{a_\delta}{2C_{\mathrm{ema}}R_2}, \frac{a_\delta}{8\delta B_\delta}, \frac{\omega_{\delta,p}}{C_{\mathrm{loc}}}\},$$

where $C_{\mathrm{loc}} = C_{\mathrm{ema}}\max\{\frac{1}{\gamma_\infty}, \frac{1}{\gamma_2}\}$ is the constant in Proposition C.1. Note that the first three constants in the definition of $\bar{\eta}_{\delta,p}$ are used to ensure the invariant conclusion of Corollary 22 holds. Since $\eta_t = \eta_0(t+1)^{-a}$, we define

$$N_{\delta,p} := \max\{T_*, t_{\mathrm{ema}}, \lceil(\frac{\eta_0}{\bar{\eta}_{\delta,p}})^{1/a}\rceil\},$$

and then it holds that $\eta_{N_{\delta,p}} \le \bar{\eta}_{\delta,p}$. We further define $c_{N_{\delta,p}} := G_0\exp(-K_0 S_{N_{\delta,p}})$ where $K_0 = R_1 C_D$. Next, we show the choice $\hat{\epsilon}_1(\delta) := \min\{1, \frac{c_{N_{\delta,p}}}{2\delta}\}$ works. Let $\tau := T_1^\delta(\epsilon) = \inf\{t \ge T_* : G_t \le \delta\epsilon\}$. For $0<\epsilon\le\hat{\epsilon}_1(\delta)$, by Lemma 12, we have for every $t \le N_{\delta,p}$ that

$$G_t \ge G_0 e^{-K_0 S_t} \ge G_0 e^{-K_0 S_{N_{\delta,p}}} = c_{N_{\delta,p}} > \frac{c_{N_{\delta,p}}}{2} \ge \delta\epsilon.$$

Therefore, the event $G_t \le \delta\epsilon$ cannot occur at or before $N_{\delta,p}$. This implies that $\tau > N_{\delta,p}$. Since the learning rate is decreasing, we have

$$\eta_\tau \le \eta_{N_{\delta,p}} \le \bar{\eta}_{\delta,p} \le \frac{\omega_{\delta,p}}{C_{\mathrm{loc}}}.$$

Consequently, it holds that $C_{\mathrm{loc}}\eta_\tau \le \omega_{\delta,p}$. For every $t \ge \tau$, Corollary 22 gives $\lambda_t = \frac{G_t}{\epsilon} \le \frac{L_t}{\epsilon} \le 2\delta$. Using Proposition C.1 and $\eta_t \le \eta_\tau$, it holds that

$$\mathfrak{r}_t^{\mathrm{grad}} \le \frac{2R_1\delta}{1+2R_1\delta} + C_{\mathrm{loc}}\eta_\tau \le A'_\delta + \omega_{\delta,p} = p.$$

Similarly, it also holds that

$$\mathfrak{r}_t^{\mathrm{sgn}} \ge \frac{1}{1+2R_1\delta} - C_{\mathrm{loc}}\eta_\tau = 1 - A'_\delta - C_{\mathrm{loc}}\eta_\tau \ge 1 - A'_\delta - \omega_{\delta,p} = q.$$

□

E.3 Proof of Proposition 3.1

Lemma 24. *For a nonnegative integer T, define $V_{T,t} := \sum_{s=T}^{t-1}\eta_s$ and $V_{T,T} = 0$. It holds that*

$$\sum_{t=T}^{\infty}\frac{\eta_t^2}{1+V_{T,t}} \le Q_\eta, \qquad Q_\eta := \eta_0^2 + \frac{1}{a}\max\{\eta_0^2, \eta_0(1-a)\}.$$

Proof. For $a=1$ (i.e. $\eta_t = \frac{\eta_0}{t+1}$), we have

$$\sum_{t=T}^{\infty}\frac{\eta_t^2}{1+V_{T,t}} \le \sum_{t=T}^{\infty}\eta_t^2 = \eta_0^2\sum_{t=T}^{\infty}\frac{1}{(t+1)^2} \le \eta_0^2\sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{6}\eta_0^2 \le 2\eta_0^2 = Q_\eta.$$

Next, we consider the case where $0 < a < 1$. Fix any $T \geq 0$ and define the functions $f(x) := \eta_0(x+1)^{-a}, x \geq T$ and $F_T(x) := \int_T^x f(r)dr$. Note that $f(t) = \eta_t$. Because f is positive and decreasing, it holds that $f(s) \geq \int_s^{s+1} f(x)dx$ for every $s \geq T$. Therefore, we have

$$V_{T,t} = \sum_{s=T}^{t-1} f(s) \geq \sum_{s=T}^{t-1} \int_s^{s+1} f(x)dx = \int_T^t f(x)dx = F_T(t).$$

From this, it holds that $\frac{\eta_t^2}{1+V_{T,t}} \leq \frac{f(t)^2}{1+F_T(t)}$.

We further define $h_T(x) := \frac{f(x)^2}{1+FT(x)}$. Differentiating this function gives

$$h_T'(x) = \frac{2f(x)f'(x)(1+F_T(x)) - f(x)^3}{(1+F_T(x))^2},$$

where we used $F_T'(x) = f(x)$. Given $f(x) > 0$ and $f'(x) < 0$, it holds that $h_T'(x) < 0$. Therefore, the function $h_T(x)$ is decreasing and consequently we have

$$\sum_{t=T}^{\infty} h_T(t) \leq h_T(T) + \int_T^{\infty} h_T(x)dx.$$

Combining this with $\frac{\eta_t^2}{1+V_{T,t}} \leq \frac{f(t)^2}{1+F_T(t)}$, we have

$$\sum_{t=T}^{\infty} \frac{\eta_t^2}{1+V_{T,t}} \leq h_T(T) + \int_T^{\infty} \frac{f(x)^2}{1+F_T(x)}dx. \tag{108}$$

Given $h_T(T) = f(T)^2 = \eta_T^2 \leq \eta_0^2$, it remains to bound the second term $I_T := \int_T^{\infty} \frac{f(x)^2}{1+F_T(x)}dx$. Given $F_T(x) = \eta_0 \int_T^x (r+1)^{-a}dr = \frac{\eta_)}{1-a}[(x+1)^{1-a} - (T+1)^{1-a}]$, it holds that

$$I_T = \int_T^{\infty} \frac{\eta_0^2(x+1)^{-2a}}{1 + \frac{\eta_0}{1-a}[(x+1)^{1-a} - (T+1)^{1-a}]}dx$$

Define $A_T := (T+1)^{1-a}$ and $y := (x+1)^{1-a} - A_T$. Then, it holds that $y \geq 0$, $(x+1)^{1-a} = A_T + y$, and $dx = \frac{(x+1)^a}{1-a}dy$. Consequently, we have

$$(x+1)^{-2a}dx = \frac{1}{1-a}(x+1)^{-a}dy = \frac{1}{1-a}(A_T+y)^{-a/(1-a)}dy.$$

Setting $p := \frac{a}{1-a}$ and $c := \frac{\eta_0}{1-a}$ and substituting these results into the expression of I_T, we obtain

$$I_T = \frac{\eta_0^2}{1-a}\int_0^{\infty} \frac{(A_T+y)^{-p}}{1+cy}dy.$$

Because $A_T = (T+1)^{1-a} \geq 1$, we have $A_T + y \geq 1 + y$. Since $p > 0$, we further have $(A_T+y)^{-p} \leq (1+y)^{-p}$. Therefore, it holds that

$$I_T \leq \frac{\eta_0^2}{1-a}\int_0^{\infty} \frac{(1+y)^{-p}}{1+cy}dy.$$

Define $m := \min\{1, c\} = \min\{1, \frac{\eta_0}{1-a}\}$. It holds that $1 + cy \geq m(1+y)$ as $y \geq 0$. Therefore, we have

$$I_T \leq \frac{\eta_0^2}{(1-a)m}\int_0^{\infty}(1+y)^{-(p+1)}dy = \frac{\eta_0^2}{(1-a)mp} = \frac{\eta_0^2}{a\min\{1, \frac{\eta_0}{1-a}\}}. \tag{109}$$

Combining (108) and (109) with the bound $h_T(T) \leq \eta_0^2$, we obtain $\sum_{t=T}^{\infty} \frac{\eta_t^2}{1+V_{T,t}} \leq \eta_0^2 + \frac{\eta_0^2}{a\min\{1, \frac{\eta_0}{1-a}\}}$. Because this holds for every $T \geq 0$, we conclude that

$$\sup_{T\geq 0} \sum_{t=T}^{\infty} \frac{\eta_t^2}{1+\sum_{s=T}^{t-1}\eta_s} \leq \eta_0^2 + \frac{\eta_0^2}{a\min\{1, \frac{\eta_0}{1-a}\}} = \eta_0^2 + \frac{1}{a}\max\{\eta_0^2, \eta_0(1-a)\} = Q_\eta.$$

Therefore, the constant Q_η works for both cases. □

Before we present the proof of Proposition 3.1, we first present a short sketch highlighting the main ideas. Let $\tau := T_1^\delta(\epsilon)$, $V_t := V_t^\delta(\epsilon) = \sum_{s=\tau}^{t-1}\eta_s$, and $H_t := H_t^\delta(\epsilon) = \sum_{s=\tau}^{t-1}\theta_s$.

Proof Sketch of Proposition 3.1 The proof of Theorem 4 gives

$$c_\delta \frac{\eta_s}{\epsilon} \le \frac{1}{L_{s+1}} - \frac{1}{L_s} \le 2A_\delta \frac{\eta_s}{\epsilon}, \qquad s \ge \tau.$$

Summing from $s = \tau$ to $t-1$ yields $\frac{\epsilon}{L_t} = \frac{\epsilon}{L_\tau} + \Theta_\delta(V_t)$. Next, we show $\frac{\epsilon}{L_\tau} \asymp_\delta 1$. This can be seen from $L_\tau = \delta\epsilon(1+o(1))$. The proof below provides a more explicit bound. Consequently, we have $\frac{\epsilon}{L_t} \asymp_\delta 1 + V_t$, and $L_t \asymp_\delta \frac{\epsilon}{1+V_t}$. Since Proposition 2.1 gives $\frac{1}{2}L_t \le G_t \le L_t$, we also obtain $G_t \asymp_\delta \frac{\epsilon}{1+V_t}$.

Next, Lemma 12 gives $\gamma_2 G_s \le \|g_s\|_2 \le R_2 G_s$. Consequently, we have $\theta_s = \frac{\eta_s\|g_s\|_2}{\epsilon} \asymp_\delta \frac{\eta_s}{1+V_s}$. Because $V_{s+1} = V_s + \eta_s$, setting $x_s := \frac{\eta_s}{1+V_s} \le \eta_0$ gives

$$\log(1+V_{s+1}) - \log(1+V_s) = \log(1+x_s) \asymp x_s.$$

Summing this telescoping relation proves $H_t \asymp_\delta \log(1+V_t)$.

Recall that $e_t := \frac{\epsilon D_t - g_t}{\|g_t\|_2}$ and $\theta_t := \frac{\eta_t\|g_t\|_2}{\epsilon}$. Therefore, it holds that $\theta_t\|e_t\|_2 = \frac{\eta_t}{\epsilon}\|\epsilon D_t - g_t\|_2$. We decompose

$$\epsilon D_t - g_t = \epsilon\big(D_t - \sigma_\epsilon(g_t)\big) + \big(\epsilon\sigma_\epsilon(g_t) - g_t\big).$$

Based on Lemma 17 and Lemma 7, we can show

$$\theta_t\|e_t\|_2 \le C_{\mathrm{ema}}\eta_t^2\frac{G_t}{\epsilon} + R_1R_2\eta_t\left(\frac{G_t}{\epsilon}\right)^2. \tag{110}$$

The proof of Theorem 4 establishes

$$c_\delta\frac{\eta_t}{\epsilon}L_t^2 \le L_t - L_{t+1}, \qquad t \ge \tau.$$

Based on this, we prove

$$\frac{G_t}{\epsilon} \le \frac{C_L}{1+V_t}, \quad \text{and} \quad \eta_t\left(\frac{G_t}{\epsilon}\right)^2 \le \frac{L_t - L_{t+1}}{c_*\epsilon}.$$

The first inequality is used in Lemma 24 to yield $\sum_{t=\tau}^\infty \eta_t^2\frac{G_t}{\epsilon} \le C_L Q_\eta$. For the second inequality, we obtain $\sum_{t=\tau}^\infty \eta_t\left(\frac{G_t}{\epsilon}\right)^2 \le \frac{L_\tau}{c_*\epsilon} \le \frac{2}{c_*}$ after telescoping and using $L_\tau \le 2\delta\epsilon \le 2\epsilon$. Substituting these two bounds into (110) gives

$$\sum_{t=\tau}^\infty \theta_t\|e_t\|_2 \le C_{\mathrm{ema}}C_L Q_\eta + \frac{2R_1R_2}{c_*} =: E_0.$$

Proof of Proposition 3.1. Let $\tau := T_1^\delta(\epsilon)$, $V_t := V_t^\delta(\epsilon) = \sum_{s=\tau}^{t-1}\eta_s$, and $H_t := H_t^\delta(\epsilon) = \sum_{s=\tau}^{t-1}\theta_s$. From the argument following (104), we have for $t \ge \tau$ that

$$c_\delta\frac{\eta_t}{\epsilon} \le \frac{1}{L_{t+1}} - \frac{1}{L_t} \le 2A_\delta\frac{\eta_t}{\epsilon},$$

where $c_\delta = \frac{\gamma_2^2}{16(1+2R_1\delta)}$ and $A_\delta = R_2^2 + \frac{a_\delta}{2} = R_2^2 + \frac{\gamma_2^2}{2(1+2R_1\delta)}$. Summing from $s=\tau$ to $t-1$ gives

$$c_\delta\frac{V_t}{\epsilon} \le \frac{1}{L_t} - \frac{1}{L_\tau} \le 2A_\delta\frac{V_t}{\epsilon}. \tag{111}$$

Let $K_0 := R_1C_D$. Define $\overline{\eta}^{\mathrm{entry}} := \min\left\{\eta_0, \frac{\log 2}{K_0}\right\}$ and $N^{\mathrm{entry}} := \max\left\{T_*, \left\lceil\left(\frac{\eta_0}{\overline{\eta}^{\mathrm{entry}}}\right)^{1/a}\right\rceil\right\}$. Then, it holds that $\eta_{N^{\mathrm{entry}}} = \frac{\eta_0}{\left(N^{\mathrm{entry}}+1\right)^a} \le \overline{\eta}^{\mathrm{entry}} \le \frac{\log 2}{K_0}$ or equivalently $K_0\eta_{N^{\mathrm{entry}}} \le \log 2$. By Lemma 12, we have $G_t \ge G_0e^{-K_0S_t} \ge G_0e^{-K_0S_{N^{\mathrm{entry}}}} =: c_{\mathrm{entry}} > 0$ for every $t \le N^{\mathrm{entry}}$. We now define $\epsilon_\delta^{\mathrm{entry}} := \min\left\{1, \frac{c_{\mathrm{entry}}}{2\delta}\right\} = \min\left\{1, \frac{G_0e^{-K_0S_{N^{\mathrm{entry}}}}}{2\delta}\right\}$. Suppose that $0 < \epsilon \le \epsilon_\delta^{\mathrm{entry}}$. Then, we have

$$\delta\epsilon \le \frac{c_{\mathrm{entry}}}{2} < c_{\mathrm{entry}} \le G_t, \qquad 0 \le t \le N^{\mathrm{entry}}.$$

Consequently, by the definition of τ, we must have $\tau > N^{\mathrm{entry}}$, from which we obtain $\eta_{\tau-1} \leq \eta_{N^{\mathrm{entry}}} \leq \frac{\log 2}{K_0}$. Consequently, it holds that $e^{-K_0\eta_{\tau-1}} \geq e^{-\log 2} = \frac{1}{2}$. By the minimality of τ, we have $G_{\tau-1} > \delta\epsilon$ and $G_\tau \leq \delta\epsilon$. Applying Lemma 12 with $r = \tau - 1$ and $t = \tau$ gives

$$G_\tau \geq e^{-K_0\eta_{\tau-1}} G_{\tau-1} \geq e^{-K_0\eta_{\tau-1}}\delta\epsilon \geq \frac{\delta\epsilon}{2}.$$

Together with the defining upper bound, we obtain $\frac{\delta\epsilon}{2} < G_\tau \leq \delta\epsilon$. Since $\tau \geq T_*$, Proposition 2.1 gives $\frac{1}{2}L_\tau \leq G_\tau \leq L_\tau$. Combining them together, we further obtain that $\frac{\delta\epsilon}{2} < L_\tau \leq 2\delta\epsilon$. Consequently, it holds that $\frac{\epsilon}{L_\tau} \asymp_\delta 1$. Therefore, by multiplying (111) by ϵ, we obtain

$$\frac{\epsilon}{L_t} = \frac{\epsilon}{L_\tau} + \Theta_\delta(V_t) \asymp_\delta 1 + V_t.$$

From this, we conclude that $L_t \asymp_\delta \frac{\epsilon}{1+V_t}$. By Proposition 2.1, we have $\frac{1}{2}L_t \leq G_t \leq L_t$ when $t \geq T_*$. This immediately gives $G_t \asymp_\delta \frac{\epsilon}{1+V_t}$ when $t \geq \tau$ (note that $\tau \geq T_*$).

By Lemma 7, it holds that $\gamma_2 G_s \leq \|g_s\|_2 \leq R_2 G_s$. Therefore, we have

$$\theta_s = \frac{\eta_s\|g_s\|_2}{\epsilon} \asymp \eta_s\frac{G_s}{\epsilon} \asymp_\delta \frac{\eta_s}{1+V_s}.$$

This implies that $H_t \asymp_\delta \sum_{s=\tau}^{t-1} \frac{\eta_s}{1+V_s}$. It remains to compare $\sum_{s=\tau}^{t-1} \frac{\eta_s}{1+V_s}$ with $\log(1+V_t)$. Since $V_{s+1} = V_s + \eta_s$, we have

$$\log(1+V_{s+1}) - \log(1+V_s) = \log(1 + \frac{\eta_s}{1+V_s}) = \log(1+x_s), \tag{112}$$

where we have let $x_s := \frac{\eta_s}{1+V_s}$. Note that $0 \leq x_s \leq \eta_s \leq \eta_0$. For $0 \leq u \leq x$, it holds that $\frac{1}{1+x} \leq \frac{1}{1+u} \leq 1$. Integrating from 0 to x gives $\frac{x}{1+x} \leq \log(1+x) \leq x$. Applying this with $x = x_s$ and using $x_s \leq \eta_0$, we obtain

$$\frac{x_s}{1+\eta_0} \leq \frac{x_s}{1+x_s} \leq \log(1+x_s) \leq x_s,$$

or equivalently,

$$\log(1+x_s) \leq x_s \leq (1+\eta_0)\log(1+x_s).$$

Substituting $x_s = \frac{\eta_s}{1+V_s}$ and using (112), it holds that

$$\log(1+V_{s+1}) - \log(1+V_s) \leq \frac{\eta_s}{1+V_s} \leq (1+\eta_0)[\log(1+V_{s+1}) - \log(1+V_s)],$$

from which we conclude that $\frac{\eta_s}{1+V_s} \asymp \log(1+V_{s+1}) - \log(1+V_s)$. Given $\sum_{s=\tau}^{t-1}[\log(1+V_{s+1}) - \log(1+V_s)] = \log(1+V_t)$, it holds that $\sum_{s=\tau}^{t-1} \frac{\eta_s}{1+V_s} \asymp \log(1+V_t)$. Therefore, we conclude that $H_t \asymp_\delta \log(1+V_t)$.

Next, we show $\sum_{t=\tau}^{\infty} \theta_t\|e_t\|2 \leq E_0$ for some constant E_0. Recall some definitions: $\theta_t = \frac{\eta_t\|g_t\|_2}{\epsilon}$ and $e_t = \frac{\epsilon D_t - g_t}{\|g_t\|_2}$. We first provide an upper bound on $\theta_t\|e_t\|_2$. By noting $\theta_t\|e_t\|_2 = \frac{\eta_t}{\epsilon}\|\epsilon D_t - g_t\|_2$, it holds that

$$\theta_t\|e_t\|_2 \leq \frac{\eta_t}{\epsilon}[\epsilon\|D_t - \sigma_\epsilon(g_t)\|_2 + \|\epsilon\sigma_\epsilon(g_t) - g_t\|_2], \tag{113}$$

where we used the decomposition $\epsilon D_t - g_t = \epsilon(D_t - \sigma_\epsilon(g_t)) + (\epsilon\sigma_\epsilon(g_t) - g_t)$. By Lemma 17, it holds that $\epsilon\|D_t - \sigma_\epsilon(g_t)\|_2 \leq C_{\mathrm{ema}}\eta_t G_t, t \geq t_{\mathrm{ema}}$. For the second term, we have for each coordinate that $\epsilon\sigma_\epsilon(g_t)[j] - g_t[j] = -g_t[j]\frac{|g_t[j]|}{|g_t[j]|+\epsilon}$. Since $\frac{|g_t[j]|}{|g_t[j]|+\epsilon} \leq \frac{\|g_t\|_\infty}{\epsilon}$, we obtain

$$\|\epsilon\sigma_\epsilon(g_t) - g_t\|_2 \leq \frac{\|g_t\|_\infty\|g_t\|_2}{\epsilon} \leq R_1R_2\frac{G_t^2}{\epsilon}.$$

Substituting this into (113), we obtain

$$\theta_t\|e_t\|_2 \leq C_{\mathrm{ema}}\eta_t^2\frac{G_t}{\epsilon} + R_1R_2\eta_t(\frac{G_t}{\epsilon})^2. \tag{114}$$

Next, we bound $\sum_{t=\tau}^{\infty} \eta_t^2 \frac{G_t}{\epsilon}$ and $\sum_{t=\tau}^{\infty} \eta_t (\frac{G_t}{\epsilon})^2$. To do so, we first control $\frac{G_t}{\epsilon}$. Part 3 of Theorem 4 establishes

$$c_\delta \frac{\eta_t}{\epsilon} L_t^2 \le L_t - L_{t+1}, \qquad t \ge \tau, \tag{115}$$

where $c_\delta = \frac{\gamma_2^2}{16(1+2R_1\delta)}$. Because $0 < \delta < 1$, it holds that $c_\delta \ge c_* := \frac{\gamma_2^2}{16(1+2R_1)} > 0$. Since $L_{t+1} \le L_t$, we have

$$\frac{1}{L_{t+1}} - \frac{1}{L_t} = \frac{L_t - L_{t+1}}{L_t L_{t+1}} \ge \frac{L_t - L_{t+1}}{L_t^2} \ge c_* \frac{\eta_t}{\epsilon}. \tag{116}$$

Summing from τ to $t-1$ leads to $\frac{1}{L_t} \ge \frac{1}{L_\tau} + c_* \frac{V_t}{\epsilon}$. Therefore, we have $\frac{L_t}{\epsilon} \le \frac{1}{\epsilon/L_\tau + c_* V_t}$. Since $L_\tau \le 2\delta\epsilon \le 2\epsilon$, it holds that $\frac{\epsilon}{L_\tau} \ge \frac{1}{2}$. Therefore, we have

$$\frac{G_t}{\epsilon} \le \frac{L_t}{\epsilon} \le \frac{1}{1/2 + c_* V_t} \le \frac{C_L}{1+V_t}, \tag{117}$$

where $C_L := \frac{1}{\min\{1/2, c_*\}}$. Next, we apply Lemma 24 to obtain

$$\sum_{t=\tau}^{\infty} \eta_t^2 \frac{G_t}{\epsilon} \le C_L \sum_{t=\tau}^{\infty} \frac{\eta_t^2}{1+V_t} \le C_L Q_\eta = C_L(\eta_0^2 + \frac{1}{a}\max\{\eta_0^2, \eta_0(1-a)\}). \tag{118}$$

From (115) and $G_t \le L_t$, it holds that

$$L_t - L_{t+1} \ge c_* \frac{\eta_t}{\epsilon} L_t^2 \ge c_* \frac{\eta_t}{\epsilon} G_t^2.$$

Therefore, we have $\eta_t (\frac{G_t}{\epsilon})^2 \le \frac{L_t - L_{t+1}}{c_* \epsilon}$. Summing this leads to

$$\sum_{t=\tau}^{\infty} \eta_t (\frac{G_t}{\epsilon})^2 \le \frac{1}{c_*\epsilon} \sum_{t=\tau}^{\infty} (L_t - L_{t+1}) \le \frac{L_\tau}{c_*\epsilon} \le \frac{2\delta}{c_*} \le \frac{2}{c_*}. \tag{119}$$

Summing (114) and applying (118) and (119), we obtain

$$\sum_{t=\tau}^{\infty} \theta_t \|e_t\|_2 \le C_{\text{ema}} \sum_{t=\tau}^{\infty} \eta_t^2 \frac{G_t}{\epsilon} + R_1 R_2 \sum_{t=\tau}^{\infty} \eta_t (\frac{G_t}{\epsilon})^2 \le C_{\text{ema}} C_L Q_\eta + \frac{2R_1 R_2}{c_*}.$$

Therefore, we can take

$$E_0 = C_{\text{ema}} C_L Q_\eta + \frac{2R_1 R_2}{c_*}, \qquad c_* = \frac{\gamma_2^2}{16(1+2R_1)}.$$

Finally, we summarize the conditions on ϵ. For (111) to hold, we require three conditions: $C_{\text{ema}} R_2 \eta_\tau \le \frac{a_\delta}{2}$, $2\delta B_\delta \eta_\tau \le \frac{a_\delta}{4}$, and $2\delta A_\delta \eta_\tau \le \frac{1}{2}$. The constants A_δ and B_δ are given in Theorem 4. The first two conditions ensure (104) holds. They can be satisfied by requiring $0 < \epsilon \le \epsilon_\delta^{\text{inv}}$, where $\epsilon_\delta^{\text{inv}}$ is given in Corollary 22. The last condition is for ensuring $L_{t+1} \ge \frac{1}{2} L_t$. Define $\overline{\eta}_\delta^{\text{rec}} := \min\left\{\eta_0, \frac{1}{4\delta A_\delta}\right\}$, and $N_\delta^{\text{rec}} := \max\left\{T_*, t_{\text{ema}}, \left\lceil \left(\frac{\eta_0}{\overline{\eta}_\delta^{\text{rec}}}\right)^{1/a} \right\rceil\right\}$. Choose $\epsilon_\delta^{\text{rec}} := \min\left\{1, \frac{G_0 e^{-K_0 S_{N_\delta^{\text{rec}}}}}{2\delta}\right\}$. If $0 < \epsilon \le \epsilon_\delta^{\text{rec}}$, then we have $G_t \ge G_0 e^{-K_0 S_{N_\delta^{\text{rec}}}} > \delta\epsilon$ when $0 \le t \le N_\delta^{\text{rec}}$. Therefore, it holds that $\tau > N_\delta^{\text{rec}}$ and $\eta_\tau \le \eta_{N_\delta^{\text{rec}}} \le \overline{\eta}_\delta^{\text{rec}} \le \frac{1}{4\delta A_\delta}$. Consequently, (105) gives $\frac{L_t - L_{t+1}}{L_t} \le 2\delta A_\delta \eta_\tau \le \frac{1}{2}$, which implies $L_{t+1} \ge \frac{1}{2} L_t$. In summary, we can let $\widetilde{\epsilon}_1(\delta) := \min\left\{\epsilon_\delta^{\text{inv}}, \epsilon_\delta^{\text{rec}}, \epsilon_\delta^{\text{entry}}\right\}$. □

F ADAM'S ℓ_∞-MARGIN OPTIMALITY

Throughout this section, we assume data separability, $0 \le \beta_1 \le \beta_2 < 1$, and $\epsilon \in (0, 1]$. Recall that $R_1 := \max_{i\in[n]} \|x_i\|_1$ and $R_2 := \max_{i\in[n]} \|x_i\|_2$. The step size schedule is $\eta_t = \eta_0 (t+1)^{-a}$ with $\eta_0 > 0$ and $a \in (1/3, 1]$. We first present some useful lemmas. They include Hoffman's error bound in Lemma 26 and the margin difference bound in Lemma 28. The margin gaps at $T_1^-(\epsilon; b)$, $T_1^\delta(\epsilon)$, and $T_1^+(\epsilon; b)$ are proved in Theorem 31, Theorem 29, and Corollary 32, respectively.

Lemma 25. *For any two nonzero vectors x and u, it holds that*

$$\left|\frac{1}{\|x\|_2}-\frac{1}{\|u\|_2}\right| \le \frac{\|x-u\|_2}{\|x\|_2\|u\|_2}.$$

Proof. It holds that

$$\left|\frac{1}{\|x\|_2}-\frac{1}{\|u\|_2}\right| = \left|\frac{\|u\|_2-\|x\|_2}{\|x\|_2\|u\|_2}\right| = \frac{|\|u\|_2-\|x\|_2|}{\|x\|_2\|u\|_2}$$

The conclusion follows from the reverse triangle inequality $|\|u\|_2-\|x\|_2| \le \|u-x\|_2$. □

Lemma 26 ((Hoffman, 1952; Peña et al., 2021))**.** *Define*

$$A_{\mathrm{H}} := \begin{pmatrix} I_d \\ -I_d \\ -Z \end{pmatrix} \in \mathbb{R}^{(2d+n)\times d}, \qquad c_{\mathrm{H}} := \begin{pmatrix} \mathbf{1}_d \\ \mathbf{1}_d \\ -\gamma_\infty \mathbf{1}_n \end{pmatrix} \in \mathbb{R}^{2d+n},$$

where the i-th row of Z is z_i^T. There exists a finite constant $H_{2,\infty}(A_H)$ (depending only on the matrix A_H) such that

$$\mathrm{dist}_2(x, \mathcal{M}_\infty) \le H_{2,\infty}(A_{\mathrm{H}}) \left\|(A_{\mathrm{H}}x - c_{\mathrm{H}})_+\right\|_\infty, \tag{120}$$

where $\mathrm{dist}_2(x,\mathcal{M}_\infty) := \min_{u\in\mathcal{M}_\infty}\|x-u\|_2$ *and* $[y]_+ := (\max\{y_k, 0\})_k$.

Proof. Recall that $\mathcal{M}_\infty = \{u \in \mathbb{R}^d : \|u\|_\infty \le 1, \rho(u) = \gamma_\infty\}$. We prove that $\mathcal{M}_\infty = \{u \in \mathbb{R}^d : A_H u \le c_H\}$. We first show that $\mathcal{M}_\infty \subseteq \{u : A_H u \le c_H\}$. For any $u \in \mathcal{M}_\infty$, it holds that $\|u\|_\infty \le 1$. Consequently, we have $u \le \mathbf{1}_d$ and $-u \le \mathbf{1}_d$. Moreover, it holds that $z_i^T u \ge \gamma_\infty, \forall i$ given $\rho(u) = \min_i z_i^T u = \gamma_\infty$. Therefore, we have $Zu \ge \gamma_\infty \mathbf{1}_n$ or equivalently $-Zu \le -\gamma_\infty \mathbf{1}_n$. Hence, all the block inequalities hold and we conclude that $A_H u \le c_H$. Next, we show $\{u : A_H u \le c_H\} \subseteq \mathcal{M}_\infty$. For any u satisfying $A_H u \le c_H$, the first two blocks imply $\|u\|_\infty \le 1$, and the third block implies $z_i^T u \ge \gamma_\infty, \forall i$ (consequently, it holds that $\rho(u) = \min_i z_i^T u \ge \gamma_\infty$). Since $\|u\|_\infty \le 1$, the definition of γ_∞ implies that $\rho(u) \le \gamma_\infty$. Therefore, we have $\rho(u) = \gamma_\infty$ and $u \in \mathcal{M}_\infty$. Since $[-1,1]^d$ is compact and ρ is continuous, the maximum defining γ_∞ is attained. Hence, $\mathcal{M}_\infty$ is nonempty. Therefore, the system $A_H u \le c_H$ is feasible. By applying Peña et al. (2021, Proposition 1 and 2), we conclude that (120) holds with the constant

$$H_{2,\infty}(A_{\mathrm{H}}) := \max_{\substack{\varnothing\neq J\subseteq[2d+n] \\ \mathrm{rank}((A_{\mathrm{H}})_J)=|J|}} \left[\min_{\substack{v\in\mathbb{R}_+^{|J|} \\ \mathbf{1}^\top v=1}} \left\|(A_{\mathrm{H}})_J^\top v\right\|_2\right]^{-1},$$

where $(A_H)_J$ denotes the submatrix formed by the rows of A_H indexed by J. Denote $\sigma_{\min}\big((A_H)_J\big) := \sqrt{\lambda_{\min}\big((A_H)_J(A_H)_J^\top\big)}$. Let $A := A_H$. The inner quantity $\alpha_J := \min_{v\ge 0,\,\|v\|_1=1}\|A_J^T v\|_2$ is strictly positive whenever A_J has full row rank. Since there are finitely many row subsets J, the constant $H_{2,\infty}(A_H)$ is finite. From the definition of $H_{2,\infty}(A)$, we can derive a simple upper bound. For every full-row-rank A_J, it holds that $\|A_J^T v\|_2 \ge \sigma_{\min}(A_J)\|v\|_2$. If $v \ge 0$ and $\|v\|_1 = 1$, then we have $\|v\|_2 \ge \frac{1}{\sqrt{|J|}}$. Therefore, it holds that $\alpha_J \ge \frac{\sigma_{\min}(A_J)}{\sqrt{|J|}}$ and consequently

$$H_{2,\infty}(A) \le \max_{\substack{\varnothing\neq J\subseteq[2d+n] \\ \mathrm{rank}(A_J)=|J|}} \frac{\sqrt{|J|}}{\sigma_{\min}(A_J)}.$$

□

Lemma 27. *Recall that $\rho(u) := \min_{i\in[n]} z_i^\top u$. It holds that the function ρ is R_1-Lipschitz in max-norm and R_2-Lipschitz in 2-norm:*

$$|\rho(u)-\rho(v)| \le R_1\|u-v\|_\infty, \qquad |\rho(u)-\rho(v)| \le R_2\|u-v\|_2.$$

Proof. Choose i_v such that $\rho(v) = z_{i_v}^T v$. By definition, it holds that $\rho(u) \le z_{i_v}^T u$. Therefore, we have

$$\begin{aligned}\rho(u) - \rho(v) &\le z_{i_v}^T u - z_{i_v}^T v = z_{i_v}^T (u - v) \\ &\le |z_{i_v}^T (u - v)| \le \|z_{i_v}\|_1 \|u - v\|_\infty \le R_1 \|u - v\|_\infty\end{aligned}$$

Similarly, we also have

$$\rho(u) - \rho(v) \le \|z_{i_v}\|_2 \|u - v\|_2 \le R_2 \|u - v\|_2.$$

Interchanging u and v gives the corresponding lower bounds. □

Lemma 28. *For any nonzero vectors x and y, it holds that*

$$|\hat{\gamma}_\infty(x) - \hat{\gamma}_\infty(y)| \le \frac{2R_1\|x - y\|_\infty}{\|y\|_\infty}, \qquad |\hat{\gamma}_2(x) - \hat{\gamma}_2(y)| \le \frac{2R_2\|x - y\|_2}{\|y\|_2}.$$

Proof. Let $R_x := \|x\|_\infty$ and $R_y := \|y\|_\infty$. It holds that

$$\frac{\rho(x)}{R_x} - \frac{\rho(y)}{R_y} = \frac{\rho(x) - \rho(y)}{R_y} + \rho(x)\left(\frac{1}{R_x} - \frac{1}{R_y}\right). \tag{121}$$

From Lemma 27, we know $|\rho(x) - \rho(y)| \le R_1\|x - y\|_\infty$. Setting $y = 0$ leads to $|\rho(x)| \le R_1\|x\|_\infty = R_1 R_x$. By the reverse triangle inequality, we also have $|R_x - R_y| \le \|x - y\|_\infty$. Substituting these inequalities into (121), we obtain

$$\begin{aligned}\left|\frac{\rho(x)}{R_x} - \frac{\rho(y)}{R_y}\right| &\le \frac{|\rho(x) - \rho(y)|}{R_y} + |\rho(x)|\frac{|R_x - R_y|}{R_x R_y} \\ &\le \frac{R_1\|x - y\|_\infty}{R_y} + R_1 R_x \frac{\|x - y\|_\infty}{R_x R_y} \\ &= \frac{2R_1\|x - y\|_\infty}{R_y}.\end{aligned}$$

Similarly, from Lemma 27, we know $|\rho(x) - \rho(y)| \le R_2\|x - y\|_2$ and $|\rho(x)| \le R_2\|x\|_2$. Consequently, we have

$$|\hat{\gamma}_2(x) - \hat{\gamma}_2(y)| \le \frac{R_2\|x - y\|_2}{\|y\|_2} + R_2\|x\|_2 \frac{\|x - y\|_2}{\|x\|_2\|y\|_2} \le \frac{2R_2\|x - y\|_2}{\|y\|_2}.$$

□

F.1 Margin Gap at T_1

Theorem 29. *Denote $r_\epsilon := \ell_\epsilon^{-a}$ when $a \in (\frac{1}{3}, 1)$ and $r_\epsilon := \frac{\log \ell_\epsilon}{\ell_\epsilon}$ when $a = 1$. For any $\delta \in (0, 1)$, there exists a constant $\tilde{\epsilon}_2(\delta)$ such that for every $0 < \epsilon \le \tilde{\epsilon}_2(\delta)$, we have*

$$\gamma_\infty - \widehat{\gamma}_\infty(w_{T_1^\delta(\epsilon)}) \le O(r_\epsilon), \qquad \gamma_2 - \widehat{\gamma}_2(w_{T_1^\delta(\epsilon)}) \ge \Delta_{\text{geom}} - O(r_\epsilon).$$

Proof of Theorem 29. Recall that $\ell_\epsilon = \log \frac{1}{\epsilon}$ and let $\tau_\epsilon := T_1^\delta(\epsilon) = \inf\{t \ge T_* : G_t \le \delta\epsilon\}$. We further define $\omega_a(l) := l^{1-a}$ when $a \in (1/3, 1)$ and $\omega_a(l) := \log l$ when $a = 1$. We note that $\frac{\omega_a(\ell_\epsilon)}{\ell_\epsilon} = r_\epsilon$ where $r_\epsilon := l_\epsilon^{-a}$ when $a \in (\frac{1}{3}, 1)$ and $r_\epsilon := \frac{\log l_\epsilon}{l_\epsilon}$ when $a = 1$. Theorem 4 already gives $S_{\tau_\epsilon} - S_{T_*} = \frac{\ell_\epsilon}{\gamma_\infty} + o(\ell_\epsilon)$.

Part 1 Define $\Lambda_\epsilon = \ell_\epsilon^a$ when $a \in (1/3, 1)$ and $\frac{\ell_\epsilon}{\log \ell_\epsilon}$ when $a = 1$. Note that $\Lambda_\epsilon \to \infty$ and $\epsilon\Lambda_\epsilon \to 0$ as $\epsilon \downarrow 0$. We also define an auxiliary stopping time $\sigma_\epsilon := \inf\{t \ge T_* : G_t \le \epsilon\Lambda_\epsilon\}$ that is useful for the proof. For sufficiently small ϵ, we have $\Lambda_\epsilon > \delta$ and consequently $\sigma_\epsilon \le \tau_\epsilon$. By the definition of σ_ϵ, it holds for $T_* \le t < \sigma_\epsilon$ that $G_t > \epsilon\Lambda_\epsilon$ or equivalently $\frac{\epsilon}{G_t} < \frac{1}{\Lambda_\epsilon}$. Using (42) and $\|g_t\|_1 \ge \gamma_\infty G_t$, it holds that

$$\langle g_t, \sigma_\epsilon(g_t)\rangle \ge \gamma_\infty G_t - d\epsilon > (\gamma_\infty - \frac{d}{\Lambda_\epsilon})G_t.$$

Substituting this into (44), we have

$$L_{t+1} \le L_t - \eta_t A_t G_t, \qquad A_t := \gamma_\infty - \frac{d}{\Lambda_\epsilon} - C_{\mathrm{md}}\eta_t.$$

Next, we define $\bar{\eta}_0 := \min\{\frac{\gamma_\infty}{4C_{\mathrm{md}}}, \frac{1}{2\gamma_\infty}\}$ and choose $t_0 := \max\{T_*, t_{\mathrm{ema}}, \lceil(\frac{\eta_0}{\bar{\eta}_0})^{1/a}\rceil\}$. Then, we have that $C_{\mathrm{md}}\eta_t \le \frac{\gamma_\infty}{4}$ and $\gamma_\infty\eta_t \le \frac{1}{2}$ when $t \ge t_0$. Given this fixed t_0, Lemma 12 gives $G_t \ge G_0\exp(-R_1C_DS_{t_0}) =: c_{t_0} > 0$ when $t \le t_0$. We choose ϵ sufficiently small such that $\epsilon\Lambda_\epsilon < c_{t_0}$. This would ensure that $G_t > \epsilon\Lambda_\epsilon$ for $0 \le t \le t_0$ and consequently we have $\sigma_\epsilon > t_0$. We also choose ϵ sufficiently small such that $\frac{d}{\Lambda_\epsilon} \le \frac{\gamma_\infty}{4}$. With the choice of t_0 and ϵ, we obtain for every $t_0 \le t < \sigma_\epsilon$

$$\gamma_\infty \ge A_t = \gamma_\infty - \frac{d}{\Lambda_\epsilon} - C_{\mathrm{md}}\eta_t \ge \gamma_\infty - \frac{\gamma_\infty}{4} - \frac{\gamma_\infty}{4} = \frac{\gamma_\infty}{2}.$$

Since $G_t/L_t \le 1$, we also have $0 \le \eta_t A_t \frac{G_t}{L_t} \le \gamma_\infty\eta_t \le \frac{1}{2}$ for $t_0 \le t < \sigma_\epsilon$. Let $y_t := \eta_t A_t \frac{G_t}{L_t}$, we then have $0 \le y_t \le \frac{1}{2}$ and $L_{t+1} \le L_t - L_t y_t = L_t(1-y_t)$. This implies that

$$\log\frac{L_t}{L_{t+1}} \ge -\log(1-y_t) \ge y_t = \eta_t A_t \frac{G_t}{L_t}, \qquad t_0 \le t < \sigma_\epsilon,$$

where we used $-\log(1-y) \ge y$ for $0 \le y < 1$. From $\frac{G_t}{L_t} \ge 1 - \frac{nL_t}{2}$, we have

$$\begin{aligned} A_t\frac{G_t}{L_t} &\ge A_t(1-\frac{nL_t}{2}) \\ &= \gamma_\infty - \frac{d}{\Lambda_\epsilon} - C_{\mathrm{md}}\eta_t - \frac{n\gamma_\infty}{2}L_t + \frac{nd}{2\Lambda_\epsilon}L_t + \frac{nC_{\mathrm{md}}}{2}\eta_t L_t \\ &\ge \gamma_\infty - \frac{d}{\Lambda_\epsilon} - C_{\mathrm{md}}\eta_t - \frac{n\gamma_\infty}{2}L_t. \end{aligned}$$

Therefore, it holds that

$$\log\frac{L_t}{L_{t+1}} \ge \gamma_\infty\eta_t - \frac{d}{\Lambda_\epsilon}\eta_t - C_{\mathrm{md}}\eta_t^2 - \frac{n\gamma_\infty}{2}\eta_t L_t, \qquad t_0 \le t < \sigma_\epsilon. \tag{122}$$

Summing (122) from t_0 to $\sigma_\epsilon - 1$, we obtain

$$\log\frac{L_{t_0}}{L_{\sigma_\epsilon}} \ge \gamma_\infty(S_{\sigma_\epsilon} - S_{t_0}) - \frac{d}{\Lambda_\epsilon}(S_{\sigma_\epsilon} - S_{t_0}) - C_{\mathrm{md}}Q_{\sigma_\epsilon;t_0} - \frac{n\gamma_\infty}{2}\sum_{t=t_0}^{\sigma_\epsilon-1}\eta_t L_t \tag{123}$$

Next, we bound the term $\sum_{t=t_0}^{\sigma_\epsilon-1}\eta_t L_t$. Starting from $L_{t+1} \le L_t - \eta_t A_t G_t$, we obtain

$$L_t - L_{t+1} \ge \eta_t A_t G_t \ge \frac{\gamma_\infty}{2}\eta_t G_t \ge \frac{\gamma_\infty}{4}\eta_t L_t, \tag{124}$$

where we used $A_t \ge \frac{\gamma_\infty}{2}$ and $G_t \ge \frac{1}{2}L_t$ (as $T_* \le t_0 \le t < \sigma_\epsilon$). Rearranging and summing (124) from t_0 to $\sigma_\epsilon - 1$, we obtain

$$\sum_{t=t_0}^{\sigma_\epsilon-1}\eta_t L_t \le \frac{4}{\gamma_\infty}\sum_{t=t_0}^{\sigma_\epsilon-1}(L_t - L_{t+1}) = \frac{4}{\gamma_\infty}(L_{t_0} - L_{\sigma_\epsilon}).$$

Since $L_{\sigma_\epsilon} > 0$, it holds that $\sum_{t=t_0}^{\sigma_\epsilon-1}\eta_t L_t \le \frac{4L_{t_0}}{\gamma_\infty} \le \frac{4\log 2}{n\gamma_\infty}$ (as $L_{t_0} \le \frac{\log 2}{n}$ when $t_0 \ge T_*$).

Denote $K_0 := R_1C_D$. We have shown that $\sigma_\epsilon > t_0$. By minimality of σ_ϵ, it holds that $G_{\sigma_\epsilon-1} > \epsilon\Lambda_\epsilon$. From Lemma 12, we obtain

$$L_{\sigma_\epsilon} \ge G_{\sigma_\epsilon} \ge e^{-K_0\eta_{\sigma_\epsilon-1}}G_{\sigma_\epsilon-1} \ge \epsilon\Lambda_\epsilon e^{-K_0\eta_{\sigma_\epsilon-1}} \ge \epsilon\Lambda_\epsilon e^{-K_0\eta_{t_0}}.$$

Consequently, it holds that $\frac{L_{t_0}}{L_{\sigma_\epsilon}} < \frac{L_{t_0}e^{K_0\eta_{t_0}}}{\epsilon\Lambda_\epsilon}$. After taking logarithms, we obtain

$$\begin{aligned} \log\frac{L_{t_0}}{L_{\sigma_\epsilon}} &\le \log(\frac{L_{t_0}e^{K_0\eta_{t_0}}}{\epsilon\Lambda_\epsilon}) = \ell_\epsilon - \log\Lambda_\epsilon + \log L_{t_0} + K_0\eta_{t_0} \\ &\le \ell_\epsilon + K_0\eta_{t_0}, \end{aligned}$$

where the second inequality is by $\Lambda_\epsilon \geq 1$ (ensured by choosing ϵ sufficiently small) and $L_{t_0} \leq \frac{\log 2}{n} < 1$. Consequently, we let $C_1 := K_0 \eta_{t_0}$ and it holds that $\log \frac{L_{t_0}}{L_{\sigma_\epsilon}} \leq \ell_\epsilon + C_1$.

By defining $V_\epsilon := S_{\sigma_\epsilon} - S_{t_0}$ and using $\sum_{t=t_0}^{\sigma_\epsilon - 1} \eta_t L_t \leq \frac{4 \log 2}{n \gamma_\infty}$ and $\log \frac{L_{t_0}}{L_{\sigma_\epsilon}} \leq \ell_\epsilon + C_1$, we obtain from (123) that

$$\gamma_\infty V_\epsilon \leq \ell_\epsilon + C_1 + \frac{d}{\Lambda_\epsilon} V_\epsilon + C_{\mathrm{md}} Q_{\sigma_\epsilon; t_0} + 2 \log 2. \tag{125}$$

From $\frac{d}{\Lambda_\epsilon} \leq \frac{\gamma_\infty}{4}$ and $C_{\mathrm{md}} \eta_{t_0} \leq \frac{\gamma_\infty}{4}$, we further obtain $\frac{\gamma_\infty}{2} V_\epsilon \leq \ell_\epsilon + C_1 + 2 \log 2$. Consequently, it holds that $V_\epsilon \leq \frac{2}{\gamma_\infty}(\ell_\epsilon + C_1 + 2 \log 2)$. We choose ϵ sufficiently small such that $\ell_\epsilon \geq e > 1$. Since $S_{\sigma_\epsilon} = S_{t_0} + V_\epsilon$ and $\ell_\epsilon \geq 1$, it holds that

$$\begin{aligned} S_{\sigma_\epsilon} &\leq S_{t_0} + \frac{2}{\gamma_\infty}(\ell_\epsilon + C_1 + 2 \log 2) \leq S_{t_0} \ell_\epsilon + \frac{2}{\gamma_\infty} \ell_\epsilon + \frac{2}{\gamma_\infty}(C_1 + 2 \log 2) \ell_\epsilon \\ &= [S_{t_0} + \frac{2}{\gamma_\infty}(1 + C_1 + 2 \log 2)] \ell_\epsilon = A_0 \ell_\epsilon, \end{aligned}$$

where $A_0 := S_{t_0} + \frac{2}{\gamma_\infty}(1 + C_1 + 2 \log 2)$. For $a < 1$, we have

$$S_{\sigma_\epsilon} = \eta_0 \sum_{k=1}^{\sigma_\epsilon} k^{-a} \geq \eta_0 \sum_{k=1}^{\sigma_\epsilon} \sigma_\epsilon^{-a} = \eta_0 \sigma_\epsilon^{1-a}.$$

Combining this with $S_{\sigma_\epsilon} \leq A_0 \ell_\epsilon$, we obtain $\sigma_\epsilon^{1-a} \leq \frac{A_0}{\eta_0} \ell_\epsilon \leq B \ell_\epsilon$ where $B := \max\{1, \frac{A_0}{\eta_0}\}$. The condition $B \geq 1$ ensures that $\log B \geq 0$.

By the definition of $Q_{\sigma_\epsilon; t_0}$, we have

$$Q_{\sigma_\epsilon; t_0} = \eta_0^2 \sum_{t=t_0}^{\sigma_\epsilon - 1} (t+1)^{-2a} \leq \eta_0^2 \sum_{t=0}^{\sigma_\epsilon - 1} (t+1)^{-2a} = \eta_0^2 \sum_{k=1}^{\sigma_\epsilon} k^{-2a}.$$

In the case where $1/3 < a < 1/2$, we have

$$\sum_{k=1}^{\sigma_\epsilon} k^{-2a} \leq 1 + \int_1^{\sigma_\epsilon} x^{-2a} dx = \frac{\sigma_\epsilon^{1-2a} - 2a}{1 - 2a} \leq \frac{\sigma_\epsilon^{1-2a}}{1 - 2a}.$$

Consequently, it holds that $Q_{\sigma_\epsilon; t_0} \leq \frac{\eta_0^2}{1-2a} \sigma_\epsilon^{1-2a}$. Define $p_a := \frac{1-2a}{1-a}$. It holds that $1 - 2a = p_a(1-a)$ and $p_a > 0$ given $a < \frac{1}{2}$. Therefore, we have $\sigma_\epsilon^{1-2a} = (\sigma_\epsilon^{1-a})^{p_a} \leq (B \ell_\epsilon)^{p_a} = B^{p_a} \ell_\epsilon^{p_a}$. Moreover, it holds that $(1-a) - p_a = \frac{a^2}{1-a} > 0$ and therefore $p_a < 1 - a$. This implies that $\ell_\epsilon^{p_a} \leq \ell_\epsilon^{1-a}$ given $\ell_\epsilon \geq 1$ and $Q_{\sigma_\epsilon; t_0} \leq \frac{\eta_0^2}{1-2a} B^{p_a} \ell_\epsilon^{1-a}$. In the case where $a = \frac{1}{2}$, we have

$$Q_{\sigma_\epsilon; t_0} \leq \eta_0^2 \sum_{k=1}^{\sigma_\epsilon} \frac{1}{k} \leq \eta_0^2 (1 + \log \sigma_\epsilon).$$

From $\sigma_\epsilon^{1-a} \leq B \ell_\epsilon$ when $a < 1$, we obtain $\sigma_\epsilon^{1/2} \leq B \ell_\epsilon$ and consequently $\sigma_\epsilon \leq B^2 \ell_\epsilon^2$. This implies that $\log \sigma_\epsilon \leq \log(B^2 \ell_\epsilon^2) = 2 \log B + 2 \log \ell_\epsilon$. Therefore, we have $Q_{\sigma_\epsilon; t_0} \leq \eta_0^2 (1 + 2 \log B + 2 \log \ell_\epsilon)$. Given $\ell_\epsilon \geq 1$, we have $1 \leq \sqrt{\ell_\epsilon}$ and $\log \ell_\epsilon \leq \sqrt{\ell_\epsilon}$. Therefore, it holds that

$$1 + 2 \log B + 2 \log \ell_\epsilon \leq (1 + 2 \log B) \sqrt{\ell_\epsilon} + 2 \sqrt{\ell_\epsilon} = (3 + 2 \log B) \sqrt{\ell_\epsilon}.$$

From this, we conclude that $Q_{\sigma_\epsilon; t_0} \leq \eta_0^2 (3 + 2 \log B) \sqrt{\ell_\epsilon}$. In the case where $\frac{1}{2} < a < 1$, given $\sum_{k=1}^{\infty} k^{-2a} \leq 1 + \int_1^{\infty} x^{-2a} dx = 1 + \frac{1}{2a-1}$, it holds that

$$Q_{\sigma_\epsilon; t_0} \leq \eta_0^2 (1 + \frac{1}{2a-1}) \leq \eta_0^2 (1 + \frac{1}{2a-1}) \omega_a(\ell_\epsilon),$$

where the last inequality is by $\omega_a(\ell_\epsilon) = \ell_\epsilon^{1-a} \geq 1$ given $\ell_\epsilon \geq 1$. Finally, in the case where $a = 1$, we have

$$Q_{\sigma_\epsilon; t_0} \leq \eta_0^2 \sum_{t=0}^{\infty} (t+1)^{-2} = \frac{\pi^2}{6} \eta_0^2 \leq \frac{\pi^2}{6} \eta_0^2 \log \ell_\epsilon = \frac{\pi^2}{6} \eta_0^2 \omega_1(\ell_\epsilon),$$

where we used $\omega_1(\ell_\epsilon) = \log \ell_\epsilon \geq 1$ given $\ell_\epsilon \geq e$. In summary, we let $C_2 = \frac{\eta_0^2}{1-2a} B^{\frac{1-2a}{1-a}}$ when $\frac{1}{3} < a < \frac{1}{2}$, $C_2 = \eta_0^2(3 + 2\log B)$ when $a = \frac{1}{2}$, $C_2 = \eta_0^2(1 + \frac{1}{2a-1})$ when $\frac{1}{2} < a < 1$, and $C_2 = \frac{\pi^2}{6}\eta_0^2$ when $a = 1$, then it holds that $Q_{\sigma_\epsilon;t_0} \leq C_2\omega_a(\ell_\epsilon)$.

Define $U_{\sigma_\epsilon} := S_{\sigma_\epsilon} - S_{T_*}$ and $B_0 := S_{t_0} - S_{T_*}$ (note that B_0 is a constant independent of ϵ). Then it holds that $V_\epsilon = U_{\sigma_\epsilon} - B_0$. Substituting this into (123), together with the bound $\sum_{t=t_0}^{\sigma_\epsilon - 1} \eta_t L_t \leq \frac{4\log 2}{n\gamma_\infty}$, we obtain

$$\begin{aligned}
\gamma_\infty U_{\sigma_\epsilon} &\leq \log \frac{L_{t_0}}{L_{\sigma_\epsilon}} + \gamma_\infty B_0 + \frac{d}{\Lambda_\epsilon} U_{\sigma_\epsilon} - \frac{dB_0}{\Lambda_\epsilon} + C_{\mathrm{md}} Q_{\sigma_\epsilon;t_0} + 2\log 2 \\
&\leq \log \frac{L_{t_0}}{L_{\sigma_\epsilon}} + \gamma_\infty B_0 + \frac{d}{\Lambda_\epsilon} U_{\sigma_\epsilon} + C_{\mathrm{md}} Q_{\sigma_\epsilon;t_0} + 2\log 2 \\
&\leq \ell_\epsilon + C_1 + \gamma_\infty B_0 + \frac{d}{\Lambda_\epsilon} U_{\sigma_\epsilon} + C_{\mathrm{md}} C_2 \omega_a(\ell_\epsilon) + 2\log 2 \\
&\leq \ell_\epsilon + \frac{d}{\Lambda_\epsilon} U_{\sigma_\epsilon} + C_{\mathrm{md}} C_2 \omega_a(\ell_\epsilon) + C_3 \\
&\leq \ell_\epsilon + C_4[1 + \frac{U_{\sigma_\epsilon}}{\Lambda_\epsilon} + \omega_a(\ell_\epsilon)], \qquad (126)
\end{aligned}$$

where $C_3 := C_1 + \gamma_\infty B_0 + 2\log 2$ and $C_4 := \max\{d, C_{\mathrm{md}} C_2, C_3\}$. Note that U_{σ_ϵ} appears in both sides of the previous equation. We remove it in two stages. We first show the coarse bound $U_{\sigma_\epsilon} = O(\ell_\epsilon)$. Choosing ϵ sufficiently small such that $\frac{C_4}{\Lambda_\epsilon} \leq \frac{\gamma_\infty}{2}$, we obtain from (126) that

$$\frac{\gamma_\infty}{2} U_{\sigma_\epsilon} \leq (\gamma_\infty - \frac{C_4}{\Lambda_\epsilon}) U_{\sigma_\epsilon} \leq \ell_\epsilon + C_4[1 + \omega_a(\ell_\epsilon)].$$

Therefore, it holds that $U_{\sigma_\epsilon} \leq \frac{2}{\gamma_\infty}[\ell_\epsilon + C_4(1 + \omega_a(\ell_\epsilon))]$. Given $\omega_a(\ell_\epsilon) \leq \ell_\epsilon$ (as $\ell_\epsilon \geq 1$), we further have

$$U_{\sigma_\epsilon} \leq \frac{2(1 + 2C_4)}{\gamma_\infty} \ell_\epsilon = C_U \ell_\epsilon,$$

where $C_U := \frac{2(1+2C_4)}{\gamma_\infty}$. This implies that $\frac{U_{\sigma_\epsilon}}{\Lambda_\epsilon} \leq C_U \frac{\ell_\epsilon}{\Lambda_\epsilon}$. By the definition of ℓ_ϵ and Λ_ϵ, it holds that $\frac{\ell_\epsilon}{\Lambda_\epsilon} = \omega_a(\ell_\epsilon)$. Therefore, we have that $\frac{U_{\sigma_\epsilon}}{\Lambda_\epsilon} \leq C_U \omega_a(\ell_\epsilon)$. Substituting this into (126), we obtain

$$\gamma_\infty U_{\sigma_\epsilon} \leq \ell_\epsilon + C_4[1 + C_U \omega_a(\ell_\epsilon) + \omega_a(\ell_\epsilon)].$$

Given $\ell_\epsilon \geq e$, it holds that $\omega_a(\ell_\epsilon) \geq 1$ and consequently $1 + (C_U + 1)\omega_a(\ell_\epsilon) \leq (C_U + 2)\omega_a(\ell_\epsilon)$. Therefore, we have $\gamma_\infty U_{\sigma_\epsilon} \leq \ell_\epsilon + C_4(C_U + 2)\omega_a(\ell_\epsilon)$. After dividing by γ_∞, we obtain

$$U_{\sigma_\epsilon} \leq \frac{\ell_\epsilon}{\gamma_\infty} + \frac{C_4(C_U + 2)}{\gamma_\infty} \omega_a(\ell_\epsilon).$$

From this, we conclude that $S_{\sigma_\epsilon} - S_{T_*} \leq \frac{\ell_\epsilon}{\gamma_\infty} + C_\sigma \omega_a(\ell_\epsilon)$ where $C_\sigma := \frac{C_4(C_U+2)}{\gamma_\infty}$. Finally, we give concrete choices of ϵ such that all conditions above are satisfied. We define:

$$M_\delta := \max\{1, 2\delta, \frac{4d}{\gamma_\infty}, \frac{2C_4}{\gamma_\infty}\}, \quad L_{1,\delta} := \begin{cases} \max\left\{\mathrm{e},\, 2\log\left(\frac{2}{c_{t_0}}\right),\, M_\delta^{1/a}\right\}, & \frac{1}{3} < a < 1, \\ \max\left\{\mathrm{e},\, 2\log\left(\frac{2}{c_{t_0}}\right),\, M_\delta^2\right\}, & a = 1. \end{cases}$$

Let $\epsilon_1(\delta) := e^{-L_{1,\delta}}$. We now verify every $0 < \epsilon \leq \epsilon_1(\delta)$ satisfies all the desired properties. Given $\epsilon \leq e^{-L_{1,\delta}}$, it holds that $\ell_\epsilon = \log \frac{1}{\epsilon} \geq L_{1,\delta}$. Consequently, it holds that $\ell_\epsilon \geq e$ and $\ell_\epsilon \geq 2\log \frac{2}{c_{t_0}}$. Given $\ell_\epsilon \geq M_\delta^{1/a}$ when $\frac{1}{3} < a < 1$, we have $\Lambda_\epsilon = \ell_\epsilon^a \geq (M_\delta^{1/a})^a = M_\delta$. Given $\ell_\epsilon \geq M_\delta^2$ when $a = 1$, we have $\Lambda_\epsilon = \frac{\ell_\epsilon}{\log \ell_\epsilon} \geq \sqrt{\ell_\epsilon} \geq M_\delta$ where we used $\log x \leq \sqrt{x}$ when $x \geq 1$. Therefore, it holds that $\Lambda_\epsilon \geq M_\delta$ in both cases. Moreover, it holds that

$$\epsilon\Lambda_\epsilon \overset{(a)}{\leq} e^{-\ell_\epsilon}\ell_\epsilon \overset{(b)}{\leq} e^{-\ell_\epsilon/2} \overset{(c)}{\leq} \frac{c_{t_0}}{2} < c_{t_0},$$

where (a) is by $\Lambda_\epsilon \le \ell_\epsilon$ (as $\ell_\epsilon \ge e > 1$), (b) is by $\log x < \frac{x}{2}$ for every $x > 0$, and (c) is by $\ell_\epsilon \ge 2\log\frac{2}{c_{t_0}}$. In summary, the following conditions simultaneously hold for every $0 < \epsilon \le \epsilon_1(\delta)$: $\ell_\epsilon \ge e > 1$, $\Lambda_\epsilon \ge 2\delta > \delta$, $\Lambda_\epsilon \ge 1$, $\frac{d}{\Lambda_\epsilon} \le \frac{\gamma_\infty}{4}$, $\frac{C_4}{\Lambda_\epsilon} \le \frac{\gamma_\infty}{2}$, and $\epsilon\Lambda_\epsilon < c_{t_0}$. Consequently, it holds that $\sigma_\epsilon > t_0$ and $\sigma_\epsilon \le \tau_\epsilon$. For the remainder of the proof, we assume $0 < \epsilon \le \epsilon_1(\delta)$ and track additional conditions on ϵ that haven't been included.

Part 2 Next, we want to show $S_{\tau_\epsilon} - S_{T_*} \le \frac{\ell_\epsilon}{\gamma_\infty} + O_\delta(\omega_a(\ell_\epsilon))$. If $\sigma_\epsilon = \tau_\epsilon$, the desired bound follows immediately. Below, we consider the nontrivial case $\sigma_\epsilon < \tau_\epsilon$. By the definition of τ_ϵ, we have

$$G_t > \delta\epsilon, \qquad \sigma_\epsilon \le t < \tau_\epsilon. \tag{127}$$

Lemma 18 gives $\langle g, \sigma_\epsilon(g)\rangle \ge \frac{\|g\|_1^2}{\|g\|_1 + d\epsilon}$. From $\|g_t\|_1 \ge \gamma_\infty G_t$ and $\|g_t\|_1 \le R_1 G_t$, we have

$$\langle g_t, \sigma_\epsilon(g_t)\rangle \ge \frac{\|g_t\|_1^2}{\|g_t\|_1 + d\epsilon} \ge \frac{\gamma_\infty^2 G_t^2}{R_1 G_t + d\epsilon} = \frac{\gamma_\infty^2}{R_1 + d\epsilon/G_t} G_t.$$

From (127), we have $\frac{d\epsilon}{G_t} < \frac{d}{\delta}$. Consequently, it holds that $\frac{1}{R_1 + d\epsilon/G_t} \ge \frac{1}{R_1 + d/\delta}$. Define $\kappa_\delta := \frac{\gamma_\infty{}^2}{R_1 + d/\delta} = \frac{\gamma_\infty^2 \delta}{R_1\delta + d} > 0$, then it holds that

$$\langle g_t, \sigma_\epsilon(g_t)\rangle \ge \kappa_\delta G_t, \qquad \sigma_\epsilon \le t < \tau_\epsilon.$$

From (44), we have

$$L_{t+1} \le L_t - \kappa_\delta \eta_t G_t + C_{\mathrm{md}} \eta_t^2 G_t = L_t - \eta_t(\kappa_\delta - C_{\mathrm{md}}\eta_t) G_t.$$

Given $\sigma_\epsilon \to \infty$, we have $\eta_{\sigma_\epsilon} \to 0$. Consequently, for sufficiently small ϵ, we have $C_{\mathrm{md}}\eta_t \le C_{\mathrm{md}}\eta_{\sigma_\epsilon} \le \frac{\kappa_\delta}{2}, t \ge \sigma_\epsilon$. This leads to

$$L_{t+1} \le L_t - \frac{\kappa_\delta}{2}\eta_t G_t, \qquad \sigma_\epsilon \le t < \tau_\epsilon.$$

For $t \ge \sigma_\epsilon \ge T_*$, we have $G_t \ge \frac{1}{2} L_t$ by Proposition 2.1. Using this, we further obtain

$$L_{t+1} \le Lt - \frac{\kappa_\delta}{4}\eta_t L_t = (1 - c_\delta \eta_t) L_t, \qquad \sigma_\epsilon \le t < \tau_\epsilon. \tag{128}$$

where $c_\delta := \frac{\kappa_\delta}{4}$. We additionally require ϵ to be sufficiently small such that $\eta_{\sigma_\epsilon} \le \frac{2}{\kappa_\delta}$. This would ensure that $\frac{1}{2} \le 1 - \frac{\kappa_\delta}{4}\eta_t \le 1$. Iterating (128) from σ_ϵ to any $\sigma_\epsilon \le t \le \tau_\epsilon$, it holds that

$$\begin{aligned} L_t &\le L_{\sigma_\epsilon} \prod_{s=\sigma_\epsilon}^{t-1} (1 - c_\delta \eta_s) \le L_{\sigma_\epsilon} \prod_{s=\sigma_\epsilon}^{t-1} e^{-c_\delta \eta_s} = L_{\sigma_\epsilon} \exp(-c_\delta \sum_{s=\sigma_\epsilon}^{t-1} \eta_s) \\ &= L_{\sigma_\epsilon} \exp[-\frac{\kappa_\delta}{4}(S_t - S_{\sigma_\epsilon})]. \end{aligned} \tag{129}$$

By the definition of σ_ϵ, we have $G_{\sigma_\epsilon} \le \epsilon\Lambda_\epsilon$. Since $\sigma_\epsilon \ge T_*$, Proposition 2.1 gives $L_{\sigma_\epsilon} \le 2G_{\sigma_\epsilon}$. Consequently, it holds that $L_{\sigma_\epsilon} \le 2\epsilon\Lambda_\epsilon$. Define $A_\epsilon := \frac{4}{\kappa_\delta}\log\frac{2\Lambda_\epsilon}{\delta}$. Given $\Lambda_\epsilon > \delta$, it holds that $A_\epsilon > 0$. Define $\bar\tau_\epsilon := \inf\{t \ge \sigma_\epsilon : S_t - S_{\sigma_\epsilon} \ge A_\epsilon\}$. This time is finite because $S_t - S_{\sigma_\epsilon} = \sum_{s=\sigma_\epsilon}^{t-1} \eta_s \to \infty$ as $t \to \infty$. We claim that $\tau_\epsilon \le \bar\tau_\epsilon$. Suppose (for contradiction) that $\bar\tau_\epsilon < \tau_\epsilon$. Applying (129) and using $L_{\sigma_\epsilon} \le 2\epsilon\Lambda_\epsilon$, we obtain

$$L_{\bar\tau_\epsilon} \le L_{\sigma_\epsilon} \exp[-\frac{\kappa_\delta}{4}(S_{\bar\tau_\epsilon} - S_{\sigma_\epsilon})] \le 2\epsilon\Lambda_\epsilon \exp[-\frac{\kappa_\delta}{4}A_\epsilon] = 2\epsilon\Lambda_\epsilon \frac{\delta}{2\Lambda_\epsilon} = \delta\epsilon.$$

Consequently, we have $G_{\bar\tau_\epsilon} \le L_{\bar\tau_\epsilon} \le \delta\epsilon$, which means that $\bar\tau_\epsilon$ belongs to the set $\{t \ge T^* : G_t \le \delta\epsilon\}$. By the definition of τ_ϵ, it holds that $\tau_\epsilon \le \bar\tau_\epsilon$, which is a contradiction. Therefore, the claim $\tau_\epsilon \le \bar\tau_\epsilon$ holds. By minimality of $\bar\tau_\epsilon$, it holds that $S_{\bar\tau_\epsilon - 1} - S_{\sigma_\epsilon} < A_\epsilon$ and $S_{\bar\tau_\epsilon} - S_{\sigma_\epsilon} \ge A_\epsilon$. This implies that $S_{\bar\tau_\epsilon} - S_{\sigma_\epsilon} = S_{\bar\tau_\epsilon - 1} - S_{\sigma_\epsilon} + \eta_{\bar\tau_\epsilon - 1} < A_\epsilon + \eta_{\bar\tau_\epsilon - 1}$. Since η_t is decreasing and $\bar\tau_\epsilon - 1 \ge \sigma_\epsilon$, we have $\eta_{\bar\tau_\epsilon - 1} \le \eta_{\sigma_\epsilon}$ and consequently $S_{\bar\tau_\epsilon} - S_{\sigma_\epsilon} < A_\epsilon + \eta_{\sigma_\epsilon}$. Given $\tau_\epsilon \le \bar\tau_\epsilon$ and S_t is increasing, we conclude that

$$S_{\tau_\epsilon} - S_{\sigma_\epsilon} \le S_{\bar\tau_\epsilon} - S_{\sigma_\epsilon} \le \frac{4}{\kappa_\delta}\log\frac{2\Lambda_\epsilon}{\delta} + \eta_{\sigma_\epsilon} \le \frac{4}{\kappa_\delta}\log\Lambda_\epsilon + \frac{4}{\kappa_\delta}\log\frac{2}{\delta} + \eta_0. \tag{130}$$

Next, we show that $\log \Lambda_\epsilon = O(\omega_a(\ell_\epsilon))$. For $a \in (\frac{1}{3}, 1)$, we have $\Lambda_\epsilon = \ell_\epsilon^a$ and $\log \Lambda_\epsilon = a \log \ell_\epsilon$. For any $p > 0$, since $u^p \geq 1$ when $u \geq 1$, it holds that

$$\log x = \int_1^x \frac{1}{u} du \leq \int_1^x u^{p-1} du = \frac{x^p - 1}{p} \leq \frac{x^p}{p}.$$

Taking $x = \ell_\epsilon$ and $p = 1 - a$ gives $\log \ell_\epsilon \leq \frac{\ell_\epsilon^{1-a}}{1-a}$. Therefore, we conclude that

$$\log \Lambda_\epsilon = a \log \ell_\epsilon \leq \frac{a}{1-a} \ell_\epsilon^{1-a} = \frac{a}{1-a} \omega_a(\ell_\epsilon).$$

For $a = 1$, we have $\Lambda_\epsilon = \frac{\ell_\epsilon}{\log \ell_\epsilon}$. For $\ell_\epsilon \geq e$, it holds that

$$\log \Lambda_\epsilon = \log \ell_\epsilon - \log \log \ell_\epsilon \leq \log \ell_\epsilon = \omega_1(\ell_\epsilon).$$

Therefore, by defining $c_a := \frac{a}{1-a}$ when $a \in (\frac{1}{3}, 1)$ and $c_a := 1$ when $a = 1$, it holds that $\log \Lambda_\epsilon \leq c_a \omega_a(\ell_\epsilon)$. Given $\omega_a(\ell_\epsilon) \geq 1$ (as $\ell_\epsilon \geq e$), we obtain via (130) that

$$S_{\tau_\epsilon} - S_{\sigma_\epsilon} \leq C_{\mathrm{tr},\delta} \omega_a(\ell_\epsilon),$$

where $C_{\mathrm{tr},\delta} := \eta_0 + \frac{4}{\kappa_\delta}(c_a + \log \frac{2}{\delta}) = \eta_0 + \frac{4(R_1 \delta + d)}{\gamma_\infty^2 \delta}(c_a + \log \frac{2}{\delta})$. Previously, we have shown that $S_{\sigma_\epsilon} - S_{T_*} \leq \frac{\ell_\epsilon}{\gamma_\infty} + C_\sigma \omega_a(\ell_\epsilon)$. Combining them together, we obtain

$$S_{\tau_\epsilon} - S_{T_*} = S_{\tau_\epsilon} - S_{\sigma_\epsilon} + S_{\sigma_\epsilon} - S_{T_*} \leq \frac{\ell_\epsilon}{\gamma_\infty} + C_{\tau,\delta} \omega_a(\ell_\epsilon), \tag{131}$$

where $C_{\tau,\delta} := C_\sigma + C_{\mathrm{tr},\delta}$. The additional conditions on the choice of ϵ in this part are $C_{\mathrm{md}} \eta_{\sigma_\epsilon} \leq \frac{\kappa_\delta}{2}$ and $\eta_{\sigma_\epsilon} \leq \frac{2}{\kappa_\delta}$. Choose $\bar{\eta}_{2,\delta} := \min\{\eta_0, \frac{\kappa_\delta}{2C_{\mathrm{md}}}, \frac{2}{\kappa_\delta}\}$ and $N_{2,\delta} := \max\{t_0, \lceil (\frac{\eta_0}{\bar{\eta}_{2,\delta}})^{1/a} \rceil\}$. It holds that $\eta_{N_{2,\delta}} = \frac{\eta_0}{(N_{2,\delta}+1)^a} \leq \bar{\eta}_{2,\delta}$. By Lemma 12, we have for every $0 \leq t \leq N_{2,\delta}$ that

$$G_t \geq G_0 e^{-K_0 S_t} \geq G_0 e^{-K_0 S_{N_{2,\delta}}} = c_{2,\delta}. \tag{132}$$

Define $\epsilon_2(\delta) := (\frac{c_{2,\delta}}{2})^2$ and choose $0 < \epsilon \leq \min\{\epsilon_1(\delta), \epsilon_2(\delta)\}$. Then, we have

$$\epsilon \Lambda_\epsilon \leq e^{-\ell_\epsilon} \ell_\epsilon \leq e^{-\ell_\epsilon/2} = \sqrt{\epsilon} \leq \frac{c_{2,\delta}}{2} < c_{2,\delta}. \tag{133}$$

Combining (132) and (133), it holds that

$$G_t > \epsilon \Lambda_\epsilon, \qquad 0 \leq t \leq N_{2,\delta}.$$

This implies that $\sigma_\epsilon > N_{2,\delta}$ and consequently $\eta_{\sigma_\epsilon} \leq \eta_{N_{2,\delta}} \leq \bar{\eta}_{2,\delta}$. Therefore, it holds that $C_{\mathrm{md}} \eta_{\sigma_\epsilon} \leq C_{\mathrm{md}} \bar{\eta}_{2,\delta} \leq \frac{\kappa_\delta}{2}$ and $\eta_{\sigma_\epsilon} \leq \bar{\eta}_{2,\delta} \leq \frac{2}{\kappa_\delta}$. For the remainder of the proof, we assume $0 < \epsilon \leq \min\{\epsilon_1(\delta), \epsilon_2(\delta)\}$ and track additional conditions on ϵ that haven't been included.

Part 3 We prove that $\|w_{\tau_\epsilon}\|_\infty \leq S_{\tau_\epsilon} - S_{T_*} + O(\omega_a(\ell_\epsilon))$. Its proof uses the radius bound of Lemma 1. Define $k_\epsilon := \lceil \ell_\epsilon \rceil$. Given ϵ is chosen such that $\ell_\epsilon \geq e$, it ensures that $k_\epsilon \leq \ell_\epsilon + 1 \leq 2\ell_\epsilon$ and $\log \ell_\epsilon \geq 1$. In the case of $a \in (\frac{1}{3}, 1)$, using integral approximation we have

$$\begin{aligned} S_{k_\epsilon} &= \eta_0 \sum_{j=1}^{k_\epsilon} j^{-a} \leq \eta_0 (1 + \int_1^{k_\epsilon} x^{-a} dx) = \eta_0 (1 + \frac{k_\epsilon^{1-a} - 1}{1-a}) = \eta_0 \frac{k_\epsilon^{1-a} - a}{1-a} \\ &\leq \frac{\eta_0}{1-a} k_\epsilon^{1-a} \leq \frac{2^{1-a} \eta_0}{1-a} \ell_\epsilon^{1-a}. \end{aligned}$$

In the case where $a = 1$, we have

$$S_{k_\epsilon} = \eta_0 \sum_{j=1}^{k_\epsilon} \frac{1}{j} \leq \eta_0 (1 + \log k_\epsilon) \leq \eta_0 (1 + \log 2 + \log \ell_\epsilon) \leq \eta_0 (2 + \log 2) \log \ell_\epsilon.$$

Define $K_S := \frac{2^{1-a} \eta_0}{1-a}$ when $a \in (\frac{1}{3}, 1)$ and $K_S := \eta_0 (2 + \log 2)$ when $a = 1$, it then holds that $S_{k_\epsilon} \leq K_S \omega_a(\ell_\epsilon)$. This implies that $\frac{S_{k_\epsilon}}{\ell_\epsilon} \leq K_S \frac{\omega_a(\ell_\epsilon)}{\ell_\epsilon} \to 0$ and therefore $S_{k_\epsilon} = o(\ell_\epsilon)$.

Define $L_{3,\delta} := \max\{T_*, 2[\log \frac{\delta}{G_0}]_+, \Omega_a\}$ where $\Omega_a := (4K_0K_S)^{1/a}$ when $\frac{1}{3} < a < 1$ and $(4K_0K_S)^2$ when $a = 1$. We show that $T_* \le k_\epsilon < \tau_\epsilon$ for every $0 < \epsilon \le \epsilon_3(\delta) := e^{-L_{3,\delta}}$. Lemma 12 gives $G_{\tau_\epsilon} \ge G_0 \exp(-K_0 S_{\tau_\epsilon})$. By the definition of τ_ϵ, we have $G_{\tau_\epsilon} \le \delta\epsilon$ and therefore $G_0 \exp(-K_0 S_{\tau_\epsilon}) \le \delta\epsilon$. This implies that $S_{\tau_\epsilon} \ge \frac{1}{K_0} \log \frac{G_0}{\delta\epsilon} = \frac{1}{K_0}(\ell_\epsilon + \log \frac{G_0}{\delta})$. Therefore, given $\ell_\epsilon \ge 2[\log \frac{\delta}{G_0}]_+$, it holds that $S_{\tau_\epsilon} \ge \frac{\ell_\epsilon}{2K_0}$. When $\frac{1}{3} < a < 1$, the condition $\ell_\epsilon \ge \Omega_a$ implies that $K_S \ell_\epsilon^{1-a} \le \frac{\ell_\epsilon}{4K_0}$, using $\log x \le \sqrt{x}, x \ge 1$, the condition $\ell_\epsilon \ge \Omega_1 = (4K_0K_S)^2$ implies that $K_S \log \ell_\epsilon \le K_S\sqrt{\ell_\epsilon} \le \frac{\ell_\epsilon}{4K_0}$. Consequently, in both cases, we have

$$S_{k_\epsilon} \le K_S \omega_a(\ell_\epsilon) \le \frac{\ell_\epsilon}{4K_0} < \frac{\ell_\epsilon}{2K_0} \le S_{\tau_\epsilon}.$$

Since S_t is strictly increasing, this implies that $k_\epsilon < \tau_\epsilon$. Given $\ell_\epsilon \ge T_*$, it holds that $k_\epsilon = \lceil \ell_\epsilon \rceil \ge \ell_\epsilon$. Therefore, we conclude that $T_* \le k_\epsilon < \tau_\epsilon$ for every $0 < \epsilon \le \epsilon_3(\delta)$.

From Lemma 11 and $w_{k_\epsilon} = w_0 - \sum_{t=0}^{k_\epsilon - 1} \eta_t D_t$, we have

$$\begin{aligned}\|w_{k_\epsilon}\|_\infty &\le \|w_0\|_\infty + \sum_{t=0}^{k_\epsilon-1} \eta_t \|D_t\|_\infty \le \|w_0\|_\infty + C_D S_{k_\epsilon} \\ &\le \|w_0\|_\infty + C_D K_S \omega_a(\ell_\epsilon) \le (\|w_0\|_\infty + C_D K_S)\omega_a(\ell_\epsilon) \\ &= K_p \omega_a(\ell_\epsilon),\end{aligned}$$

where $K_p := \|w_0\|_\infty + C_D K_S$ and we used $\omega_a(\ell_\epsilon) \ge 1$ given $\ell_\epsilon \ge e$. Therefore, it holds that $\|w_{k_\epsilon}\|_\infty = O(\omega_a(\ell_\epsilon))$.

Given $|g_t[j]| \le R_1 G_t \le R_1, \forall j$, we can set $B_g = R_1$ in Lemma 1. With this B_g, we define $H_\epsilon := 1 + \log \frac{R_1^2 + \epsilon^2}{(1-\beta_2)\epsilon^2}$. It then holds that

$$H_\epsilon = 2\ell_\epsilon + 1 + \log \frac{R_1^2 + \epsilon^2}{1 - \beta_2} \le 2\ell_\epsilon + H_0, \tag{134}$$

where $H_0 := 1 + \log \frac{R_1^2 + 1}{1 - \beta_2}$ and we used $R_1^2 + \epsilon^2 \le R_1^2 + 1$ given $0 < \epsilon \le 1$.

Next, we bound the term $\eta_{k_\epsilon} H_\epsilon$. Because $k_\epsilon + 1 \ge \ell_\epsilon$, it holds that $\eta_{k_\epsilon} = \eta_0 \frac{1}{(k_\epsilon + 1)^a} \le \eta_0 \ell_\epsilon^{-a}$ and consequently $\eta_{k_\epsilon} H_\epsilon \le \eta_0 \ell_\epsilon^{-a}(2\ell_\epsilon + H_0)$ by (134). For $a \in (\frac{1}{3}, 1)$, we have $\ell_\epsilon^{-a} \le \ell_\epsilon^{1-a}$ since $\ell_\epsilon \ge 1$. Therefore, it holds that

$$\eta_{k_\epsilon} H_\epsilon \le 2\eta_0 \ell_\epsilon^{1-a} + H_0 \eta_0 \ell_\epsilon^{-a} \le \eta_0(2 + H_0)\ell_\epsilon^{1-a}$$

In the case where $a = 1$, we have

$$\begin{aligned}\eta_{k_\epsilon} H_\epsilon &\le \eta_0 \ell_\epsilon^{-1}(2\ell_\epsilon + H_0) = \eta_0(2 + \frac{H_0}{\ell_\epsilon}) \\ &\le \eta_0(2 + H_0) \le \eta_0(2 + H_0)\log \ell_\epsilon,\end{aligned}$$

where we used $\ell_\epsilon \ge 1$ and $\log \ell_\epsilon \ge 1$. Therefore, in both cases, we have $\eta_{k_\epsilon} H_\epsilon \le \eta_0(2 + H_0)\omega_a(\ell_\epsilon)$. Defining $K_r := C_{\mathrm{rad}} \eta_0 (2 + H_0)$ and multiplying by C_{rad}, we conclude

$$C_{\mathrm{rad}} \eta_{k_\epsilon} H_\epsilon \le K_r \omega_a(\ell_\epsilon). \tag{135}$$

Because $k_\epsilon < \tau_\epsilon$, we apply Lemma 1 with $r = k_\epsilon$ and $T = \tau_\epsilon$ to obtain

$$\begin{aligned}\|\sum_{t=k_\epsilon}^{\tau_\epsilon - 1} \eta_t D_t\|_\infty &\le S_{\tau_\epsilon} - S_{k_\epsilon} + C_{\mathrm{rad}} \eta_{k_\epsilon} H_\epsilon \\ &\le S_{\tau_\epsilon} - S_{k_\epsilon} + K_r \omega_a(\ell_\epsilon).\end{aligned}$$

Based on $\|w_{k_\epsilon}\|_\infty \le K_p \omega_a(\ell_\epsilon)$ and $w_{\tau_\epsilon} = w_{k_\epsilon} - \sum_{t=k_\epsilon}^{\tau_\epsilon - 1} \eta_t D_t$, it holds that

$$\begin{aligned}\|w_{\tau_\epsilon}\|_\infty &\le \|w_{k_\epsilon}\|_\infty + \|\sum_{t=k_\epsilon}^{\tau_\epsilon - 1} \eta_t D_t\|_\infty \\ &\le K_p \omega_a(\ell_\epsilon) + S_{\tau_\epsilon} - S_{k_\epsilon} + K_r \omega_a(\ell_\epsilon) \\ &= S_{\tau_\epsilon} - S_{k_\epsilon} + (K_p + K_r)\omega_a(\ell_\epsilon).\end{aligned}$$

Because $k_\epsilon \geq T_*$, it holds that $S_{\tau_\epsilon} - S_{k_\epsilon} \leq S_{\tau_\epsilon} - S_{T_*}$. By defining $K_\tau := K_p + K_r = \|w_0\|_\infty + C_D K_S + C_{\text{rad}}\eta_0[3 + \log \frac{R_1^2+1}{1-\beta_2}]$, we conclude that

$$\|w_{\tau_\epsilon}\|_\infty \leq S_{\tau_\epsilon} - S_{T^*} + K_\tau \omega_a(\ell_\epsilon). \tag{136}$$

Combining this with $S_{\tau_\epsilon} - S_{T_*} \leq \frac{\ell_\epsilon}{\gamma_\infty} + C_{\tau,\delta}\omega_a(\ell_\epsilon)$ (shown above), we conclude that

$$\|w_{\tau_\epsilon}\|_\infty \leq \frac{\ell_\epsilon}{\gamma_\infty} + C_{w,\delta}\omega_a(\ell_\epsilon), \tag{137}$$

where $C_{w,\delta} := C_{\tau,\delta} + K_\tau$. So far, the thresholds on ϵ that have been introduced are $\epsilon_1(\delta)$, $\epsilon_2(\delta)$, and $\epsilon_3(\delta)$. Consequently, we set $\tilde{\epsilon}_2(\delta) = \min\{\epsilon_1(\delta), \epsilon_2(\delta), \epsilon_3(\delta)\}$ and choose $0 < \epsilon \leq \tilde{\epsilon}_2(\delta)$. This ensures all the requirements on ϵ appeared in Parts 1–3 are satisfied and the conclusion in (137) holds.

Part 4 By definition, we have $\gamma_\infty = \max_{\|u\|_\infty \leq 1} \rho(u)$. It holds that $\rho(w) \leq \gamma_\infty \|w\|_\infty$ for every w. By definition of τ_ϵ, we have $G_{\tau_\epsilon} \leq \delta\epsilon$. Since $\tau_\epsilon \geq T_*$, Proposition 2.1 gives $L_{\tau_\epsilon} \leq 2G_{\tau_\epsilon} \leq 2\delta\epsilon$. From Lemma 8, it also holds that $L_{\tau_\epsilon} \geq \frac{\log 2}{n} e^{-\rho(w_{\tau_\epsilon})}$. Consequently, we have $\frac{\log 2}{n} e^{-\rho(w_{\tau_\epsilon})} \leq 2\delta\epsilon$. Rearranging and taking logarithms lead to

$$\rho(w_{\tau_\epsilon}) \geq \ell_\epsilon - \log \frac{2n\delta}{\log 2}. \tag{138}$$

By defining $A_\delta := [\log \frac{2n\delta}{\log 2}]_+$, we further obtain $\rho(w_{\tau_\epsilon}) \geq \ell_\epsilon - A_\delta$ given $A_\delta = \max\{\log \frac{2n\delta}{\log 2}, 0\} \geq \log \frac{2n\delta}{\log 2}$. Define $R_\epsilon := \|w_{\tau_\epsilon}\|_\infty$. By (137), it holds that $R_\epsilon \leq \frac{\ell_\epsilon}{\gamma_\infty} + C_{w,\delta}\omega_a(\ell_\epsilon)$. Then, we have

$$\hat{\gamma}_\infty(w_{\tau_\epsilon}) = \frac{\rho(w_{\tau_\epsilon})}{R_\epsilon} \geq \frac{\ell_\epsilon - A_\delta}{R_\epsilon} \geq \frac{\ell_\epsilon - A_\delta}{\ell_\epsilon/\gamma_\infty + C_{w,\delta}\omega_a(\ell_\epsilon)}.$$

Consequently, this leads to

$$\begin{aligned}\gamma_\infty - \hat{\gamma}_\infty(w_{\tau_\epsilon}) &\leq \gamma_\infty - \frac{\ell_\epsilon - A_\delta}{\ell_\epsilon/\gamma_\infty + C_{w,\delta}\omega_a(\ell_\epsilon)} \\ &= \frac{\gamma_\infty C_{w,\delta}\omega_a(\ell_\epsilon) + A_\delta}{\ell_\epsilon/\gamma_\infty + C_{w,\delta}\omega_a(\ell_\epsilon)}.\end{aligned}$$

Given $C_{w,\delta} \geq 0$ and $\omega_a(\ell_\epsilon) \geq 0$, it holds that $\frac{\ell_\epsilon}{\gamma_\infty} + C_{w,\delta}\omega_a(\ell_\epsilon) \geq \frac{\ell_\epsilon}{\gamma_\infty}$. Consequently, we have

$$\gamma_\infty - \hat{\gamma}_\infty(w_{\tau_\epsilon}) \leq \gamma_\infty \frac{\gamma_\infty C_{w,\delta}\omega_a(\ell_\epsilon) + A_\delta}{\ell_\epsilon} = \gamma_\infty^2 C_{w,\delta}\frac{\omega_a(\ell_\epsilon)}{\ell_\epsilon} + \gamma_\infty A_\delta \frac{1}{\ell_\epsilon}.$$

Recall that $r_\epsilon = \frac{\omega_a(\ell_\epsilon)}{\ell_\epsilon}$. Given $\omega_a(\ell_\epsilon) \geq 1$, it holds that $r_\epsilon \geq \frac{1}{\ell_\epsilon}$. Then, we have

$$0 \leq \gamma_\infty - \hat{\gamma}_\infty(w_{\tau_\epsilon}) \leq \gamma_\infty^2 C_{w,\delta} r_\epsilon + \gamma_\infty A_\delta r_\epsilon = C_{\infty,\delta} r_\epsilon, \tag{139}$$

where $C_{\infty,\delta} := \gamma_\infty^2 C_{w,\delta} + \gamma_\infty A_\delta$. For $a \in (\frac{1}{3}, 1)$, it holds that $r_\epsilon = \ell_\epsilon^{-a} \to 0$ because $a > 0$ and $\ell_\epsilon \to \infty$. For $a = 1$, we again have $r_\epsilon = \frac{\log \ell_\epsilon}{\ell_\epsilon} \to 0$. Consequently, by (139), it holds that $\hat{\gamma}_\infty(w_{T_1^\delta(\epsilon)}) \to \gamma_\infty$ as $\epsilon \downarrow 0$.

From $\rho(w_{\tau_\epsilon}) \leq \gamma_\infty \|w_{\tau_\epsilon}\|_\infty$, we further obtain $\|w_{\tau_\epsilon}\|_\infty \geq \frac{\ell_\epsilon}{\gamma_\infty} - \frac{1}{\gamma_\infty}\log \frac{2n\delta}{\log 2}$ from (138). By defining $A_\delta := [\log \frac{2n\delta}{\log 2}]_+$, it holds that

$$\|w_{\tau_\epsilon}\|_\infty \geq \frac{\ell_\epsilon}{\gamma_\infty} - \frac{A_\delta}{\gamma_\infty} \geq \frac{\ell_\epsilon}{\gamma_\infty} - \frac{A_\delta}{\gamma_\infty}\omega_a(\ell_\epsilon),$$

where we used $\omega_a(\ell_\epsilon) \geq 1$ in the last inequality. Combining this with (137), we obtain $|\|w_{\tau_\epsilon}\|_\infty - \frac{\ell_\epsilon}{\gamma_\infty}| \leq C_\delta^+ \omega_a(\ell_\epsilon)$ where $C_\delta^+ := \max\{C_{w,\delta}, \frac{A_\delta}{\gamma_\infty}\}$. Therefore, we conclude that

$$\|w_{\tau_\epsilon}\|_\infty = \frac{\ell_\epsilon}{\gamma_\infty} + O_\delta(\omega_a(\ell_\epsilon)). \tag{140}$$

Part 5 Denote $w_\epsilon := w_{\tau_\epsilon}$. Since $\ell_\epsilon \geq e$ and $\delta \in (0,1)$, we have $G_{\tau_\epsilon} \leq \delta\epsilon < \frac{1}{2} = G(0)$. Therefore, we have $w_\epsilon \neq 0$. Define $x_\epsilon := \frac{w_\epsilon}{\|w_\epsilon\|_\infty}$. Fixing any $0 < \epsilon \leq \tilde{\epsilon}_2(\delta)$, the conclusion $0 \leq \gamma_\infty - \hat{\gamma}_\infty(w_\epsilon) \leq C_{\infty,\delta} r_\epsilon$ in (139) holds. The margin is positively homogeneous as $\rho(cw) = \min_i z_i^T(cw) = c\min_i z_i^T w = c\rho(w)$, $c \geq 0$. Therefore, it holds that $\rho(x_\epsilon) = \rho(\frac{w_\epsilon}{\|w_\epsilon\|_\infty}) = \frac{\rho(w_\epsilon)}{\|w_\epsilon\|_\infty} = \hat{\gamma}_\infty(w_\epsilon)$. Therefore, we have

$$0 \leq \gamma_\infty - \rho(x_\epsilon) \leq C_{\infty,\delta} r_\epsilon.$$

From Lemma 26, we know $\mathcal{M}_\infty = \{u : Au \leq b\}$ where

$$A := \begin{pmatrix} I_d \\ -I_d \\ -Z \end{pmatrix}, \qquad b := \begin{pmatrix} \mathbf{1}_d \\ \mathbf{1}_d \\ -\gamma_\infty \mathbf{1}_n \end{pmatrix}.$$

We next compute $\|(Ax_\epsilon - b)_+\|_\infty$. Given $\|x_\epsilon\|_\infty = 1$, it holds that $x_\epsilon[j] - 1 \leq 0$ and $-x_\epsilon[j] - 1 \leq 0$ for all $j \in [d]$. Therefore, we have

$$[x_\epsilon - \mathbf{1}_d]_+ = 0, \qquad [-x_\epsilon - \mathbf{1}_d]_+ = 0.$$

The remaining part is $[-Zx_\epsilon + \gamma_\infty \mathbf{1}_n]_+ = ([\gamma_\infty - z_i^T x_\epsilon]_+)_{i=1}^n$. Therefore, it holds that

$$\|(Ax_\epsilon - b)_+\|_\infty = \max_i [\gamma_\infty - z_i^T x_\epsilon]_+$$

Given $\min_i z_i^T x_\epsilon = \rho(x_\epsilon)$, it holds that

$$\max_i(\gamma_\infty - z_i^T x_\epsilon) = \gamma_\infty - \min_i z_i^T x_\epsilon = \gamma_\infty - \rho(x_\epsilon) \geq 0.$$

Consequently, we have $\max_i(\gamma_\infty - z_i^T x_\epsilon) = \max_i [\gamma_\infty - z_i^T x_\epsilon]_+$. Applying Lemma 26, we have

$$\mathrm{dist}_2(x_\epsilon, \mathcal{M}_\infty) \leq H_{2,\infty}(A) \left\|(Ax_\epsilon - b)_+\right\|_\infty = H_{2,\infty}[\gamma_\infty - \rho(x_\epsilon)] \leq H_{2,\infty} C_{\infty,\delta} r_\epsilon,$$

where we denote $H_{2,\infty} := H_{2,\infty}(A)$ for simplicity. By choosing $u_\epsilon \in \arg\min_{u \in \mathcal{M}_\infty} \|x_\epsilon - u\|_2$, it holds that

$$\|x_\epsilon - u_\epsilon\|_2 \leq H_{2,\infty} C_{\infty,\delta} r_\epsilon.$$

Next, we prove that every $u \in \mathcal{M}_\infty$ satisfies $\|u\|_\infty = 1$. Suppose instead that $c := \|u\|_\infty < 1$. Since $\rho(u) = \gamma_\infty > 0$, we have $c > 0$, and hence $c \in (0,1)$. Define $\tilde{u} := u/c$. It then holds that $\rho(\tilde{u}) = \frac{\rho(u)}{c} = \frac{\gamma_\infty}{c} > \gamma_\infty$, which contradicts the definition of γ_∞. This implies that $\|u_\epsilon\|_\infty = 1$. Given $\|v\|_2^2 \geq \|v\|_\infty^2$ for every vector v, we also have $\|x_\epsilon\|_2 \geq 1$ and $\|u_\epsilon\|_2 \geq 1$.

Define $F(v) := \frac{\rho(v)}{\|v\|_2}, v \neq 0$. We now compare $F(x_\epsilon)$ and $F(u_\epsilon)$. By definition, it holds that

$$F(x_\epsilon) - F(u_\epsilon) = \frac{\rho(x_\epsilon) - \rho(u_\epsilon)}{\|x_\epsilon\|_2} + \rho(u_\epsilon)[\frac{1}{\|x_\epsilon\|_2} - \frac{1}{\|u_\epsilon\|_2}].$$

By Lemma 27, it holds that $|\rho(x_\epsilon) - \rho(u_\epsilon)| \leq R_2 \|x_\epsilon - u_\epsilon\|_2$. For the first term, using this and $\|x_\epsilon\|_2 \geq 1$, we obtain

$$\frac{\rho(x_\epsilon) - \rho(u_\epsilon)}{\|x_\epsilon\|_2} \leq \frac{|\rho(x_\epsilon) - \rho(u_\epsilon)|}{\|x_\epsilon\|_2} \leq R_2 \|x_\epsilon - u_\epsilon\|_2.$$

Using Lemma 25 and $\|x_\epsilon\|_2 \|u_\epsilon\|_2 \geq 1$, it holds that

$$\left|\frac{1}{\|x_\epsilon\|_2} - \frac{1}{\|u_\epsilon\|_2}\right| \leq \|x_\epsilon - u_\epsilon\|_2.$$

Given this and using $\rho(u_\epsilon) = \gamma_\infty > 0$, we have

$$\begin{aligned} \rho(u_\epsilon)\left(\frac{1}{\|x_\epsilon\|_2} - \frac{1}{\|u_\epsilon\|_2}\right) &\leq \gamma_\infty \left|\frac{1}{\|x_\epsilon\|_2} - \frac{1}{\|u_\epsilon\|_2}\right| \\ &\leq \gamma_\infty \|x_\epsilon - u_\epsilon\|_2. \end{aligned}$$

Putting things together, we conclude that

$$F(x_\epsilon) \le F(u_\epsilon) + (R_2+\gamma_\infty)\|x_\epsilon - u_\epsilon\|_2 \le F(u_\epsilon) + (R_2+\gamma_\infty)H_{2,\infty}C_{\infty,\delta}r_\epsilon.$$

Given $u_\epsilon \in \mathcal{M}_\infty$, the definition $\gamma_{2|\infty} = \max_{u\in\mathcal{M}_\infty} \frac{\rho(u)}{\|u\|_2}$ implies that $F(u_\epsilon) = \frac{\rho(u_\epsilon)}{\|u_\epsilon\|_2} \le \gamma_{2|\infty}$. Using this, we conclude that

$$F(x_\epsilon) \le \gamma_{2|\infty} + C_{2,\delta}r_\epsilon, \qquad C_{2,\delta} := (R_2+\gamma_\infty)H_{2,\infty}C_{\infty,\delta}. \tag{141}$$

Given $x_\epsilon = \frac{w_\epsilon}{\|w_\epsilon\|_\infty}$, we have $\rho(x_\epsilon) = \frac{\rho(w_\epsilon)}{\|w_\epsilon\|_\infty}$ and $\|x_\epsilon\|_2 = \frac{\|w_\epsilon\|_2}{\|w_\epsilon\|_\infty}$. Consequently, it holds that $F(x_\epsilon) = \frac{\rho(w_\epsilon)}{\|w_\epsilon\|_2} = \hat{\gamma}_2(w_\epsilon)$. Substituting this into (141), we obtain $\hat{\gamma}_2(w_{T_1^\delta(\epsilon)}) \le \gamma_{2|\infty} + C_{2,\delta}r_\epsilon$. Subtracting both sides from γ_2, it holds that

$$\gamma_2 - \hat{\gamma}_2(w_{T_1^\delta(\epsilon)}) \ge \gamma_2 - \gamma_{2|\infty} - C_{2,\delta}r_\epsilon = \triangle_{\mathrm{geom}} - C_{2,\delta}r_\epsilon,$$

where $\triangle_{\mathrm{geom}} = \gamma_2 - \gamma_{2|\infty}$. This is non-negative. For any $v = \frac{u}{\|u\|_2}$, it holds that $\|v\|_2 = 1$ and $\rho(v) = \frac{\rho(u)}{\|u\|_2} = F(u) \le \gamma_2$. Taking the maximum over $u \in \mathcal{M}_\infty$ leads to $\gamma_{2|\infty} \le \gamma_2$. □

Corollary 30. *In the same setting as Theorem 29, if $\Delta_{\mathrm{geom}} > 0$, then it holds that $\liminf_{\epsilon\downarrow 0}[\gamma_2 - \hat{\gamma}_2(w_{T_1^\delta(\epsilon)})] \ge \Delta_{\mathrm{geom}} > 0$. Moreover, there exists a constant $\hat{\epsilon}_2(\delta) \le \tilde{\epsilon}_2(\delta)$ such that for every $0 < \epsilon \le \hat{\epsilon}_2(\delta)$, we have $\gamma_2 - \hat{\gamma}_2(w_{T_1^\delta(\epsilon)}) \ge \frac{\triangle_{\mathrm{geom}}}{2}$.*

Proof of Corollary 30. Let $\tau_\epsilon := T_1^\delta(\epsilon)$ and $\triangle := \triangle_{\mathrm{geom}} > 0$. From Theorem 29, we have $\gamma_2 - \hat{\gamma}_2(w_{\tau_\epsilon}) \ge \triangle - C_{2,\delta}r_\epsilon$. Taking lower limits and using $r_\epsilon \to 0$, we obtain

$$\liminf_{\epsilon\downarrow 0}[\gamma_2 - \hat{\gamma}_2(w_{\tau_\epsilon})] \ge \liminf_{\epsilon\downarrow 0}[\triangle - C_{2,\delta}r_\epsilon] = \triangle > 0.$$

For the second conclusion, the sufficient conditions on ϵ are

$$0 < \epsilon \le \tilde{\epsilon}_2(\delta), \quad \text{and} \quad C_{2,\delta}r_\epsilon \le \frac{\triangle_{\mathrm{geom}}}{2}.$$

In the case where $\frac{1}{3} < a < 1$, we have $r_\epsilon = (\log\frac{1}{\epsilon})^{-a}$. Therefore, it is sufficient to require $\epsilon \le \exp[-(\frac{2C_{2,\delta}}{\triangle_{\mathrm{geom}}})^{1/a}]$ to ensure $C_{2,\delta}(\log\frac{1}{\epsilon})^{-a} \le \frac{\triangle_{\mathrm{geom}}}{2}$. In the case where $a = 1$, we have $r_\epsilon = \frac{\log\ell_\epsilon}{\ell_\epsilon}$. For $\ell_\epsilon \ge 1$ (ensured via $\epsilon \le \tilde{\epsilon}_2(\delta)$), it holds that $\log\ell_\epsilon \le \sqrt{\ell_\epsilon}$ and consequently $C_{2,\delta}\frac{\log\ell_\epsilon}{\ell_\epsilon} \le \frac{C_{2,\delta}}{\sqrt{\ell_\epsilon}}$. It is therefore sufficient to require $\ell_\epsilon \ge (\frac{2C_{2,\delta}}{\triangle_{\mathrm{geom}}})^2$. Therefore, we can define $L_{3,\delta} := (\frac{2C_{2,\delta}}{\triangle_{\mathrm{geom}}})^{1/a}$ when $\frac{1}{3} < a < 1$ and $L_{3,\delta} := (\frac{2C_{2,\delta}}{\triangle_{\mathrm{geom}}})^2$ when $a = 1$. Then, we set $\hat{\epsilon}_2(\delta) := \min\{\tilde{\epsilon}_2(\delta), e^{-L_{3,\delta}}\}$. □

F.2 Margin Gap at T_1^-

Theorem 31. *Denote $r_\epsilon := \ell_\epsilon^{-a}$ when $a \in (\frac{1}{3}, 1)$ and $r_\epsilon := \frac{\log\ell_\epsilon}{\ell_\epsilon}$ when $a = 1$. For any $\delta, b \in (0,1)$, there exists a constant $\epsilon_2'(\delta, b)$ such that for every $0 < \epsilon \le \epsilon_2'(\delta, b)$, we have*

$$\gamma_\infty - \widehat{\gamma}_\infty(w_{T_1^-(\epsilon;b)}) \le O(r_\epsilon), \qquad \gamma_2 - \widehat{\gamma}_2(w_{T_1^-(\epsilon;b)}) \ge \Delta_{\mathrm{geom}} - O(r_\epsilon).$$

Proof of Theorem 31. Let $\tau_\epsilon^- := T_1^-(\epsilon; b)$, $\tau_\epsilon^\delta := T_1^\delta(\epsilon)$, and $\ell_\epsilon := \log\frac{1}{\epsilon}$. We first specify a sufficient choice of the threshold on ϵ. From Proposition 19, we have

$$L_{t+1} \le L_t - \eta_t\langle g_t, \sigma_\epsilon(g_t)\rangle + C_{\mathrm{md}}\eta_t^2 G_t, \qquad C_{\mathrm{md}} := C_H + C_{\mathrm{ema}}.$$

Define $\kappa_\delta := \frac{\gamma_\infty^2}{R_1+d/\delta}$, $C_D := \sqrt{\frac{1-\beta_1}{1-\beta_2}}$, $K_0 := R_1 C_D$, and $\bar{\eta}_\delta := \min\left\{\eta_0, \frac{\kappa_\delta}{2C_{\mathrm{md}}}, \frac{2}{\kappa_\delta}\right\}$. Let

$$N_\delta := \max\left\{T_*, \left\lceil\left(\frac{\eta_0}{\bar{\eta}_\delta}\right)^{1/a}\right\rceil\right\}.$$

Since $G_{T_*} \ge g_*$ by Proposition 2.1, Lemma 12 gives

$$G_t \ge G_{T_*}\exp\left(-K_0(S_t - S_{T_*})\right) \ge g_*\exp\left(-K_0(S_{N_\delta} - S_{T_*})\right) =: g_{\mathrm{burn},\delta}$$

for every $T_* \le t \le N_\delta$. Set $B_n := \log \frac{2n}{\log 2}$ and

$$L_{\delta,b} := \max\left\{e,\, 2\left[\log \frac{2}{g_{\mathrm{burn},\delta}}\right]_+,\, \left(b + \sqrt{b^2 + 2B_n}\right)^2\right\}.$$

A sufficient choice for $\epsilon_2'(\delta, b)$ is $\epsilon_2'(\delta, b) := \min\left\{\epsilon_0(b,\delta), \tilde{\epsilon}_2(\delta),\, e^{-L_{\delta,b}}\right\}$, where $\tilde{\epsilon}_2(\delta)$ and $\epsilon_0(b,\delta)$ are the thresholds in Theorem 29 and Proposition C.2, respectively. We henceforth fix $0 < \epsilon \le \epsilon_2'(\delta, b)$.

We first lower-bound the norm of $w_{\tau_\epsilon^-}$. Since $\ell_\epsilon \ge L_{\delta,b} \ge e$, we have $\log \ell_\epsilon \le \sqrt{\ell_\epsilon}$. Moreover, the definition of $L_{\delta,b}$ implies $\sqrt{\ell_\epsilon} \ge b + \sqrt{b^2 + 2B_n}$. Consequently, it holds that

$$b \log \ell_\epsilon + B_n \le b\sqrt{\ell_\epsilon} + B_n \le \frac{\ell_\epsilon}{2}. \tag{142}$$

By the definition of τ_ϵ^- and Proposition 2.1, we have $L_{\tau_\epsilon^-} \le 2G_{\tau_\epsilon^-} \le 2\epsilon\ell_\epsilon^b$. The loss–margin bound in Lemma 8 therefore yields

$$\rho(w_{\tau_\epsilon^-}) \ge \log \frac{\log 2}{nL_{\tau_\epsilon^-}} \ge \log \frac{\log 2}{2n\epsilon\ell_\epsilon^b} = \ell_\epsilon - b\log \ell_\epsilon - B_n \ge \frac{\ell_\epsilon}{2}, \tag{143}$$

where the last inequality follows from (142). In particular, $w_{\tau_\epsilon^-} \ne 0$. Since $\rho(w) \le \gamma_\infty \|w\|_\infty$ for every w, we obtain

$$\|w_{\tau_\epsilon^-}\|_\infty \ge \frac{\ell_\epsilon}{2\gamma_\infty}, \qquad \|w_{\tau_\epsilon^-}\|_2 \ge \|w_{\tau_\epsilon^-}\|_\infty \ge \frac{\ell_\epsilon}{2\gamma_\infty}. \tag{144}$$

We next control the displacement between τ_ϵ^- and τ_ϵ^δ. Since $\ell_\epsilon \ge e$ and $b < 1$, we have

$$\epsilon\ell_\epsilon^b = \exp\left(-\ell_\epsilon + b\log \ell_\epsilon\right) \le \exp\left(-\frac{\ell_\epsilon}{2}\right) = \sqrt{\epsilon}.$$

The definition of $L_{\delta,b}$ further gives $\sqrt{\epsilon} \le \frac{g_{\mathrm{burn},\delta}}{2}$. It follows that $G_t \ge g_{\mathrm{burn},\delta} > \epsilon\ell_\epsilon^b$ when $T_* \le t \le N_\delta$ and hence $\tau_\epsilon^- > N_\delta$. Therefore, by the choice of N_δ, we have for every $t \ge \tau_\epsilon^-$ that

$$C_{\mathrm{md}}\eta_t \le \frac{\kappa_\delta}{2}, \qquad \eta_t \le \frac{2}{\kappa_\delta}. \tag{145}$$

Since $\epsilon \le \epsilon_0(b,\delta)$, we have $\tau_\epsilon^- < \tau_\epsilon^\delta$. For $\tau_\epsilon^- \le t < \tau_\epsilon^\delta$, the definition of τ_ϵ^δ gives $G_t > \delta\epsilon$. As in Part 2 of the proof of Theorem 4, Lemmas 7 and 18 give $\langle g_t, \sigma_\epsilon(g_t)\rangle \ge \frac{\gamma_\infty^2}{R_1 + d/\delta} G_t = \kappa_\delta G_t$. Combining this with Equation (44) and (145), we obtain

$$L_{t+1} \le L_t - \eta_t\left(\kappa_\delta - C_{\mathrm{md}}\eta_t\right) G_t \le L_t - \frac{\kappa_\delta}{2}\eta_t G_t \le \left(1 - \frac{\kappa_\delta}{4}\eta_t\right) L_t.$$

Define

$$B_\delta := \frac{4}{\kappa_\delta}\log\frac{2}{\delta} + \eta_0, \qquad q_a := \begin{cases} \dfrac{1}{1-a}, & a \in (1/3, 1), \\ 1, & a = 1, \end{cases} \qquad \mathcal{K}_{a,\delta,b} := \frac{4bq_a}{\kappa_\delta} + B_\delta.$$

Repeating the argument in Part 2 of the proof of Theorem 4 gives

$$S_{\tau_\epsilon^\delta} - S_{\tau_\epsilon^-} \le \frac{4}{\kappa_\delta}\log\frac{2\ell_\epsilon^b}{\delta} + \eta_{\tau_\epsilon^-} \le \frac{4b}{\kappa_\delta}\log\ell_\epsilon + \frac{4}{\kappa_\delta}\log\frac{2}{\delta} + \eta_0 = \frac{4b}{\kappa_\delta}\log\ell_\epsilon + B_\delta.$$

Using the uniform update bound $\|D_t\|_\infty \le C_D$ from Lemma 11, we obtain

$$\|w_{\tau_\epsilon^\delta} - w_{\tau_\epsilon^-}\|_\infty \le \sum_{t=\tau_\epsilon^-}^{\tau_\epsilon^\delta - 1} \eta_t \|D_t\|_\infty \le C_D\left(S_{\tau_\epsilon^\delta} - S_{\tau_\epsilon^-}\right) \le C_D\left(\frac{4b}{\kappa_\delta}\log\ell_\epsilon + B_\delta\right),$$

and hence

$$\|w_{\tau_\epsilon^\delta} - w_{\tau_\epsilon^-}\|_2 \le \sqrt{d}\, C_D \left(\frac{4b}{\kappa_\delta}\log \ell_\epsilon + B_\delta\right).$$

Applying Lemma 28 with $x = w_{\tau_\epsilon^\delta}$ and $y = w_{\tau_\epsilon^-}$, and using (144), gives

$$\begin{aligned}\left|\widehat{\gamma}_\infty(w_{\tau_\epsilon^\delta}) - \widehat{\gamma}_\infty(w_{\tau_\epsilon^-})\right| &\le \frac{2R_1\|w_{\tau_\epsilon^\delta} - w_{\tau_\epsilon^-}\|_\infty}{\|w_{\tau_\epsilon^-}\|_\infty} \le 4R_1\gamma_\infty C_D\left(\frac{4b}{\kappa_\delta}\frac{\log \ell_\epsilon}{\ell_\epsilon} + \frac{B_\delta}{\ell_\epsilon}\right)\\ &\le 4R_1\gamma_\infty C_D \mathcal{K}_{a,\delta,b} r_\epsilon, \end{aligned} \tag{146}$$

and

$$\begin{aligned}\left|\widehat{\gamma}_2(w_{\tau_\epsilon^\delta}) - \widehat{\gamma}_2(w_{\tau_\epsilon^-})\right| &\le \frac{2R_2\|w_{\tau_\epsilon^\delta} - w_{\tau_\epsilon^-}\|_2}{\|w_{\tau_\epsilon^-}\|_2} \le 4R_2\gamma_\infty\sqrt{d}\, C_D\left(\frac{4b}{\kappa_\delta}\frac{\log \ell_\epsilon}{\ell_\epsilon} + \frac{B_\delta}{\ell_\epsilon}\right)\\ &\le 4R_2\gamma_\infty\sqrt{d}\, C_D \mathcal{K}_{a,\delta,b} r_\epsilon. \end{aligned} \tag{147}$$

The last inequalities in (146) and (147) follow from $\frac{1}{\ell_\epsilon} \le r_\epsilon$ and $\frac{\log \ell_\epsilon}{\ell_\epsilon} \le q_a r_\epsilon$. Indeed, when $a \in (1/3, 1)$, we have $\frac{\log \ell_\epsilon}{\ell_\epsilon} \le \frac{1}{1-a}\ell_\epsilon^{-a} = q_a r_\epsilon$ and $\frac{1}{\ell_\epsilon} \le \ell_\epsilon^{-a} = r_\epsilon$; when $a = 1$, we have $\frac{\log \ell_\epsilon}{\ell_\epsilon} = r_\epsilon$ and $\frac{1}{\ell_\epsilon} \le \frac{\log \ell_\epsilon}{\ell_\epsilon} = r_\epsilon$ since $\ell_\epsilon \ge e$.

Finally, because $\epsilon \le \tilde{\epsilon}_2(\delta)$, Theorem 29 gives

$$0 \le \gamma_\infty - \widehat{\gamma}_\infty(w_{\tau_\epsilon^\delta}) \le C_{\infty,\delta} r_\epsilon, \qquad \gamma_2 - \widehat{\gamma}_2(w_{\tau_\epsilon^\delta}) \ge \Delta_{\mathrm{geom}} - C_{2,\delta} r_\epsilon.$$

Combining these bounds with (146) gives

$$\begin{aligned}\gamma_\infty - \widehat{\gamma}_\infty(w_{\tau_\epsilon^-}) &\le \gamma_\infty - \widehat{\gamma}_\infty(w_{\tau_\epsilon^\delta}) + \left|\widehat{\gamma}_\infty(w_{\tau_\epsilon^\delta}) - \widehat{\gamma}_\infty(w_{\tau_\epsilon^-})\right|\\ &\le \left(C_{\infty,\delta} + 4R_1\gamma_\infty C_D \mathcal{K}_{a,\delta,b}\right) r_\epsilon = C^-_{\infty,a,\delta,b} r_\epsilon,\end{aligned}$$

where we let $C^-_{\infty,a,\delta,b} := C_{\infty,\delta} + 4R_1\gamma_\infty C_D \mathcal{K}_{a,\delta,b}$. Similarly, (147) gives

$$\begin{aligned}\gamma_2 - \widehat{\gamma}_2(w_{\tau_\epsilon^-}) &\ge \gamma_2 - \widehat{\gamma}_2(w_{\tau_\epsilon^\delta}) - \left|\widehat{\gamma}_2(w_{\tau_\epsilon^\delta}) - \widehat{\gamma}_2(w_{\tau_\epsilon^-})\right|\\ &\ge \Delta_{\mathrm{geom}} - \left(C_{2,\delta} + 4R_2\gamma_\infty\sqrt{d}\, C_D \mathcal{K}_{a,\delta,b}\right) r_\epsilon = \Delta_{\mathrm{geom}} - C^-_{2,a,\delta,b} r_\epsilon,\end{aligned}$$

where we let $C^-_{2,a,\delta,b} := C_{2,\delta} + 4R_2\gamma_\infty\sqrt{d}\, C_D \mathcal{K}_{a,\delta,b}$. □

F.3 Margin Gap at T_1^+

We first prove Proposition 3.2 and then specialize it to $T_1^+(\epsilon; b)$.

Proof of Proposition 3.2. Denote $\tau := T_1^\delta(\epsilon)$, $H_t := H_t^\delta(\epsilon) = \sum_{s=\tau}^{t-1}\theta_s$, $R_\tau := \|w_\tau\|_\infty$, and $R_t := \|w_t\|_\infty$. Let $\epsilon_\delta^{\mathrm{inv}}$ denote the threshold in Corollary 22. Define $A_\delta = [\log \frac{2n\delta}{\log 2}]_+$ and $\bar{l}_\delta = \max\{e, 2A_\delta\}$. We choose $0 < \epsilon \le \tilde{\epsilon}_3(\delta) := \min\{\tilde{\epsilon}_1(\delta), \tilde{\epsilon}_2(\delta), \epsilon_\delta^{\mathrm{inv}}, e^{-\bar{l}_\delta}\}$ (note that $\tilde{\epsilon}_1(\delta)$ and $\tilde{\epsilon}_2(\delta)$ are the thresholds in the statements of Proposition 3.1 and Theorem 29, respectively). This ensures that the conclusions of Proposition 3.1 and Theorem 29 hold. Summing the post-crossover recursion (11) from τ to $t-1$ gives

$$w_t - w_\tau = -\sum_{s=\tau}^{t-1}\theta_s u_s - \sum_{s=\tau}^{t-1}\theta_s e_s.$$

Given $u_s = \frac{g_s}{\|g_s\|_2}$, it holds that $\|u_s\|_\infty \le \|u_s\|_2 \le 1$ and

$$\|\sum_{s=\tau}^{t-1}\theta_s u_s\|_\infty \le \sum_{s=\tau}^{t-1}\theta_s\|u_s\|_\infty \le \sum_{s=\tau}^{t-1}\theta_s = H_t. \tag{148}$$

We also have that

$$\|\sum_{s=\tau}^{t-1}\theta_s e_s\|_\infty \le \sum_{s=\tau}^{t-1}\theta_s\|e_s\|_\infty \le \sum_{s=\tau}^{t-1}\theta_s\|e_s\|_2 \le E_0. \tag{149}$$

Denoting $\triangle_t := \|w_t - w_\tau\|_\infty$, and combining (148) and (149) together, we obtain

$$\triangle_t = \|w_t - w_\tau\|_\infty \le H_t + E_0, \qquad t \ge \tau. \tag{150}$$

From Corollary 22, we know $G_t \le L_t \le 2\delta\epsilon, t \ge \tau$. Given $t \ge \tau \ge T_*$, Proposition 2.1 guarantees $L_t \le \frac{\log 2}{n}$. By Lemma 8, we also have $L_t \ge \frac{\log 2}{n} e^{-\rho(w_t)}$. Combining them together, we obtain $\frac{\log 2}{n} e^{-\rho(w_t)} \le 2\delta\epsilon$. Taking logarithms and rearranging gives $\rho(w_t) \ge \log \frac{\log 2}{2n\delta\epsilon} = \ell_\epsilon - \log \frac{2n\delta}{\log 2} \ge \ell_\epsilon - A_\delta$. The condition on ϵ ensures that $\ell_\epsilon \ge \bar{\ell}_\delta \ge \max\{e, 2A_\delta\}$. Consequently, we obtain

$$\rho(w_t) \ge \frac{1}{2}\ell_\epsilon > 0, \qquad t \ge \tau. \tag{151}$$

Since $\rho(0) = 0$, (151) implies that $w_t \ne 0$ for every $t \ge \tau$.

For any nonzero w, it holds that $\rho(w) = \|w\|_\infty \rho(\frac{w}{\|w\|_\infty}) \le \gamma_\infty \|w\|_\infty$. Combining this and (151), we have

$$R_\tau = \|w_\tau\|_\infty \ge \frac{\ell_\epsilon}{2\gamma_\infty}. \tag{152}$$

Applying Lemma 28 with $x = w_t$ and $y = w_\tau$ and using the inequalities (150) and (152), it holds that

$$|\hat{\gamma}_\infty(w_t) - \hat{\gamma}_\infty(w_\tau)| \le \frac{2R_1\|w_t - w_\tau\|_\infty}{R_\tau} \le \frac{2R_1(H_t + E_0)}{R_\tau} \le 4R_1\gamma_\infty \frac{H_t + E_0}{\ell_\epsilon}.$$

Since $\epsilon \le \tilde{\epsilon}_2(\delta)$, Theorem 29 gives $0 \le \gamma_\infty - \hat{\gamma}_\infty(w_\tau) \le C_{\infty,\delta} r_\epsilon$. Consequently, it holds that

$$\gamma_\infty - \hat{\gamma}_\infty(w_t) \le C_{\infty,\delta} r_\epsilon + 4R_1\gamma_\infty \frac{H_t + E_0}{\ell_\epsilon}.$$

The left-hand side is nonnegative because $\hat{\gamma}_\infty(w_t) = \rho(\frac{w_t}{\|w_t\|_\infty}) \le \gamma_\infty$. Define $C_{\mathrm{p},\delta} := \max\{C_{\infty,\delta}, 4R_1\gamma_\infty\}$. We conclude that

$$0 \le \gamma_\infty - \hat{\gamma}_\infty(w_t) \le C_{\mathrm{p},\delta}(r_\epsilon + \frac{H_t + E_0}{\ell_\epsilon}), \qquad t \ge \tau. \tag{153}$$

We next prove the second inequality. Using (148) and (149), we have

$$\|w_t - w_\tau\|_2 \le H_t + E_0, \qquad t \ge \tau. \tag{154}$$

At $t = \tau$, (151) gives $\rho(w_\tau) \ge \frac{\ell_\epsilon}{2}$. Since $\rho(w) \le \gamma_2\|w\|_2$ for every w, it holds that

$$\|w_\tau\|_2 \ge \frac{\rho(w_\tau)}{\gamma_2} \ge \frac{\ell_\epsilon}{2\gamma_2}. \tag{155}$$

Applying Lemma 28 with $x = w_t$ and $y = w_\tau$, and using (154) and (155), we obtain

$$|\widehat{\gamma}_2(w_t) - \widehat{\gamma}_2(w_\tau)| \le \frac{2R_2\|w_t - w_\tau\|_2}{\|w_\tau\|_2} \le 4R_2\gamma_2 \frac{H_t + E_0}{\ell_\epsilon}. \tag{156}$$

Because $\epsilon \le \tilde{\epsilon}_2(\delta)$, Theorem 29 gives $\gamma_2 - \widehat{\gamma}_2(w_\tau) \ge \Delta_{\mathrm{geom}} - C_{2,\delta} r_\epsilon$. Combining this bound with (156), we obtain

$$\begin{aligned}\gamma_2 - \widehat{\gamma}_2(w_t) &\ge \gamma_2 - \widehat{\gamma}_2(w_\tau) - |\widehat{\gamma}_2(w_t) - \widehat{\gamma}_2(w_\tau)| \\ &\ge \Delta_{\mathrm{geom}} - C_{2,\delta} r_\epsilon - 4R_2\gamma_2 \frac{H_t + E_0}{\ell_\epsilon}.\end{aligned} \tag{157}$$

The definition of $\tilde{\epsilon}_3(\delta)$ ensures that $\ell_\epsilon \geq e$. Hence, we have $\frac{1}{\ell_\epsilon} \leq r_\epsilon$ for both $a \in (1/3, 1)$ and $a = 1$. It follows from (157) that

$$\begin{aligned}\gamma_2 - \widehat{\gamma}_2(w_t) &\geq \Delta_{\mathrm{geom}} - (C_{2,\delta} + 4R_2\gamma_2 E_0)\, r_\epsilon - 4R_2\gamma_2 \frac{H_t}{\ell_\epsilon} \\ &\geq \Delta_{\mathrm{geom}} - C_{2,\delta}^{\mathrm{post}} \left(r_\epsilon + \frac{H_t^\delta(\epsilon)}{\ell_\epsilon} \right), \qquad t \geq T_1^\delta(\epsilon),\end{aligned}$$

where we let $C_{2,\delta}^{\mathrm{post}} := \max\{C_{2,\delta} + 4R_2\gamma_2 E_0,\, 4R_2\gamma_2\}$. □

Corollary 32. *Denote $r_\epsilon := \ell_\epsilon^{-a}$ when $a \in (\frac{1}{3}, 1)$ and $r_\epsilon := \frac{\log \ell_\epsilon}{\ell_\epsilon}$ when $a = 1$. For any $\delta, b \in (0,1)$, there exists a constant $\epsilon_2''(\delta, b)$ such that for every $0 < \epsilon \leq \epsilon_2''(\delta, b)$, we have*

$$\gamma_\infty - \widehat{\gamma}_\infty(w_{T_1^+(\epsilon;b)}) \leq O(r_\epsilon), \qquad \gamma_2 - \widehat{\gamma}_2(w_{T_1^+(\epsilon;b)}) \geq \Delta_{\mathrm{geom}} - O(r_\epsilon).$$

Proof. Let $\ell_\epsilon := \log \frac{1}{\epsilon}$, $C_D := \sqrt{\frac{1-\beta_1}{1-\beta_2}}$ and $c_\delta := \frac{\gamma_2^2}{16(1+2R_1\delta)}$. Set $\epsilon_2''(\delta, b) := \min\{\tilde{\epsilon}_3(\delta), \epsilon_0(\delta, b)\}$ where $\epsilon_0(\delta, b)$ is the threshold in Proposition C.2. Fix $0 < \epsilon \leq \epsilon_2''(\delta, b)$ and let $\tau := T_1^\delta(\epsilon)$, $t_+ := T_1^+(\epsilon; b)$, $V_+ := V_{t_+}^\delta(\epsilon)$, and $H_+ := H_{t_+}^\delta(\epsilon)$. Since $\epsilon \leq \epsilon_0(\delta, b)$, we have $\tau < t_+$ by Proposition C.2. Moreover, the condition $\epsilon \leq \tilde{\epsilon}_3(\delta)$ ensures that the conclusions of Proposition 3.1 apply from time τ onward. In particular, we have $\tau > T_*$ and $K_0\eta_{\tau-1} \leq \log 2$ where $K_0 := R_1 C_D$. By the minimality of t_+ and Lemma 12, it holds that

$$L_{t_+} \geq G_{t_+} \geq \exp(-K_0\eta_{t_+-1}) G_{t_+-1} \geq e^{-\log 2} G_{t_+-1} > \frac{\epsilon}{2}\ell_\epsilon^{-b}.$$

Consequently, (111) gives

$$c_\delta \frac{V_+}{\epsilon} \leq \frac{1}{L_{t_+}} - \frac{1}{L_\tau} \leq \frac{1}{L_{t_+}} < \frac{2}{\epsilon}\ell_\epsilon^b.$$

Thus, we have $V_+ \leq \frac{2}{c_\delta}\ell_\epsilon^b$. Denote $V_t := \sum_{s=\tau}^{t-1} \eta_s$ and $H_t := \sum_{s=\tau}^{t-1} \theta_s$. By (117), we know $\frac{G_s}{\epsilon} \leq \frac{C_L}{1+V_s}$ where $C_L = 1/\min\{\frac{1}{2}, \frac{\gamma_2^2}{16(1+2R_1)}\}$. Hence, it holds that

$$\theta_s = \frac{\eta_s \|g_s\|_2}{\epsilon} \leq R_2\eta_s \frac{G_s}{\epsilon} \leq R_2 C_L \frac{\eta_s}{1+V_s}.$$

Consequently, we have $H_t \leq R_2 C_L \sum_{s=\tau}^{t-1} \frac{\eta_s}{1+V_s}$ for any $t > \tau$. Setting $x_s := \frac{\eta_s}{1+V_s}$, we obtain $\log(1+V_{s+1}) - \log(1+V_s) = \log(1+x_s)$ and $0 \leq x_s \leq \eta_s \leq \eta_0$. For every $x \geq 0$, we have $\log(1+x) \geq \frac{x}{1+x}$. This leads to

$$\frac{\eta_s}{1+V_s} = x_s \leq (1+\eta_0)\left[\log(1+V_{s+1}) - \log(1+V_s)\right].$$

Summing this inequality gives

$$\begin{aligned}\sum_{s=\tau}^{t-1} \frac{\eta_s}{1+V_s} &\leq (1+\eta_0) \sum_{s=\tau}^{t-1} \left[\log(1+V_{s+1}) - \log(1+V_s)\right] \\ &= (1+\eta_0)\left[\log(1+V_t) - \log(1+V_\tau)\right].\end{aligned}$$

Because $V_\tau = 0$, it follows that $\sum_{s=\tau}^{t-1} \frac{\eta_s}{1+V_s} \leq (1+\eta_0)\log(1+V_t)$. Consequently, we obtain $H_t \leq R_2 C_L(1+\eta_0)\log(1+V_t) = C_{\mathrm{HV}} \log(1+V_t)$ for any $t > \tau$ where $C_{\mathrm{HV}} := R_2 C_L(1+\eta_0)$. Since $V_+ \leq \frac{2}{c_\delta}\ell_\epsilon^b$ and $t_+ > \tau$, it holds that

$$H_+ \leq C_{\mathrm{HV}} \log\left(1 + \frac{2}{c_\delta}\ell_\epsilon^b\right) \leq C_{\mathrm{HV}}\left[\log\left(1 + \frac{2}{c_\delta}\right) + b\log \ell_\epsilon\right].$$

Define $q_a := \dfrac{1}{1-a}$ when $a \in (1/3, 1)$ and $q_a := 1$ when $a = 1$. In both cases, we have $\frac{\log \ell_\epsilon}{\ell_\epsilon} \leq q_a r_\epsilon$. We also $1/\ell_\epsilon \leq r_\epsilon$. Consequently, we obtain

$$\frac{H_+}{\ell_\epsilon} \leq K_{H,a,\delta,b}^+ r_\epsilon, \qquad K_{H,a,\delta,b}^+ := C_{\mathrm{HV}} \log\left(1 + \frac{2}{c_\delta}\right) + bq_a.$$

Applying Proposition 3.2 at $t = t_+$ yields

$$\gamma_\infty - \widehat{\gamma}_\infty(w_{t_+}) \le C_{\infty,\delta} r_\epsilon + 4R_1\gamma_\infty \frac{H_+ + E_0}{\ell_\epsilon}$$
$$\le \left[C_{\infty,\delta} + 4R_1\gamma_\infty\left(E_0 + K^+_{H,a,\delta,b}\right)\right] r_\epsilon = C^+_{\infty,a,\delta,b} r_\epsilon,$$

where $C^+_{\infty,a,\delta,b} := C_{\infty,\delta} + 4R_1\gamma_\infty\left(E_0 + K^+_{H,a,\delta,b}\right)$. Similarly, the second inequality of Proposition 3.2 gives

$$\gamma_2 - \widehat{\gamma}_2(w_{t_+}) \ge \Delta_{\text{geom}} - C_{2,\delta} r_\epsilon - 4R_2\gamma_2 \frac{H_+ + E_0}{\ell_\epsilon}$$
$$\ge \Delta_{\text{geom}} - \left[C_{2,\delta} + 4R_2\gamma_2\left(E_0 + K^+_{H,a,\delta,b}\right)\right] r_\epsilon = \Delta_{\text{geom}} - C^+_{2,a,\delta,b} r_\epsilon,$$

where $C^+_{2,a,\delta,b} := C_{2,\delta} + 4R_2\gamma_2\left(E_0 + K^+_{H,a,\delta,b}\right)$. □

Proof of Theorem 3. We can let $\tilde{\epsilon}_2(\delta, b) := \min\{\tilde{\epsilon}_2(\delta), \epsilon'_2(\delta, b), \epsilon''_2(\delta, b)\}$, where $\tilde{\epsilon}_2(\delta)$, $\epsilon'_2(\delta, b)$, and $\epsilon''_2(\delta, b)\}$ are the thresholds in Theorem 29, Theorem 31, and Corollary 32, respectively. The conclusion then follows.

□

G Adam's Euclidean Optimality

We first provide a proof sketch of Theorem 5. Let $\tau := T_1^\delta(\epsilon)$, $V_t := \sum_{s=\tau}^{t-1} \eta_s$, and $H_t := \sum_{s=\tau}^{t-1} \theta_s$. Recall the post-crossover update can be written as

$$w_{s+1} = w_s - \theta_s d_s, \qquad d_s := u_s + e_s, \qquad u_s := \frac{g_s}{\|g_s\|_2}, \tag{158}$$

where $\theta_s := \frac{\eta_s \|g_s\|_2}{\epsilon}$ and $\|u_s\|_2 = 1$. Using Taylor's theorem and following the same approach as Lemma 13, we show

$$L_{s+1} \le L_s - \theta_s \langle g_s, d_s\rangle + C_H L_s \left(\theta_s \|d_s\|_2\right)^2,$$

for some constant C_H. From this, we further obtain $\frac{L_{s+1}}{L_s} \le 1 - \gamma_2\theta_s + r_s$, where

$$r_s := \gamma_2\theta_s\left(1 - \frac{G_s}{L_s}\right) + R_2\theta_s\|e_s\|_2 + 2C_H\left[\theta_s^2 + \left(\theta_s\|e_s\|_2\right)^2\right].$$

Based on Proposition 3.1, we can show $\sum_\tau^\infty \theta_s^2 \le K_\theta^2$ for some constant K_θ, $\sum_{s=\tau}^\infty \theta_s\|e_s\|_2 \le E_0$, $\sum_{s=\tau}^\infty (\theta_s\|e_s\|_2)^2 \le E_0^2$, and $\sum_{s=\tau}^\infty \theta_s(1 - \frac{G_s}{L_s}) \le K_p$ for some constant K_p. Consequently, we have $\sum_{s=\tau}^\infty r_s \le A_0 < \infty$ for some constant A_0. Algebraic manipulation leads to $L_t \le L_\tau \exp(A_0 - \gamma_2 H_t)$. Hence, the loss decreases exponentially in H_t. Using Proposition 2.1 and Lemma 8, it follows that $\rho(w_t) \ge \gamma_2 H_t - A_0$. On the other hand, unrolling the update gives $\|w_t\|_2 \le \|w_\tau\|_2 + H_t + E_0$. Combining them together (with a separate trivial treatment of the case $H_t < A_0/\gamma_2$), we obtain

$$\gamma_2 - \widehat{\gamma}_2(w_t) \le C_\gamma \frac{1 + \|w_\tau\|_2}{1 + H_t},$$

where C_γ is some constant. This proves part (i) of Theorem 5. For part (ii), we further denote $\Delta := \Delta_{\text{geom}}$, $\tau_\zeta := \tau_\zeta^\delta(\epsilon)$, $R_\tau := \|w_\tau\|_2$. By Proposition 3.1, we have $H_t \asymp_\delta \log(1 + V_t)$. Since $V_t \to \infty$, it holds that $H_t \to \infty$. Define $\overline{\tau}_\zeta := \inf\left\{t \ge \tau : H_t \ge \frac{C_\gamma}{\zeta}(1 + R_\tau)\right\}$. Part (i) implies that $\gamma_2 - \widehat{\gamma}_2\left(w_{\overline{\tau}_\zeta}\right) \le C_\gamma \frac{1+R_\tau}{1+H_{\overline{\tau}_\zeta}} \le \zeta$. Therefore, we have $\tau_\zeta \le \overline{\tau}_\zeta$. By the minimality of $\overline{\tau}_\zeta$, it holds that $H_{\overline{\tau}_\zeta - 1} < \frac{C_\gamma}{\zeta}(1 + R_\tau)$. Since $\theta_s \le K_\theta$, we obtain

$$H_{\tau_\zeta} \le H_{\overline{\tau}_\zeta} = H_{\overline{\tau}_\zeta - 1} + \theta_{\overline{\tau}_\zeta - 1} < \frac{C_\gamma}{\zeta}(1 + R_\tau) + K_\theta \le \left(\frac{C_\gamma}{\zeta} + K_\theta\right)(1 + R_\tau).$$

This proves the upper bound in part (ii). For the lower bound, Theorem 29 gives $\gamma_2 - \widehat{\gamma}_2(w_\tau) \geq \Delta - C_{2,\delta} r_\epsilon$. Since $r_\epsilon \to 0$, we choose ϵ sufficiently small such that $C_{2,\delta} r_\epsilon \leq \frac{\Delta-\zeta}{2}$. Then, we have

$$\gamma_2 - \widehat{\gamma}_2(w_\tau) \geq \Delta - \frac{\Delta-\zeta}{2} = \frac{\Delta+\zeta}{2} > \zeta.$$

Consequently, it holds that $\tau_\zeta > \tau$. At time τ_ζ, we also have $\gamma_2 - \widehat{\gamma}_2(w_{\tau_\zeta}) \leq \zeta$. Therefore, combining them together gives

$$\widehat{\gamma}_2(w_{\tau_\zeta}) - \widehat{\gamma}_2(w_\tau) = [\gamma_2 - \widehat{\gamma}_2(w_\tau)] - [\gamma_2 - \widehat{\gamma}_2(w_{\tau_\zeta})] \geq \frac{\Delta+\zeta}{2} - \zeta = \frac{\Delta-\zeta}{2}.$$

Applying Lemma 28 with $x = w_{\tau_\zeta}$ and $y = w_\tau$, we obtain $\frac{\Delta-\zeta}{2} \leq \frac{2R_2\|w_{\tau_\zeta} - w_\tau\|_2}{R_\tau}$. Moreover, unrolling the post-crossover update gives $\|w_{\tau_\zeta} - w_\tau\|_2 \leq H_{\tau_\zeta} + E_0$. Consequently, it holds that $H_{\tau_\zeta} \geq \frac{\Delta-\zeta}{4R_2} R_\tau - E_0$. By (155), we know $R_\tau \to \infty$ as $\epsilon \downarrow 0$. Hence, after decreasing ϵ, we can ensure that $E_0 \leq \frac{\Delta-\zeta}{8R_2} R_\tau$. Together, they imply that $H_{\tau_\zeta} \geq \frac{\Delta-\zeta}{8R_2} R_\tau$. Therefore, we can set $c_\zeta := \frac{\Delta-\zeta}{8R_2}$ and $C_\zeta := \max\left\{\frac{C_\gamma}{\zeta} + K_\theta,\, c_\zeta\right\}$, and it holds that $c_\zeta\|w_\tau\|_2 \leq H_{\tau_\zeta} \leq C_\zeta\big(1 + \|w_\tau\|_2\big)$.

Proof of Theorem 5. Denote $\tau := T_1^\delta(\epsilon)$ and $H_t := H_t^\delta(\epsilon) = \sum_{s=\tau}^{t-1} \theta_s$. We first prove part (i). Set $\tilde{\epsilon}_4^{(i)}(\delta) := \min\{\epsilon_\delta^{\mathrm{inv}}, \frac{\log 2}{4\delta}\}$, where $\epsilon_\delta^{\mathrm{inv}}$ is the constant appearing in Corollary 22. We choose ϵ such that $0 < \epsilon \leq \tilde{\epsilon}_4^{(i)}(\delta)$. The condition $0 < \epsilon \leq \epsilon_\delta^{\mathrm{inv}}$ ensures that $G_t \leq L_t \leq 2\delta\epsilon,\ t \geq \tau$. The extra condition $\epsilon \leq \frac{\log 2}{4\delta}$ implies that $G_t \leq L_t \leq 2\delta\epsilon \leq \frac{\log 2}{2} < \frac{1}{2},\ t \geq \tau$. Since $G(0) = \frac{1}{2}$, this guarantees that $w_t \neq 0,\ t \geq \tau$. Therefore, $\hat{\gamma}_2(w_t)$ is well-defined.

We know from the proof of Proposition 3.1 ((117) and (115)) that

$$L_s - L_{s+1} \geq c_* \frac{\eta_s}{\epsilon} L_s^2, \qquad \frac{G_s}{\epsilon} \leq \frac{C_L}{1+V_s}, \qquad s \geq \tau \tag{159}$$

where $C_L = \frac{1}{\min\{1/2, c_*\}}$ and $c_* = \frac{\gamma_2^2}{16(1+2R_1)}$. Using this and $\|g_s\|_2 \leq R_2 G_s$, it holds that

$$\theta_s = \frac{\eta_s\|g_s\|_2}{\epsilon} \leq R_2\eta_s \frac{G_s}{\epsilon} \leq R_2 C_L \frac{\eta_s}{1+V_s}.$$

Given this, we can further obtain

$$\sum_{s=\tau}^{\infty} \theta_s^2 \leq R_2^2 C_L^2 \sum_{s=\tau}^{\infty} \frac{\eta_s^2}{(1+V_s)^2} \leq R_2^2 C_L^2 \sum_{s=\tau}^{\infty} \frac{\eta_s^2}{1+V_s}.$$

Lemma 24 gives $\sum_{s=\tau}^{\infty} \frac{\eta_s^2}{1+V_s} \leq Q_\eta$ where $Q_\eta = \eta_0^2 + \frac{1}{a}\max\{\eta_0^2, \eta_0(1-a)\}$. Therefore, it holds that

$$\sum_{s=\tau}^{\infty} \theta_s^2 \leq K_\theta^2, \qquad K_\theta = R_2 C_L \sqrt{Q_\eta}. \tag{160}$$

In particular, it holds that $\sup_{s\geq\tau} \theta_s \leq K_\theta$.

From Proposition 3.1, we know

$$\sum_{s=\tau}^{\infty} \theta_s \|e_s\|_2 \leq E_0, \qquad E_0 = C_{\mathrm{ema}} C_L Q_\eta + \frac{2R_1R_2}{c_*}. \tag{161}$$

Let $d_s = u_s + e_s$. Since $\|u_s\|_2 = 1$, it holds that $\theta_s\|d_s\|_2 \leq \theta_s + \theta_s\|e_s\|_2$. Consequently, it holds that

$$\sup_{s\geq\tau} \theta_s\|d_s\|_2 \leq K_\theta + E_0 =: B_0 < \infty.$$

Define $b_s := \theta_s\|e_s\|_2 \geq 0$. Since $\sum_{s=\tau}^{\infty} b_s \leq E_0$, we have

$$\sum_{s=\tau}^{\infty} (\theta_s\|e_s\|_2)^2 = \sum_{s=\tau}^{\infty} b_s^2 \leq (\sum_{s=\tau}^{\infty} b_s)^2 \leq E_0^2. \tag{162}$$

Next, we bound the term $\sum_{s=\tau}^{\infty}\theta_s(1-\frac{G_s}{L_s})$. We first bound $\sum_{s=\tau}^{\infty}\theta_s L_s$. From $\theta_s L_s = \frac{\eta_s\|g_s\|_2}{\epsilon}L_s \le R_2\frac{\eta_s}{\epsilon}G_s L_s \le R_2\frac{\eta_s}{\epsilon}L_s^2$ and (159), we obtain $\theta_s L_s \le \frac{R_2}{c_*}(L_s - L_{s+1})$. Consequently, it holds that

$$\sum_{s=\tau}^{\infty}\theta_s L_s \le \frac{R_2}{c_*}\sum_{s=\tau}^{\infty}(L_s - L_{s+1}) \le \frac{R_2 L_\tau}{c_*}.$$

Since $\tau \ge T_*$, we know $L_\tau \le \frac{\log 2}{n}$ from Proposition 2.1. Lemma 8 further gives $0 \le 1-\frac{G_s}{L_s} \le \frac{nL_s}{2}$. Based on these inequalities, we obtain

$$\sum_{s=\tau}^{\infty}\theta_s(1-\frac{G_s}{L_s}) \le \frac{n}{2}\sum_{s=\tau}^{\infty}\theta_s L_s \le \frac{R_2\log 2}{2c_*} =: K_p. \tag{163}$$

For $s \ge \tau$, the update is $w_{s+1} = w_s - \theta_s d_s$. For $\xi \in [0,1]$, we define $w_{s,\xi} = w_s - \xi\theta_s d_s$. From Lemma 13, we know the Hessian of the logistic loss is $\nabla^2 L(w) = \frac{1}{n}\sum_{i=1}^n \ell''(z_i^T w)z_i z_i^T$ with $0 \le \ell''(r) \le -\ell'(r)$. This implies that

$$d_s^T\nabla^2 L(w_{s,\xi})d_s = \frac{1}{n}\sum_{i=1}^n \ell''(z_i^T w_{s,\xi})(z_i^T d_s)^2 \le \frac{1}{n}\sum_{i=1}^n a_i(w_{s,\xi})(z_i^T d_s)^2.$$

For every sample i, it holds that $|z_i^T(w_{s,\xi}-w_s)| \le \|z_i\|_2\theta_s\|d_s\|_2 \le R_2 B_0$. From Lemma 10, we have $\frac{a_i(w_{s,\xi})}{a_i(w_s)} = \frac{-\ell'(z_i^T w_{s,\xi})}{-\ell'(z_i^T w_s)} \le \exp(|z_i^T w_{s,\xi} - z_i^T w_s|) \le e^{R_2 B_0}$. Consequently, it holds that

$$G(w_{s,\xi}) = \frac{1}{n}\sum_{i=1}^n a_i(w_{s,\xi}) \le e^{R_2 B_0}\frac{1}{n}\sum_{i=1}^n a_i(w_s) = e^{R_2 B_0}G_s.$$

Moreover, we have $(z_i^T d_s)^2 \le R_2^2\|d_s\|_2^2$ given $|z_i^T d_s| \le \|z_i\|_2\|d_s\|_2 \le R_2\|d_s\|_2$. Putting things together, we obtain

$$d_s^T\nabla^2 L(w_{s,\xi})d_s \le \frac{1}{n}\sum_{i=1}^n a_i(w_{s,\xi})R_2^2\|d_s\|_2^2 = R_2^2 G(w_{s,\xi})\|d_s\|_2^2 \le e^{R_2 B_0}R_2^2 G_s\|d_s\|_2^2 \le e^{R_2 B_0}R_2^2 L_s\|d_s\|_2^2,$$

where we used $G_s \le L_s$. Following the same approach as Lemma 13, we have

$$\begin{aligned}
L_{s+1} &= L_s - \theta_s\langle g_s, d_s\rangle + \theta_s^2\int_0^1 (1-\xi)d_s^T\nabla^2 L(w_{s,\xi})d_s d\xi \\
&\le L_s - \theta_s\langle g_s, d_s\rangle + \theta_s^2\int_0^1 (1-\xi)e^{R_2 B_0}R_2^2 L_s\|d_s\|_2^2 d\xi \\
&= L_s - \theta_s\langle g_s, d_s\rangle + \frac{1}{2}e^{R_2 B_0}R_2^2 L_s\theta_s^2\|d_s\|_2^2 \\
&= L_s - \theta_s\langle g_s, d_s\rangle + C_H L_s(\theta_s\|d_s\|_2)^2,
\end{aligned}$$

where $C_H = \frac{1}{2}R_2^2 e^{R_2 B_0}$.

Next, we focus on the first-order term. Given $\langle g_s, d_s\rangle = \langle g_s, u_s\rangle + \langle g_s, e_s\rangle = \|g_s\|_2 + \langle g_s, e_s\rangle$, it holds that $-\theta_s\langle g_s, d_s\rangle \le -\theta_s\|g_s\|_2 + \theta_s\|g_s\|_2\|e_s\|_2$. From Lemma 7, we know $\gamma_2 G_s \le \|g_s\|_2 \le R_2 G_s$. This implies

$$-\theta_s\|g_s\|_2 \le -\gamma_2\theta_s G_s = -\gamma_2\theta_s L_s + \gamma_2\theta_s(L_s - G_s), \qquad \theta_s\|g_s\|_2\|e_s\|_2 \le R_2\theta_s L_s\|e_s\|_2.$$

Substituting these bounds into above, we obtain

$$L_{s+1} \le L_s\big[1 - \gamma_2\theta_s + \gamma_2\theta_s(1-\frac{G_s}{L_s}) + R_2\theta_s\|e_s\|_2 + C_H(\theta_s\|d_s\|_2)^2\big]. \tag{164}$$

Since $\theta_s\|d_s\|_2 \le \theta_s + \theta_s\|e_s\|_2$, we have $(\theta_s\|d_s\|_2)^2 \le 2\theta_s^2 + 2(\theta_s\|e_s\|_2)^2$. Define

$$r_s := \gamma_2\theta_s(1-\frac{G_s}{L_s}) + R_2\theta_s\|e_s\|_2 + 2C_H[\theta_s^2 + (\theta_s\|e_s\|_2)^2] \tag{165}$$

Then, (164) becomes

$$\frac{L_{s+1}}{L_s} \leq 1 - \gamma_2\theta_s + r_s. \tag{166}$$

Because the left side is positive, the right side of (166) is also positive. Using $\log(1+x) \leq x$ for $x > -1$, we obtain

$$\log \frac{L_{s+1}}{L_s} \leq \log(1 - \gamma_2\theta_s + r_s) \leq -\gamma_2\theta_s + r_s.$$

Equivalently, we have

$$\log \frac{L_s}{L_{s+1}} \geq \gamma_2\theta_s - r_s. \tag{167}$$

Using (160), (161), (162), and (163), we obtain from (165) after summing that

$$\sum_{s=\tau}^{\infty} r_s \leq \gamma_2 K_p + R_2 E_0 + 2C_H(K_\theta^2 + E_0^2).$$

Define $A_0 := \gamma_2 K_p + R_2 E_0 + 2C_H(K_\theta^2 + E_0^2)$. Summing (167) from $s = \tau$ to $t-1$ gives

$$\log \frac{L_\tau}{L_t} \geq \gamma_2 \sum_{s=\tau}^{t-1} \theta_s - \sum_{s=\tau}^{t-1} r_s \geq \gamma_2 H_t - A_0.$$

Therefore, it holds that

$$L_t \leq L_\tau \exp(A_0 - \gamma_2 H_t), \qquad t \geq \tau. \tag{168}$$

For every $t \geq \tau \geq T_*$, Proposition 2.1 gives $L_t \leq \frac{\log 2}{n}$. Given this, Lemma 8 gives $L_t \geq \frac{\log 2}{n} e^{-\rho(w_t)}$. Denote $c_{\log} := \frac{\log 2}{n}$. Then, it holds that $\rho(w_t) \geq \log \frac{c_{\log}}{L_t}$. Using (168), we obtain for $t \geq \tau$ that

$$\begin{aligned} \rho(w_t) &\geq \log \frac{c_{\log}}{L_\tau} + \gamma_2 H_t - A_0 \\ &\geq \gamma_2 H_t - A_0, \qquad t \geq \tau, \end{aligned} \tag{169}$$

where we used $\log \frac{c_{\log}}{L_\tau} \geq 0$ as $L_\tau \leq c_{\log}$.

Unrolling (158) gives $w_t = w_\tau - \sum_{s=\tau}^{t-1} \theta_s u_s - \sum_{s=\tau}^{t-1} \theta_s e_s$. Since $\|u_s\|_2 = 1$, we have $\|\sum_{s=\tau}^{t-1} \theta_s u_s\|_2 \leq \sum_{s=\tau}^{t-1} \theta_s = H_t$. Moreover, (161) gives $\|\sum_{s=\tau}^{t-1} \theta_s e_s\|_2 \leq \sum_{s=\tau}^{t-1} \theta_s \|e_s\|_2 \leq E_0$. Therefore, we have

$$\|w_t\|_2 \leq H_t + \|w_\tau\|_2 + E_0 \leq H_t + B_\tau, \tag{170}$$

where $B_\tau := \|w_\tau\|_2 + E_0 + 1$.

Next, we split the analysis into two cases. We first note that $\rho(w_t) \geq 0$, $t \geq \tau$ by Lemma 8. Also, by the definition of γ_2, it holds that $\hat{\gamma}_2(w_t) = \rho(\frac{w_t}{\|w_t\|_2}) \leq \gamma_2$. Consequently, it holds that $0 \leq \gamma_2 - \hat{\gamma}_2(w_t) \leq \gamma_2$. We first consider the case where $H_t \geq \frac{A_0}{\gamma_2}$. It then holds that $\gamma_2 H_t - A_0 \geq 0$. Using (169) and (170), we have $\hat{\gamma}_2(w_t) = \frac{\rho(w_t)}{\|w_t\|_2} \geq \frac{\gamma_2 H_t - A_0}{H_t + B_\tau}$. Consequently, it holds that

$$\gamma_2 - \hat{\gamma}_2(w_t) \leq \gamma_2 - \frac{\gamma_2 H_t - A_0}{H_t + B_\tau} = \frac{\gamma_2 B_\tau + A_0}{H_t + B_\tau}.$$

Because $B_\tau \geq 1$, we have $H_t + B_\tau \geq 1 + H_t$. Therefore, it holds that $\gamma_2 - \hat{\gamma}_2(w_t) \leq \frac{\gamma_2 B_\tau + A_0}{1 + H_t}$. Moreover, we note that

$$\gamma_2 B_\tau + A_0 = \gamma_2(1 + \|w_\tau\|_2 + E_0) + A_0 \leq [\gamma_2(1 + E_0) + A_0 + \gamma_2](1 + \|w_\tau\|_2).$$

If we define $C_\gamma = \gamma_2(1 + E_0) + A_0 + \gamma_2$, then it holds that $\gamma_2 B_\tau + A_0 \leq C_\gamma(1 + \|w_\tau\|_2)$. Consequently, we conclude for the case $H_t \geq \frac{A_0}{\gamma_2}$ that

$$\gamma_2 - \hat{\gamma}_2(w_t) \leq C_\gamma \frac{1 + \|w_\tau\|_2}{1 + H_t}. \tag{171}$$

For the case where $H_t < \frac{A_0}{\gamma_2}$, we only use $0 \le \gamma_2 - \hat{\gamma}_2(w_t) \le \gamma_2$. Since $1 + H_t < 1 + \frac{A_0}{\gamma_2}$, we obtain

$$\gamma_2 - \hat{\gamma}_2(w_t) \le \gamma_2 \le \gamma_2(1 + \frac{A_0}{\gamma_2})\frac{1}{1+H_t} \le (\gamma_2 + A_0)\frac{1 + \|w_\tau\|_2}{1+H_t}. \tag{172}$$

Combining (171) and (172) together, we conclude that $\gamma_2 - \hat{\gamma}_2(w_t) \le C_\gamma \frac{1+\|w_\tau\|_2}{1+H_t}$.

Next, we prove part (ii). Denote $\triangle := \triangle_{\text{geom}} > 0$, $\tau_\zeta := \tau_\zeta^\delta(\epsilon) = \inf\{t \ge \tau : \gamma_2 - \hat{\gamma}_2(w_t) \le \zeta\}$, $R_\tau := \|w_\tau\|_2$, and $V_t := V_t^\delta(\epsilon) = \sum_{s=\tau}^{t-1} \eta_s$. Fix any $0 < \zeta < \triangle$. Proposition 3.1 gives $H_t \asymp_\delta \log(1 + V_t)$. Given $V_t = \sum_{s=\tau}^{t-1} \eta_s \to \infty$ as $t \to \infty$, it holds that $H_t \to \infty$. Define $A_\zeta := \frac{C_\gamma}{\zeta}(1 + R_\tau)$ and $\bar{\tau}_\zeta := \inf\{t \ge \tau : H_t \ge A_\zeta\}$. Since $H_t \to \infty$, this time is finite. At $t = \bar{\tau}_\zeta$, from part (i), we obtain

$$\gamma_2 - \hat{\gamma}_2(w_{\bar{\tau}_\zeta}) \le C_\gamma \frac{1 + R_\tau}{1 + H_{\bar{\tau}_\zeta}} \le C_\gamma \frac{1 + R_\tau}{H_{\bar{\tau}_\zeta}} \le \zeta.$$

Therefore, $\bar{\tau}_\zeta$ belongs to the set defining τ_ζ, and consequently we have $\tau_\zeta \le \bar{\tau}_\zeta$. In particular, this shows that $\tau_\zeta < \infty$. Since H_t is increasing, it holds that $H_{\tau_\zeta} \le H_{\bar{\tau}_\zeta}$. By minimality of $\bar{\tau}_\zeta$, it also holds that $H_{\bar{\tau}_\zeta - 1} < A_\zeta$. Using $H_{\bar{\tau}_\zeta} = H_{\bar{\tau}_\zeta - 1} + \theta_{\bar{\tau}_\zeta - 1}$, we obtain $H_{\bar{\tau}_\zeta} < A_\zeta + \theta_{\bar{\tau}_\zeta - 1}$. Because $H_\tau = 0 < A_\zeta$, we must have $\bar{\tau}_\zeta \ge \tau + 1$ or equivalently $\bar{\tau}_\zeta - 1 \ge \tau$. Given $\sup_{s \ge \tau} \theta_s \le K_\theta$ (proved in (i)), we have $\theta_{\bar{\tau}_\zeta - 1} \le K_\theta$. These inequalities lead to

$$H_{\tau_\zeta} \le H_{\bar{\tau}_\zeta} < \frac{C_\gamma}{\zeta}(1 + R_\tau) + K_\theta \le (\frac{C_\gamma}{\zeta} + K_\theta)(1 + R_\tau).$$

Next, we show the lower bound on H_{τ_ζ}. From Theorem 29, we have $\gamma_2 - \hat{\gamma}_2(w_\tau) \ge \triangle - C_{2,\delta} r_\epsilon$. Choose ϵ sufficiently small such that $C_{2,\delta} r_\epsilon \le \frac{\triangle - \zeta}{2}$. Then, it holds that

$$\gamma_2 - \hat{\gamma}_2(w_\tau) \ge \triangle - \frac{\triangle - \zeta}{2} = \frac{\triangle + \zeta}{2}.$$

In particular, we have $\gamma_2 - \hat{\gamma}_2(w_\tau) > \zeta$. This implies that $\tau_\zeta > \tau$. At $t = \tau_\zeta$, we have $\gamma_2 - \hat{\gamma}_2(w_{\tau_\zeta}) \le \zeta$. Consequently, it holds that

$$\begin{aligned}\hat{\gamma}_2(w_{\tau_\zeta}) - \hat{\gamma}_2(w_\tau) &= [\gamma_2 - \hat{\gamma}_2(w_\tau)] - [\gamma_2 - \hat{\gamma}_2(w_{\tau_\zeta})] \\ &\ge \frac{\triangle + \zeta}{2} - \zeta = \frac{\triangle - \zeta}{2}.\end{aligned}$$

Applying Lemma 28 with $x = w_{\tau_\zeta}$ and $y = w_\tau$, we obtain

$$\frac{\triangle - \zeta}{2} \le \frac{2R_2 \|w_{\tau_\zeta} - w_\tau\|_2}{R_\tau}. \tag{173}$$

Based on the post-crossover update $w_{s+1} = w_s - \theta_s(u_s + e_s)$ with $\|u_s\|_2 = 1$, we unroll from τ to $\tau_\zeta - 1$ to obtain $w_{\tau_\zeta} - w_\tau = -\sum_{s=\tau}^{\tau_\zeta - 1} \theta_s u_s - \sum_{s=\tau}^{\tau_\zeta - 1} \theta_s e_s$. Therefore, it holds that

$$\|w_{\tau_\zeta} - w_\tau\|_2 \le \sum_{s=\tau}^{\tau_\zeta - 1} \theta_s \|u_s\|_2 + \sum_{s=\tau}^{\tau_\zeta - 1} \theta_s \|e_s\|_2 \le H_{\tau_\zeta} + E_0. \tag{174}$$

Combining (173) and (174), we obtain $\frac{\triangle - \zeta}{2} \le \frac{2R_2(H_{\tau_\zeta} + E_0)}{R_\tau}$. This is equivalent to

$$H_{\tau_\zeta} \ge \frac{\triangle - \zeta}{4R_2} R_\tau - E_0. \tag{175}$$

It remains to absorb the constant E_0.

From the proof of Proposition 3.2, at time τ, we have $\rho(w_\tau) \ge \ell_\epsilon - A_\delta$ where $A_\delta := [\log \frac{2n\delta}{\log 2}]_+$. Because $\rho(w) \le \gamma_2 \|w\|_2$, we have $R_\tau \ge \frac{\ell_\epsilon - A_\delta}{\gamma_2}$. Therefore, it holds that $R_\tau \to \infty$ as $\epsilon \downarrow 0$. In particular, we choose ϵ sufficiently small such that $R_\tau \ge \frac{8R_2 E_0}{\triangle - \zeta}$. This implies that $E_0 \le \frac{\triangle - \zeta}{8R_2} R_\tau$. Substituting this into (175), we obtain

$$H_{\tau_\zeta} \ge \frac{\triangle - \zeta}{4R_2} R_\tau - \frac{\triangle - \zeta}{8R_2} R_\tau = \frac{\triangle - \zeta}{8R_2} R_\tau.$$

Therefore, we can take

$$c_\zeta = \frac{\triangle - \zeta}{8R_2}, \qquad C_\zeta = \max\{\frac{C_\gamma}{\zeta} + K_\theta, \frac{\triangle - \zeta}{8R_2}\}.$$

Then, it holds that $0 < c_\zeta \le C_\zeta < \infty$ and

$$c_\zeta \|w_{T_1^\delta(\epsilon)}\|_2 \le H^\delta_{\tau_\zeta^\delta(\epsilon)}(\epsilon) \le C_\zeta(1 + \|w_{T_1^\delta(\epsilon)}\|_2). \tag{176}$$

In the proof of Theorem 29, we have shown $\|w_{\tau_\epsilon}\|_\infty = \frac{\ell_\epsilon}{\gamma_\infty} + O_\delta(\omega_a(\ell_\epsilon))$. Given this, we can write $\|w_{\tau_\epsilon}\|_\infty = \frac{\ell_\epsilon}{\gamma_\infty} + R_\epsilon$ with $|R_\epsilon| \le C_\delta \omega_a(\ell_\epsilon)$ when ϵ is sufficiently small. Given $r_\epsilon \to 0$, it holds that $C_\delta \gamma_\infty r_\epsilon \le \frac{1}{2}$ or equivalently $C_\delta \omega_a(\ell_\epsilon) \le \frac{\ell_\epsilon}{2\gamma_\infty}$ when ϵ is sufficiently small. Combining them together, for sufficiently small ϵ, it holds that $|R_\epsilon| \le \frac{\ell_\epsilon}{2\gamma_\infty}$. This leads to $\|w_{\tau_\epsilon}\|_\infty = \frac{\ell_\epsilon}{\gamma_\infty} + R_\epsilon \ge \frac{\ell_\epsilon}{\gamma_\infty} - |R_\epsilon| \ge \frac{\ell_\epsilon}{\gamma_\infty} - \frac{\ell_\epsilon}{2\gamma_\infty} = \frac{\ell_\epsilon}{2\gamma_\infty}$ and $\|w_{\tau_\epsilon}\|_\infty = \frac{\ell_\epsilon}{\gamma_\infty} + R_\epsilon \le \frac{\ell_\epsilon}{\gamma_\infty} + |R_\epsilon| \le \frac{\ell_\epsilon}{\gamma_\infty} + \frac{\ell_\epsilon}{2\gamma_\infty} \le \frac{2\ell_\epsilon}{\gamma_\infty}$. Therefore, we conclude for sufficiently small ϵ that

$$\frac{\ell_\epsilon}{2\gamma_\infty} \le \|w_{\tau_\epsilon}\|_\infty \le \frac{2\ell_\epsilon}{\gamma_\infty}.$$

Using $\|w\|_\infty \le \|w\|_2 \le \sqrt{d}\|w\|_\infty$, we further obtain $\frac{\ell_\epsilon}{2\gamma_\infty} \le \|w_{\tau_\epsilon}\|_2 \le \frac{2\sqrt{d}}{\gamma_\infty}\ell_\epsilon$. Substituting this into (176), we conclude that

$$H^\delta_{\tau_\zeta^\delta(\epsilon)}(\epsilon) = \Theta(\ell_\epsilon) = \Theta(\log\frac{1}{\epsilon}). \tag{177}$$

Based on $H_t \asymp_\delta \log(1 + V_t)$ (shown in Proposition 3.1), we know there exist constants $0 < \bar{c}_\delta \le \bar{C}_\delta < \infty$ such that

$$\bar{c}_\delta \log(1 + V_t) \le H_t \le \bar{C}_\delta \log(1 + V_t), \qquad t \ge \tau.$$

At $t = \tau_\zeta$, (177) implies that there exist constants $0 < h_\zeta^- \le h_\zeta^+ < \infty$ such that $h_\zeta^-\ell_\epsilon \le H_{\tau_\zeta} \le h_\zeta^+\ell_\epsilon$. Combining them, we obtain

$$\frac{h_\zeta^-}{\bar{C}_\delta}\ell_\epsilon \le \log(1 + V_{\tau_\zeta}) \le \frac{h_\zeta^+}{\bar{c}_\delta}\ell_\epsilon. \tag{178}$$

Define $\alpha_\zeta := \frac{h_\zeta^-}{\bar{C}_\delta} > 0$ and $\beta_\zeta := \frac{h_\zeta^+}{\bar{c}_\delta} < \infty$. Exponentiating (178) and using $e^{\ell_\epsilon} = \frac{1}{\epsilon}$, we obtain $\epsilon^{-\alpha_\zeta} \le 1 + V_{\tau_\zeta} \le \epsilon^{-\beta_\zeta}$. For sufficiently small ϵ, it holds that $\epsilon^{-\alpha_\zeta} \ge 2$. Consequently, we have $\frac{1}{2}\epsilon^{-\alpha_\zeta} \le V_{\tau_\zeta} \le \epsilon^{-\beta_\zeta}$. Therefore, we conclude that $V^\delta_{\tau_\zeta^\delta(\epsilon)}(\epsilon) = \epsilon^{-\Theta(1)}$.

Finally, we specify the threshold $\tilde{\epsilon}_4(\delta, \zeta)$ that incorporates the conditions on ϵ from both parts (i) and (ii). We define

$$L_2 := \begin{cases} \max\left\{1, \left(\frac{2C_{2,\delta}}{\Delta - \zeta}\right)^{1/a}\right\}, & \frac{1}{3} < a < 1, \\ \max\left\{e, \left(\frac{2C_{2,\delta}}{\Delta - \zeta}\right)^{2}\right\}, & a = 1, \end{cases}$$

and $L_{\delta,\zeta} := \max\{2A_\delta, \frac{16\gamma_2 R_2 E_0}{\triangle - \zeta}, L_2\}$. We set $\tilde{\epsilon}_4(\delta, \zeta) := \min\{\tilde{\epsilon}_1(\delta), \tilde{\epsilon}_4^{(i)}(\delta), \tilde{\epsilon}_2(\delta), e^{-L_{\delta,\zeta}}\}$. Then, any $0 < \epsilon \le \tilde{\epsilon}_4(\delta, \zeta)$ satisfies the required conditions. Indeed, $L_{\delta,\zeta} \ge L_2$ guarantees that $C_{2,\delta} r_\epsilon \le \frac{\triangle - \zeta}{2}$ (for $a = 1$, we used the fact $\frac{\log \ell_\epsilon}{\ell_\epsilon} \le \frac{1}{\sqrt{\ell_\epsilon}}$); $L_{\delta,\zeta} \ge 2A_\delta$ guarantees $R_\tau \ge \frac{\ell_\epsilon}{2\gamma_2}$; and $L_{\delta,\zeta} \ge \frac{16\gamma_2 R_2 E_0}{\triangle - \zeta}$ guarantees $R_\tau \ge \frac{8R_2E_0}{\triangle - \zeta}$. Moreover, we have ensured that $\tilde{\epsilon}_4(\delta, \zeta) \le \tilde{\epsilon}_1(\delta)$ so we can apply the results of Proposition 3.1. □

H Proof of Corollary 6

Proof of Corollary 6. Denote $\tau_\epsilon := T_1^\delta(\epsilon)$ and $\tau_\zeta := \tau_\zeta^\delta(\epsilon)$. From Theorem 4, we know $S_{\tau_\epsilon} - S_{T_*} = \frac{\ell_\epsilon}{\gamma_\infty} + o(\ell_\epsilon)$. Since S_{T_*} is a fixed constant independent of ϵ, we can write $S_{\tau_\epsilon} = \frac{1}{\gamma_\infty}\ell_\epsilon + o(\ell_\epsilon)$. From

Theorem 5, we proved there exist constants $0 < \alpha_{\delta,\zeta} \le \beta_{\delta,\zeta} < \infty$ such that $\frac{1}{2}\epsilon^{-\alpha_{\delta,\zeta}} \le V^{\delta}_{\tau_\zeta}(\epsilon) \le \epsilon^{-\beta_{\delta,\zeta}}$ when ϵ is sufficiently small. For simplicity, we let $\alpha := \alpha_{\delta,\zeta}$, $\beta := \beta_{\delta,\zeta}$, and $V_\epsilon := V^{\delta}_{\tau_\zeta}(\epsilon)$.

Since $\eta_s = \eta_0(s+1)^{-a}$, we have $S_t = \eta_0 \sum_{k=1}^{t} k^{-a}$. For the $0 < a < 1$ case, since the function x^{-a} is decreasing, it holds that $\int_1^{t+1} x^{-a}dx \le \sum_{k=1}^{t} k^{-a} \le 1 + \int_1^t x^{-a}dx$. Evaluating the integrals leads to

$$\frac{(t+1)^{1-a}-1}{1-a} \le \sum_{k=1}^{t} k^{-a} \le 1 + \frac{t^{1-a}-1}{1-a}. \tag{179}$$

Define $M_t := \frac{t^{1-a}}{1-a}$. Then, the upper bound satisfies $1 + \frac{t^{1-a}}{1-a} = M_t - \frac{a}{1-a}$. For the lower bound, we have

$$\frac{(t+1)^{1-a}-1}{1-a} - M_t = \frac{(t+1)^{1-a} - t^{1-a} - 1}{1-a}. \tag{180}$$

Applying the mean value theorem to $h(x) = x^{1-a}$, there exists some $\nu_t \in (t, t+1)$ such that $(t+1)^{1-a} - t^{1-a} = h'(\nu_t) = (1-a)\nu_t^{-a}$. For $t \ge 1$, we have $\nu_t \ge 1$ and hence $0 < \nu_t^{-a} \le 1$. Consequently, it holds that $0 < (t+1)^{1-a} - t^{1-a} \le 1 - a$. Substituting this into (180), we obtain

$$-\frac{1}{1-a} < \frac{(t+1)^{1-a}-1}{1-a} - M_t \le -\frac{a}{1-a}.$$

Combining this with (179), we obtain

$$\frac{t^{1-a}}{1-a} - \frac{1}{1-a} \le \sum_{k=1}^{t} k^{-a} \le \frac{t^{1-a}}{1-a} - \frac{a}{1-a}.$$

If we define $E_t := \sum_{k=1}^{t} k^{-a} - \frac{t^{1-a}}{1-a}$, then it holds that $-\frac{1}{1-a} \le E_t \le -\frac{a}{1-a}$. This implies that $|E_t| \le \frac{1}{1-a}$ and hence $E_t = O(1)$. From this, we obtain $\sum_{k=1}^{t} k^{-a} = \frac{t^{1-a}}{1-a} + O(1)$ and consequently

$$S_t = \frac{\eta_0}{1-a} t^{1-a} + O(1), \qquad 0 < a < 1 \tag{181}$$

In the case where $a = 1$, we have $\int_1^{t+1} \frac{dx}{x} \le \sum_{k=1}^{t} \frac{1}{k} \le 1 + \int_1^t \frac{dx}{x}$. Evaluating the integral leads to $\log(t+1) \le \sum_{k=1}^{t} \frac{1}{k} \le 1 + \log t$. If we let $\Omega_t := \sum_{k=1}^{t} \frac{1}{k}$, then we have $\log t \le \Omega_t \le \log t + 1$ and consequently $\Omega_t = \log t + O(1)$. From this, we conclude that

$$S_t = \eta_0 \log t + O(1), \qquad a = 1. \tag{182}$$

Now, we prove part (i) where the learning rate schedule is $\eta_t = \eta_0(t+1)^{-a}$, $a \in (1/3, 1)$. Applying (181) at $t = \tau_\epsilon$ and using $S_{\tau_\epsilon} = \frac{1}{\gamma_\infty}\ell_\epsilon + o(\ell_\epsilon)$, we obtain

$$\frac{\eta_0}{1-a}\tau_\epsilon^{1-a} + O(1) = \frac{1}{\gamma_\infty}\ell_\epsilon + o(\ell_\epsilon).$$

Since $\ell_\epsilon \to \infty$, the $O(1)$ term is $o(\ell_\epsilon)$ and it holds that $\tau_\epsilon^{1-a} = \frac{1-a}{\eta_0\gamma_\infty}\ell_\epsilon + o(\ell_\epsilon)$. Equivalently, we can write $\tau_\epsilon^{1-a} = \frac{1-a}{\eta_0\gamma_\infty}\ell_\epsilon(1 + o(1))$. Taking the $1/(1-a)$-th power gives

$$T_1^{\delta}(\epsilon) = \big(\frac{1-a}{\eta_0\gamma_\infty}\log\frac{1}{\epsilon}\big)^{1/(1-a)}\big(1 + o(1)\big) = \Theta\big((\log\frac{1}{\epsilon})^{1/(1-a)}\big).$$

By definition, it holds that $S_{\tau_\zeta} = S_{\tau_\epsilon} + V_\epsilon$. From $S_{\tau_\epsilon} = \frac{1}{\gamma_\infty}\ell_\epsilon + o(\ell_\epsilon)$, we know $S_{\tau_\epsilon} = O(\ell_\epsilon)$. On the other hand, we have $V_\epsilon \ge \frac{1}{2}\epsilon^{-\alpha}$. Consequently, it holds that $\frac{S_{\tau_\epsilon}}{V_\epsilon} \le O(\ell_\epsilon\epsilon^{\alpha}) \to 0$ given $\epsilon^{\alpha}\log\frac{1}{\epsilon} \to 0$ for every $\alpha > 0$. This implies that $S_{\tau_\zeta} = V_\epsilon(1 + o(1))$. Applying (181) at $t = \tau_\zeta$, we obtain $\frac{\eta_0}{1-a}\tau_\zeta^{1-a} + O(1) = V_\epsilon(1 + o(1))$. Because $V_\epsilon \to \infty$, the O(1) term is $o(V_\epsilon)$. Therefore, after rearranging we obtain

$$\tau_\zeta^{1-a} = \frac{1-a}{\eta_0} V_\epsilon\big(1 + o(1)\big).$$

Define $K_a := \frac{1-a}{\eta_0} > 0$. We can write $\tau_\zeta^{1-a} = K_a V_\epsilon(1+r_\epsilon)$ where $r_\epsilon \to 0$ as $\epsilon \downarrow 0$. Consequently, we have $|r_\epsilon| \le \frac{1}{2}$ when ϵ is sufficiently small. Consequently, it holds that $\frac{1}{2} \le 1 + r_\epsilon \le \frac{3}{2}$. Substituting this into $\tau_\zeta^{1-a} = K_a V_\epsilon(1+r_\epsilon)$, we obtain $\frac{K_a}{2} V_\epsilon \le \tau_\zeta^{1-a} \le \frac{3K_a}{2} V_\epsilon$. Combining this with $\frac{1}{2}\epsilon^{-\alpha} \le V_\epsilon \le \epsilon^{-\beta}$, we obtain for sufficiently small ϵ that

$$\frac{K_a}{2}\frac{1}{2}\epsilon^{-\alpha} \le \frac{K_a}{2} V_\epsilon \le \tau_\zeta^{1-a} \le \frac{3K_a}{2} V_\epsilon \le \frac{3K_a}{2}\epsilon^{-\beta}.$$

Consequently, it holds that $\frac{K_a}{4}\epsilon^{-\alpha} \le \tau_\zeta^{1-a} \le \frac{3K_a}{2}\epsilon^{-\beta}$ when ϵ is sufficiently small. Given $a \in (\frac{1}{3}, 1)$, the function $x^{1/(1-a)}$ is strictly increasing. Applying this function, we obtain

$$(\frac{K_a}{4})^{1/(1-a)}\epsilon^{-\alpha/(1-a)} \le \tau_\zeta \le (\frac{3K_a}{2})^{1/(1-a)}\epsilon^{-\beta/(1-a)}.$$

Therefore, we conclude that $\tau_\zeta^\delta(\epsilon) = \epsilon^{-\Theta(1)/(1-a)}$.

Now, we prove part (ii) where $\eta_t = \eta_0(t+1)^{-1}$. Applying (182) at $t = \tau_\epsilon$ and using $S_{\tau_\epsilon} = \frac{1}{\gamma_\infty}\ell_\epsilon + o(\ell_\epsilon)$, we obtain

$$\eta_0 \log \tau_\epsilon + O(1) = \frac{\ell_\epsilon}{\gamma_\infty} + o(\ell_\epsilon).$$

Given $O(1) = o(\ell_\epsilon)$, it holds that $\log \tau_\epsilon = \frac{1}{\eta_0\gamma_\infty}\ell_\epsilon + o(\ell_\epsilon) = (\frac{1}{\eta_0\gamma_\infty} + o(1))\log\frac{1}{\epsilon}$. Exponentiating leads to $T_1^\delta(\epsilon) = \epsilon^{-1/(\eta_0\gamma_\infty)+o(1)}$. Given $S_{\tau_\zeta} = V_\epsilon(1+o(1))$ (shown above), using (182) at $t = \tau_\zeta$, we obtain

$$\eta_0 \log \tau_\zeta + O(1) = V_\epsilon(1+o(1)).$$

We can rearrange this equation and obtain $\eta_0 \log \tau_\zeta = V_\epsilon(1 + o(1) - \frac{O(1)}{V_\epsilon})$. Since $V_\epsilon \to \infty$, we have $\frac{O(1)}{V_\epsilon} \to 0$ and consequently it holds that $\log \tau_\zeta = \frac{1}{\eta_0} V_\epsilon(1+o(1))$. We can write $\log \tau_\zeta = \frac{V_\epsilon}{\eta_0}(1+r_\epsilon)$ where $r_\epsilon \to 0$. For sufficiently small ϵ, it holds that $-\frac{1}{2} \le r_\epsilon \le \frac{1}{2}$. This leads to $\frac{1}{2} \le 1 + r_\epsilon \le \frac{3}{2}$ and consequently $\frac{1}{2\eta_0} V_\epsilon \le \log \tau_\zeta \le \frac{3}{2\eta_0} V_\epsilon$. Using the bound $\frac{1}{2}\epsilon^{-\alpha} \le V_\epsilon \le \epsilon^{-\beta}$, we obtain

$$\frac{1}{4\eta_0}\epsilon^{-\alpha} = \frac{1}{2\eta_0}(\frac{1}{2}\epsilon^{-\alpha}) \le \frac{1}{2\eta_0} V_\epsilon \le \log \tau_\zeta \le \frac{3}{2\eta_0} V_\epsilon \le \frac{3}{2\eta_0}\epsilon^{-\beta}.$$

Exponentiating leads to $\exp(\frac{1}{4\eta_0}\epsilon^{-\alpha}) \le \tau_\zeta \le \exp(\frac{3}{2\eta_0}\epsilon^{-\beta})$. Therefore, we conclude that $\tau_\zeta^\delta(\epsilon) = \exp(\epsilon^{-\Theta(1)})$. □

I Multiclass Extension

I.1 Setup and notation

We let the training set be $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in [K]$ (with $K \ge 2$). The classifier is $W \in \mathbb{R}^{K\times d}$ and we denote

$$p_i(W) := \mathrm{softmax}(Wx_i), \qquad p_{ic}(W) := [p_i(W)]_c.$$

For every $i \in [n]$ and $c \ne y_i$, we define the pairwise feature matrix and the pairwise margin, respectively, as

$$Z_{ic} := (e_{y_i} - e_c)x_i^\top, \qquad m_{ic}(W) := \langle W, Z_{ic}\rangle = (e_{y_i} - e_c)^\top W x_i.$$

The cross-entropy loss (on the training set) is defined as

$$L(W) := -\frac{1}{n}\sum_{i=1}^n \log p_{i,y_i}(W) = \frac{1}{n}\sum_{i=1}^n \log\left(1 + \sum_{c\ne y_i} e^{-m_{ic}(W)}\right). \tag{183}$$

Following the construction in Fan et al. (2025), the multiclass proxy (analogous to the binary proxy) is defined as

$$G(W) := \frac{1}{n}\sum_{i=1}^{n}\big(1 - p_{i,y_i}(W)\big) = \frac{1}{n}\sum_{i=1}^{n}\sum_{c\neq y_i} p_{ic}(W).$$

For simplicity, we let $L_t := L(W_t)$, $G_t = G(W_t)$, and $g_t := \nabla L(W_t)$. The multiclass unnormalized margin is

$$\rho(W) := \min_{i\in[n],\, c\neq y_i} m_{ic}(W).$$

Matrix norms relevant in the multiclass setting are

$$\text{max-norm: } \|A\|_{\max} := \max_{c\in[K],\, j\in[d]} |A[c,j]|, \qquad \text{sum-norm: } \|A\|_{\text{sum}} := \sum_{c=1}^{K}\sum_{j=1}^{d}|A[c,j]|,$$

and

$$\text{Frobenius-norm: } \|A\|_F := \left(\sum_{c=1}^{K}\sum_{j=1}^{d} A[c,j]^2\right)^{1/2} = \sqrt{\langle A, A\rangle}.$$

The max-norm and the sum-norm are dual to each other. We define the max-norm margin and the Frobenius-norm margin as

$$\gamma_{\max} := \max_{\|U\|_{\max}\le 1} \rho(U), \qquad \gamma_F := \max_{\|U\|_F\le 1} \rho(U). \tag{184}$$

Similar to the binary case, we assume the multiclass data is separable: there exists $W \in \mathbb{R}^{k\times d}$ such that $\min_{c\neq y_i}(e_{y_i} - e_c)^T W x_i > 0$ for all $i \in [n]$, and the momentum parameters satisfy $0 \le \beta_1 \le \beta_2 < 1$. We study the same step size schedule $\eta_t = \eta_0(t+1)^{-a}$ with $a \in (1/3, 1]$. Denote $\max_i \|x_i\|_1 = B_1$ and $\max_i \|x_i\|_2 = B_2$.

We set $R_1 := 2B_1$ and $R_2 := \sqrt{2}\,B_2$. Importantly, the constants R_1 and R_2 play the same roles as the ones for binary data. Using the outer-product identities $\|uv^\top\|_{\text{sum}} = \|u\|_1\|v\|_1$ and $\|uv^\top\|_F = \|u\|_2\|v\|_2$ together with $\|e_{y_i} - e_c\|_1 = 2$ and $\|e_{y_i} - e_c\|_2 = \sqrt{2}$, we obtain $\|Z_{ic}\|_{\text{sum}} = \|e_{y_i} - e_c\|_1\|x_i\|_1 = 2\|x_i\|_1$ and $\|Z_{ic}\|_F = \|e_{y_i} - e_c\|_2\|x_i\|_2 = \sqrt{2}\|x_i\|_2$. Consequently, it holds that

$$\max_{i,\, c\neq y_i}\|Z_{ic}\|_{\text{sum}} = 2\max_{i\in[n]}\|x_i\|_1 \le 2B_1 = R_1, \quad \max_{i,\, c\neq y_i}\|Z_{ic}\|_F = \sqrt{2}\max_{i\in[n]}\|x_i\|_2 \le \sqrt{2}B_2 = R_2.$$

All products, quotients, powers, square roots, and absolute values in defining Adam's updates below are entrywise (as in the binary case). Therefore, the parameter contains $d_{\text{mul}} := Kd$ scalar coordinates. The proof in the multiclass setting is very similar to that in the binary setting. Once the analogues of the lemmas in App. B and App. F are established, the remaining arguments follow the same steps, with mostly notational changes. We prove the required analogous lemmas in the next section and then state the multiclass results.

I.2 Section B Analogous Lemmas

The following lemma is analogous to Lemma 7. Its proof can be found in Fan et al. (2025, Lemma 16). We provide the proof here for completeness.

Lemma 33. *For every $W \in \mathbb{R}^{K\times d}$, it holds that*

$$\nabla L(W) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c\neq y_i} p_{ic}(W) Z_{ic}. \tag{185}$$

Moreover, we have

$$\gamma_{\max} G(W) \le \|\nabla L(W)\|_{\text{sum}} \le R_1 G(W), \qquad \gamma_F G(W) \le \|\nabla L(W)\|_F \le R_2 G(W). \tag{186}$$

In particular, it holds that $G(W) > 0$ and $\nabla L(W) \neq 0$ at every finite W.

Proof. Differentiating the ith cross-entropy term gives

$$\nabla_W[-\log p_{i,y_i}(W)] = (p_i(W) - e_{y_i})x_i^\top = -\sum_{c\neq y_i} p_{ic}(W)Z_{ic},$$

which proves (185). Let $U_{\max}$ attain the first maximum in (184). Then, it holds that

$$\begin{aligned} -\langle \nabla L(W), U_{\max}\rangle &= \frac{1}{n}\sum_i \sum_{c\neq y_i} p_{ic}(W)\,\langle Z_{ic}, U_{\max}\rangle \\ &\geq \frac{\gamma_{\max}}{n}\sum_i\sum_{c\neq y_i} p_{ic}(W) = \gamma_{\max}G(W). \end{aligned}$$

By duality and $\|U_{\max}\|_{\max} \leq 1$, we have $-\langle \nabla L(W), U_{\max}\rangle \leq \|\nabla L(W)\|_{\text{sum}}$. This proves the lower bound on the sum-norm in (186). The upper bound on the sum-norm follows from (185) and the triangle inequality:

$$\|\nabla L(W)\|_{\text{sum}} \leq \frac{1}{n}\sum_i\sum_{c\neq y_i} p_{ic}(W)\,\|Z_{ic}\|_{\text{sum}} \leq R_1 G(W).$$

The Frobenius bounds follow from the same argument with a Frobenius max-margin solution U_F (note that the dual of Frobenius norm is itself). Finally, every wrong class has strictly positive softmax probability at finite W, which leads to $G(W) > 0$; the lower bounds then imply $\nabla L(W) \neq 0$. □

The following lemma is analogous to Lemma 8. The proofs of (187) and (188) can be found in Fan et al. (2025, Lemma 18). We provide the proof here for completeness.

Lemma 34. *For every $W \in \mathbb{R}^{K\times d}$, it holds that*

$$G(W) \leq L(W), \qquad \frac{G(W)}{L(W)} \geq 1 - \frac{nL(W)}{2}. \tag{187}$$

If either $G(W) \leq 1/(2n)$ or $L(W) \leq (\log 2)/n$ holds, then we have

$$L(W) \leq 2G(W). \tag{188}$$

Furthermore, $L(W) \leq (\log 2)/n$ implies that

$$\rho(W) \geq 0, \quad \textit{and} \quad \frac{\log 2}{n}e^{-\rho(W)} \leq L(W). \tag{189}$$

Finally, it holds that $L(W) \leq (K-1)e^{-\rho(W)}$ for every $W \in \mathbb{R}^{K\times d}$.

Proof. Let $\ell_i(W) := -\log p_{iy_i}(W)$ and $q_i(W) := 1 - p_{iy_i}(W)$. The key multiclass identity is

$$q_i(W) = 1 - e^{-\ell_i(W)}. \tag{190}$$

For $x \geq 0$, it holds that $x - \frac{x^2}{2} \leq 1 - e^{-x} \leq x$. Averaging this inequality and using $\sum_i \ell_i^2 \leq (\sum_i \ell_i)^2 = n^2 L(W)^2$ gives

$$L(W) - \frac{n}{2}L(W)^2 \leq G(W) \leq L(W),$$

which proves (187).

If $G(W) \leq \frac{1}{2n}$, then we have $q_i(W) \leq nG(W) \leq \frac{1}{2}$ for every i. If instead $L(W) \leq \frac{\log 2}{n}$, then we have $\ell_i(W) \leq nL(W) \leq \log 2$, and (190) again gives $q_i(W) \leq \frac{1}{2}$. For $x \in [0, \frac{1}{2}]$, it holds that $-\log(1-x) \leq 2x$. Consequently, we have for either case that

$$L(W) = \frac{1}{n}\sum_i -\log(1-q_i(W)) \leq \frac{2}{n}\sum_i q_i(W) = 2G(W).$$

Now suppose $L(W) \leq (\log 2)/n$. For every i, we have

$$\log\left(1 + \sum_{c\neq y_i} e^{-m_{ic}(W)}\right) = \ell_i(W) \leq \log 2.$$

Therefore, it holds that $\sum_{c\neq y_i} e^{-m_{ic}(W)} \leq 1$. This implies that every $m_{ic}(W) \geq 0$ and therefore $\rho(W) \geq 0$. Choose $(i_\star, c_\star)$ attaining $\rho(W)$. Then, it holds that

$$L(W) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \sum_{c\neq y_i} e^{-m_{ic}(W)}\right) \geq \frac{1}{n}\log(1 + e^{-m_{i_\star c_\star}}) = \frac{1}{n}\log\bigl(1 + e^{-\rho(W)}\bigr).$$

For $x \in [0,1]$, the concavity of $\log(1+x)$ gives $\log(1+x) \geq x\log 2$. Substituting $x = e^{-\rho(W)}$ proves the lower bound in (189). Finally, $\log(1+x) \leq x$ gives

$$L(W) \leq \frac{1}{n}\sum_{i}\sum_{c\neq y_i} e^{-m_{ic}(W)} \leq (K-1)e^{-\rho(W)},$$

which proves the upper bound. □

The following lemma is analogous to Lemma 12 in Section B. It will be used in the proof of Lemma 37 below.

Lemma 35. *For any $U, V \in \mathbb{R}^{K\times d}$, every $i \in [n]$, and every $c \in [K]$, we have*

$$e^{-R_1\|U-V\|_{\max}} \leq \frac{p_{ic}(U)}{p_{ic}(V)} \leq e^{R_1\|U-V\|_{\max}}, \tag{191}$$

and

$$e^{-R_2\|U-V\|_F} \leq \frac{p_{ic}(U)}{p_{ic}(V)} \leq e^{R_2\|U-V\|_F}. \tag{192}$$

Consequently, denote $R_{\max} := R_1$, $R_F := R_2$, the following holds for $p \in \{\max, F\}$

$$e^{-R_p\|U-V\|_p} G(V) \leq G(U) \leq e^{R_p\|U-V\|_p} G(V). \tag{193}$$

Further assume that $0 \leq \beta_1 \leq \beta_2 < 1$. It holds for every $0 \leq r \leq t$ that

$$e^{-R_1 C_D \sum_{s=r}^{t-1}\eta_s} G_r \leq G_t \leq e^{R_1 C_D \sum_{s=r}^{t-1}\eta_s} G_r. \tag{194}$$

Proof. Denote $\Delta := U - V$. Fix a sample $i \in [n]$ and a class $c \in [K]$. For each $r \in [K]$, we define $z_r := (Vx_i)_r$ and $a_r := (\Delta x_i)_r$. Then, we have $(Ux_i)_r = z_r + a_r$. By the definition of softmax, we have

$$p_{ic}(U) = \frac{e^{z_c+a_c}}{\sum_{r=1}^K e^{z_r+a_r}}, \qquad p_{ic}(V) = \frac{e^{z_c}}{\sum_{r=1}^K e^{z_r}}.$$

Therefore, it holds that

$$\begin{aligned}\frac{p_{ic}(U)}{p_{ic}(V)} &= \frac{e^{z_c+a_c}}{\sum_{r=1}^K e^{z_r+a_r}} \frac{\sum_{r=1}^K e^{z_r}}{e^{z_c}} = e^{a_c}\frac{\sum_{r=1}^K e^{z_r}}{\sum_{r=1}^K e^{z_r}e^{a_r}} \\ &= \frac{e^{a_c}}{\sum_{r=1}^K \frac{e^{z_r}}{\sum_{s=1}^K e^{z_s}} e^{a_r}} = \frac{e^{a_c}}{\sum_{r=1}^K p_{ir}(V)e^{a_r}}.\end{aligned} \tag{195}$$

We further define $a_{\min} := \min_{r\in[K]} a_r$ and $a_{\max} := \max_{r\in[K]} a_r$. Because $p_{ir}(V) \geq 0$ and $\sum_{r=1}^K p_{ir}(V) = 1$, we have $e^{a_{\min}} \leq \sum_{r=1}^K p_{ir}(V)e^{a_r} \leq e^{a_{\max}}$. Combining this with (195) gives

$$e^{a_c - a_{\max}} \leq \frac{p_{ic}(U)}{p_{ic}(V)} \leq e^{a_c - a_{\min}}.$$

Since $a_{\min} \leq a_c \leq a_{\max}$, we have $a_c - a_{\max} \geq -(a_{\max} - a_{\min})$ and $a_c - a_{\min} \leq a_{\max} - a_{\min}$. Consequently, it holds that

$$e^{-(a_{\max}-a_{\min})} \leq \frac{p_{ic}(U)}{p_{ic}(V)} \leq e^{a_{\max}-a_{\min}}. \tag{196}$$

It remains to control the difference $a_{\max} - a_{\min}$. For any $r, s \in [K]$, it holds that

$$a_r - a_s = (e_r - e_s)^\top \Delta x_i = \left\langle \Delta, (e_r - e_s)x_i^\top \right\rangle_F. \tag{197}$$

By duality, we have

$$|a_r - a_s| \le \|\Delta\|_{\max} \left\|(e_r - e_s)x_i^\top\right\|_{\text{sum}} \le 2\,\|\Delta\|_{\max}\,\|x_i\|_1 \le R_1\,\|\Delta\|_{\max}. \tag{198}$$

It follows that

$$a_{\max} - a_{\min} = \max_{r,s\in[K]} (a_r - a_s) \le R_1\,\|U - V\|_{\max}.$$

Substituting this bound into (196) yields

$$e^{-R_1\|U-V\|_{\max}} \le \frac{p_{ic}(U)}{p_{ic}(V)} \le e^{R_1\|U-V\|_{\max}}.$$

Similarly, by Frobenius norm self-duality, we have

$$|a_r - a_s| \le \|\Delta\|_F \left\|(e_r - e_s)x_i^\top\right\|_F \le \sqrt{2}\,\|\Delta\|_F\,\|x_i\|_2 \le R_2\,\|\Delta\|_F. \tag{199}$$

Therefore, it holds that $a_{\max} - a_{\min} \le R_2\,\|U - V\|_F$.

Using (196) again gives

$$e^{-R_2\|U-V\|_F} \le \frac{p_{ic}(U)}{p_{ic}(V)} \le e^{R_2\|U-V\|_F}.$$

Finally, we let $p \in \{\max, F\}$ where $R_{\max} := R_1$ and $R_F := R_2$. The preceding probability-ratio bounds imply that for every $i \in [n]$ and $c \ne y_i$,

$$e^{-R_p\|U-V\|_p} p_{ic}(V) \le p_{ic}(U) \le e^{R_p\|U-V\|_p} p_{ic}(V).$$

Summing over $i \in [n]$ and $c \ne y_i$, and then dividing by n, we obtain

$$G(U) = \frac{1}{n}\sum_{i=1}^{n}\sum_{c\ne y_i} p_{ic}(U) \le e^{R_p\|U-V\|_p}\frac{1}{n}\sum_{i=1}^{n}\sum_{c\ne y_i} p_{ic}(V) = e^{R_p\|U-V\|_p} G(V),$$

and similarly $G(U) \ge e^{-R_p\|U-V\|_p} G(V)$. Hence, it holds that

$$e^{-R_p\|U-V\|_p} G(V) \le G(U) \le e^{R_p\|U-V\|_p} G(V).$$

Applying this inequality with $p = \max$, $U = W_t$, and $V = W_r$, we obtain

$$e^{-R_1\|W_t - W_r\|_{\max}} G_r \le G_t \le e^{R_1\|W_t - W_r\|_{\max}} G_r. \tag{200}$$

By the triangle inequality and Lemma 38, it holds that

$$\|W_t - W_r\|_{\max} \le \sum_{s=r}^{t-1} \eta_s \|D_s\|_{\max} \le C_D \sum_{s=r}^{t-1} \eta_s.$$

Substituting this into (200) finishes the proof of (194). □

The probability simplex is defined as

$$\Delta^{K-1} := \left\{ s \in \mathbb{R}_+^K : \sum_{c=1}^{K} s_c = 1 \right\}.$$

The following lemma is a refinement of Fan et al. (2025, Lemma 14), specialized to the max-norm and Frobenius-norm cases. Direct application of Fan et al. (2025, Lemma 14) results a factor of $2R_2^2$ in the Frobenius bound (204). Lemma 36 will be used in the proof of Lemma 37.

Lemma 36. *For every $s \in \Delta^{K-1}$, $y \in [K]$, $A \in \mathbb{R}^{K\times d}$, and $h \in \mathbb{R}^d$, we have*

$$0 \le h^\top A^\top \big(\operatorname{diag}(s) - ss^\top\big) Ah \le 4(1 - s_y)\|h\|_1^2 \|A\|_{\max}^2, \tag{201}$$

and

$$0 \le h^\top A^\top \big(\operatorname{diag}(s) - ss^\top\big) Ah \le 2(1 - s_y)\|h\|_2^2 \|A\|_F^2. \tag{202}$$

Consequently, for every training example x_i, we have

$$x_i^\top A^\top \big(\operatorname{diag}(s) - ss^\top\big) Ax_i \le R_1^2(1 - s_y)\|A\|_{\max}^2, \tag{203}$$

and

$$x_i^\top A^\top \big(\operatorname{diag}(s) - ss^\top\big) Ax_i \le R_2^2(1 - s_y)\|A\|_F^2. \tag{204}$$

Proof. Let $s \in \Delta^{K-1}$. Denote $v := Ah \in \mathbb{R}^K$ and $C_s := \mathrm{diag}(s) - ss^\top$. Let C be a random class distributed according to s, i.e. $\mathbb{P}(C = c) = s_c$, $c \in [K]$. Then, we have

$$v^\top C_s v = \sum_{c=1}^{K} s_c v_c^2 - \left(\sum_{c=1}^{K} s_c v_c \right)^2 = \mathrm{Var}(v_C) \geq 0. \tag{205}$$

For any $a \in \mathbb{R}$, it holds that

$$\mathbb{E}\big[(v_C - a)^2\big] = \mathbb{E}\left[\big(v_C - \mathbb{E}[v_C] + \mathbb{E}[v_C] - a\big)^2\right] = \mathrm{Var}(v_C) + \big(\mathbb{E}[v_C] - a\big)^2 \geq \mathrm{Var}(v_C). \tag{206}$$

Choosing $a = v_y$ in (206), we obtain

$$v^\top C_s v = \mathrm{Var}(v_C) \leq \mathbb{E}\big[(v_C - v_y)^2\big] = \sum_{c=1}^{K} s_c (v_c - v_y)^2 = \sum_{c \neq y} s_c (v_c - v_y)^2. \tag{207}$$

For the max-norm bound, it holds for every $c \neq y$ that

$$\begin{aligned} |v_c - v_y| &= |(e_c - e_y)^\top Ah| = \big|\big\langle A, (e_c - e_y)h^\top \big\rangle\big| \\ &\leq \|A\|_{\max} \big\|(e_c - e_y)h^\top\big\|_{\mathrm{sum}} = 2\|A\|_{\max}\|h\|_1. \end{aligned} \tag{208}$$

Therefore, we conclude that $(v_c - v_y)^2 \leq 4\|A\|_{\max}^2 \|h\|_1^2$. Substituting this into (207) gives

$$v^\top C_s v \leq 4\|A\|_{\max}^2 \|h\|_1^2 \sum_{c \neq y} s_c = 4(1 - s_y)\|h\|_1^2 \|A\|_{\max}^2. \tag{209}$$

For the Frobenius-norm bound, it holds for every $c \neq y$ that

$$|v_c - v_y| = \big|\big\langle A, (e_c - e_y)h^\top \big\rangle_F\big| \leq \|A\|_F \big\|(e_c - e_y)h^\top\big\|_F = \sqrt{2}\|A\|_F \|h\|_2. \tag{210}$$

Thus, it holds that $(v_c - v_y)^2 \leq 2\|A\|_F^2 \|h\|_2^2$. Substituting this inequality into (207) yields

$$v^\top C_s v \leq 2\|A\|_F^2 \|h\|_2^2 \sum_{c \neq y} s_c = 2(1 - s_y)\|h\|_2^2 \|A\|_F^2. \tag{211}$$

□

The notation $\nabla^2 L(W)[A, A]$ denotes the second directional derivative of L at W in the matrix direction A:

$$\nabla^2 L(W)[A, A] := \frac{\mathrm{d}^2}{\mathrm{d}s^2} L(W + sA)\Big|_{s=0}.$$

Equivalently, after vectorizing the matrices, we have

$$\nabla^2 L(W)[A, A] = \mathrm{vec}(A)^\top \nabla^2_{\mathrm{vec}(W)} L(W)\, \mathrm{vec}(A).$$

The Hessian identity in (212) follows from Fan et al. (2025, Lemma 10). Applying Lemma 36 to (212) leads to (213) and (214). Finally, we derive (215) via Taylor expansion. Lemma 37 is analogous to Lemma 13 in Section B.

Lemma 37. *For every $W, A \in \mathbb{R}^{K \times d}$, we have*

$$\nabla^2 L(W)[A, A] = \frac{1}{n} \sum_{i=1}^{n} (Ax_i)^\top \big(\mathrm{diag}(p_i(W)) - p_i(W)p_i(W)^\top\big)(Ax_i). \tag{212}$$

It satisfies

$$0 \leq \nabla^2 L(W)[A, A] \leq R_1^2 G(W)\, \|A\|_{\max}^2 \tag{213}$$

and

$$0 \leq \nabla^2 L(W)[A, A] \leq R_2^2 G(W)\, \|A\|_F^2 . \tag{214}$$

Consequently, for $p \in \{\max, F\}$, it holds that

$$L(U) \leq L(V) + \langle \nabla L(V), U - V \rangle + \frac{R_p^2}{2} e^{R_p \|U - V\|_p} G(V)\, \|U - V\|_p^2 , \tag{215}$$

where $R_{\max} := R_1$ and $R_F := R_2$.

Proof. Denote $\ell_y(z) := -\log\big(\mathrm{softmax}(z)_y\big)$. Following Fan et al. (2025, Lemma 10), we have

$$\nabla^2 \ell_y(z) = \mathrm{diag}(\mathrm{softmax}(z)) - \mathrm{softmax}(z)\,\mathrm{softmax}(z)^\top.$$

Applying this identity with $z = Wx_i$ and directional perturbation Ax_i, and then averaging over $i \in [n]$, we obtain

$$\nabla^2 L(W)[A, A] = \frac{1}{n}\sum_{i=1}^{n} (Ax_i)^\top \left(\mathrm{diag}(p_i(W)) - p_i(W)p_i(W)^\top\right)(Ax_i).$$

For each $i \in [n]$, we apply Lemma 36 with $s = p_i(W)$, $y = y_i$, and $h = x_i$. The max-norm part of Lemma 36 gives

$$0 \le (Ax_i)^\top \left(\mathrm{diag}(p_i(W)) - p_i(W)p_i(W)^\top\right)(Ax_i) \le R_1^2\big(1 - p_{i,y_i}(W)\big)\|A\|_{\max}^2.$$

Averaging this inequality over $i \in [n]$ and using the Hessian identity, we obtain

$$\nabla^2 L(W)[A, A] \le \frac{R_1^2\|A\|_{\max}^2}{n}\sum_{i=1}^{n}\big(1 - p_{i,y_i}(W)\big) = R_1^2 G(W)\|A\|_{\max}^2,$$

where the last equality follows from $G(W) = \frac{1}{n}\sum_{i=1}^{n}\big(1 - p_{i,y_i}(W)\big)$.

Similarly, the Frobenius-norm part of Lemma 36 gives

$$0 \le (Ax_i)^\top \left(\mathrm{diag}(p_i(W)) - p_i(W)p_i(W)^\top\right)(Ax_i) \le R_2^2\big(1 - p_{i,y_i}(W)\big)\|A\|_F^2.$$

Averaging over $i \in [n]$ yields

$$\nabla^2 L(W)[A, A] \le \frac{R_2^2\|A\|_F^2}{n}\sum_{i=1}^{n}\big(1 - p_{i,y_i}(W)\big) = R_2^2 G(W)\|A\|_F^2.$$

For the final claim, we fix $U, V \in \mathbb{R}^{K\times d}$ and define $\Delta := U - V$. Consider the scalar function $\phi(\tau) := L(V + \tau\Delta)$ where $\tau \in [0, 1]$. Then, it holds that $\phi(0) = L(V)$, $\phi(1) = L(U)$, $\phi'(0) = \langle \nabla L(V), \Delta\rangle_F$, and $\phi''(\tau) = \nabla^2 L(V + \tau\Delta)[\Delta, \Delta]$. Therefore, from Taylor's theorem, we obtain

$$L(U) = L(V) + \langle \nabla L(V), \Delta\rangle_F + \int_0^1 (1 - \tau)\nabla^2 L(V + \tau\Delta)[\Delta, \Delta]\,\mathrm{d}\tau. \tag{216}$$

Let $p \in \{\max, F\}$ and define $R_{\max} := R_1$, and $R_F := R_2$. Applying the corresponding Hessian bound at $V + \tau\Delta$ gives

$$\nabla^2 L(V + \tau\Delta)[\Delta, \Delta] \le R_p^2 G(V + \tau\Delta)\|\Delta\|_p^2.$$

By Lemma 35,

$$G(V + \tau\Delta) \le \exp\left(\tau R_p\|\Delta\|_p\right) G(V) \le \exp\left(R_p\|\Delta\|_p\right) G(V),$$

where the last inequality follows from $0 \le \tau \le 1$. Consequently, it holds that

$$\nabla^2 L(V + \tau\Delta)[\Delta, \Delta] \le R_p^2 e^{R_p\|\Delta\|_p} G(V)\|\Delta\|_p^2.$$

Substituting this bound into (216) yields

$$L(U) \le L(V) + \langle \nabla L(V), \Delta\rangle_F + R_p^2 e^{R_p\|\Delta\|_p} G(V)\|\Delta\|_p^2 \int_0^1 (1 - \tau)\,\mathrm{d}\tau.$$

Since $\int_0^1 (1 - \tau)\,\mathrm{d}\tau = \frac{1}{2}$, we conclude that

$$L(U) \le L(V) + \langle \nabla L(V), U - V\rangle_F + \frac{R_p^2}{2} e^{R_p\|U - V\|_p} G(V)\|U - V\|_p^2.$$

□

For the multiclass setting, we consider the same Adam update as (3), now applied entrywise to $K \times d$ matrices. The bound from Lemma 11 remains valid.

Lemma 38. *In the multiclass setting, Adam's update satisfies* $\|D_t\|_{\max} \le C_D$ *where* $C_D := \sqrt{\frac{1-\beta_1}{1-\beta_2}}$.

We first derive an intermediate result that will be useful in the proof of Lemma 40.

Lemma 39. *For any* $U, V \in \mathbb{R}^{K\times d}$*, we have*

$$\|\nabla L(U) - \nabla L(V)\|_{\text{sum}} \le R_1 G(V)\left(e^{R_1\|U-V\|_{\max}} - 1\right), \tag{217}$$

and

$$\|\nabla L(U) - \nabla L(V)\|_F \le R_2 G(V)\left(e^{R_2\|U-V\|_F} - 1\right). \tag{218}$$

Proof. Denote $r := R_1\|U-V\|_{\max}$. Equation (191) implies that

$$|p_{ic}(U) - p_{ic}(V)| \le p_{ic}(V)(e^r - 1).$$

Using the gradient decomposition from Lemma 33,

$$\begin{aligned}\|\nabla L(U) - \nabla L(V)\|_{\text{sum}} &\le \frac{1}{n}\sum_i \sum_{c\neq y_i} |p_{ic}(U) - p_{ic}(V)|\,\|Z_{ic}\|_{\text{sum}} \\ &\le R_1 G(V)(e^r - 1),\end{aligned}$$

which proves (217). For the Frobenius norm, the argument is the same after using (192) and $\|Z_{ic}\|_F \le R_2$. □

The following lemma is analogous to Lemma 14 in Section B. It will be used in the proof of Lemma 41.

Lemma 40. *Suppose that* $0 \le \beta_1 \le \beta_2 < 1$*. For every* $0 \le r \le t$*, we have*

$$\|g_{t-r} - g_t\|_{\text{sum}} \le R_1 G_t\left[\exp\left(R_1 C_D \sum_{s=t-r}^{t-1}\eta_s\right) - 1\right]. \tag{219}$$

Proof. By Lemma 38, we obtain

$$\|W_t - W_{t-r}\|_{\max} \le C_D \sum_{s=t-r}^{t-1}\eta_s.$$

Applying (217) with $U = W_{t-r}$ and $V = W_t$ finishes the proof. □

Lemma 41 is analogous to Lemma 15 and Lemma 16.

Lemma 41. *Suppose that* $0 \le \beta_1 \le \beta_2 < 1$*. There exist a finite time* t_{ema} *and constants* $C_m, C_v < \infty$ *(independent of* ϵ*), such that for every* $t \ge t_{\text{ema}}$*,*

$$\|m_t - g_t\|_{\text{sum}} \le C_m \eta_t G_t \tag{220}$$

and

$$\|\sqrt{v_t} - |g_t|\|_{\text{sum}} \le C_v \eta_t G_t. \tag{221}$$

Proof. The only loss-specific ingredient needed in the binary proofs of Lemma 15 and Lemma 16 is the gradient-difference bound in Lemma 14. That ingredient is now supplied by Lemma 40. Apply binary Lemma 15 with $q = \beta_1$, and binary Lemma 16 with $q = \sqrt{\beta_2}$, in each case with $c = R_1 C_D$. Their EMA calculations are unchanged after vectorizing the $K \times d$ gradient. Therefore, (220) and (221) hold after $t \ge t_m := t_0(\beta_1, R_1 C_D, a, \eta_0)$ and $t \ge t_v := t_0(\sqrt{\beta_2}, R_1 C_D, a, \eta_0)$, where $t_0(q, c, a, \eta_0)$ is defined in Proposition 9. The constants C_m and C_v are $C_m = R_1 C_{\text{geo}}(\beta_1, R_1 C_D, a, \eta_0)$ and $C_v = R_1 C_{\text{geo}}(\sqrt{\beta_2}, R_1 C_D, a, \eta_0)$. Setting $t_{\text{ema}} := \max\{t_m, t_v\}$ finishes the proof. □

As in the binary case, we can define the entrywise matrix soft-sign map for any matrix $A \in \mathbb{R}^{K\times d}$.

$$\sigma_\epsilon(A)_{kj} := \frac{A_{kj}}{|A_{kj}| + \epsilon}, \qquad \epsilon > 0.$$

The following lemma is analogous to Lemma 17 in Section B.

Lemma 42. *There is a constant $C_{\mathrm{ema}} < \infty$ (independent of ϵ) such that the following holds for every $t \geq t_{\mathrm{ema}}$*

$$\|g_t \odot (D_t - \sigma_\epsilon(g_t))\|_{\mathrm{sum}} \leq C_{\mathrm{ema}} \eta_t G_t, \tag{222}$$

$$\epsilon \, \|D_t - \sigma_\epsilon(g_t)\|_F \leq C_{\mathrm{ema}} \eta_t G_t, \tag{223}$$

$$|\langle g_t, D_t - \sigma_\epsilon(g_t)\rangle| \leq C_{\mathrm{ema}} \eta_t G_t. \tag{224}$$

Proof. We first vectorize the $K \times d$ matrices. The proof of Lemma 17 in the binary setting is entirely coordinatewise and uses only (220), (221), and the bound $|m_t[j]| \leq C_D\sqrt{v_t[j]}$ from Lemma 11. Applying that argument to the Kd vectorized coordinates proves (222) and its inner-product consequence. The same coordinatewise decomposition multiplied by ϵ (followed by $\|A\|_F \leq \|A\|_{\mathrm{sum}}$) proves (223). □

Lemma 43. *For any $A \in \mathbb{R}^{K\times d}$, it holds that*

$$\langle A, \sigma_\epsilon(A)\rangle \geq \|A\|_{\mathrm{sum}} - Kd\epsilon, \tag{225}$$

$$\langle A, \sigma_\epsilon(A)\rangle \geq \frac{\|A\|_{\mathrm{sum}}^2}{\|A\|_{\mathrm{sum}} + Kd\epsilon}. \tag{226}$$

Proof. Based on the definition of $\sigma_\epsilon(A)$, we have $\langle A, \sigma_\epsilon(A)\rangle = \sum_{k=1}^K \sum_{j=1}^d \frac{A_{kj}^2}{|A_{kj}|+\epsilon}$. For (225), we have the following entrywise inequality

$$\frac{A_{kj}^2}{|A_{kj}| + \epsilon} = |A_{kj}| - \epsilon \frac{|A_{kj}|}{|A_{kj}| + \epsilon} \geq |A_{kj}| - \epsilon,$$

where we used $0 \leq \frac{|A_{kj}|}{|A_{kj}|+\epsilon} \leq 1$. Summing over all Kd entries yields $\langle A, \sigma_\epsilon(A)\rangle_F \geq \|A\|_{\mathrm{sum}} - Kd\epsilon$.

For (226), we can write

$$\|A\|_{\mathrm{sum}} = \sum_{k=1}^K \sum_{j=1}^d |A_{kj}| = \sum_{k=1}^K \sum_{j=1}^d \frac{|A_{kj}|}{\sqrt{|A_{kj}| + \epsilon}} \sqrt{|A_{kj}| + \epsilon}.$$

By the Cauchy–Schwarz inequality, it holds that

$$\|A\|_{\mathrm{sum}}^2 \leq \left(\sum_{k=1}^K \sum_{j=1}^d \frac{A_{kj}^2}{|A_{kj}| + \epsilon}\right) \left(\sum_{k=1}^K \sum_{j=1}^d \big(|A_{kj}| + \epsilon\big)\right) = \langle A, \sigma_\epsilon(A)\rangle \left(\|A\|_{\mathrm{sum}} + Kd\epsilon\right).$$

Since $\|A\|_{\mathrm{sum}} + Kd\epsilon > 0$, dividing both sides by this quantity gives

$$\langle A, \sigma_\epsilon(A)\rangle_F \geq \frac{\|A\|_{\mathrm{sum}}^2}{\|A\|_{\mathrm{sum}} + Kd\epsilon}.$$

□

Proposition 44. *For every $t \geq t_{\mathrm{ema}}$, it holds that*

$$L_{t+1} \leq L_t - \eta_t \langle g_t, \sigma_\epsilon(g_t)\rangle_F + C_{\mathrm{md}}^{\mathrm{mc}} \eta_t^2 G_t,$$

where $C_{\mathrm{md}}^{\mathrm{mc}} = C_{\mathrm{ema}} + \frac{R_1^2 C_D^2}{2} e^{R_1 C_D \eta_0}$.

Proof. Fix $\epsilon > 0$. Denote $g_t := \nabla L(W_t)$. Applying Lemma 37 with $U = W_{t+1}$, $V = W_t$, and $p = \max$, and noting that $U - V = -\eta_t D_t$, we obtain

$$\begin{aligned} L_{t+1} &\le L_t + \langle g_t, -\eta_t D_t \rangle_F + \frac{R_1^2}{2} \exp(R_1 \eta_t \|D_t\|_{\max})\, G_t \eta_t^2 \|D_t\|_{\max}^2 \\ &= L_t - \eta_t \langle g_t, D_t \rangle_F + \frac{R_1^2}{2} \exp(R_1 \eta_t \|D_t\|_{\max})\, G_t \eta_t^2 \|D_t\|_{\max}^2. \end{aligned} \tag{227}$$

The coordinatewise Adam bound in Lemma 38 gives $\|D_t\|_{\max} \le C_D$ where $C_D := \sqrt{\frac{1-\beta_1}{1-\beta_2}}$. Since $\eta_t \le \eta_0$, it follows that $\exp(R_1 \eta_t \|D_t\|_{\max}) \le \exp(R_1 C_D \eta_0)$ and $\|D_t\|_{\max}^2 \le C_D^2$. Therefore, (227) implies

$$L_{t+1} \le L_t - \eta_t \langle g_t, D_t \rangle_F + C_H^{\mathrm{mc}} \eta_t^2 G_t, \tag{228}$$

where $C_H^{\mathrm{mc}} := \frac{R_1^2 C_D^2}{2} \exp(R_1 C_D \eta_0)$. Next, we decompose the first-order term as

$$\langle g_t, D_t \rangle_F = \langle g_t, \sigma_\epsilon(g_t) \rangle_F + \langle g_t, D_t - \sigma_\epsilon(g_t) \rangle_F .$$

Define $\delta_t := \langle g_t, D_t - \sigma_\epsilon(g_t) \rangle_F$. Lemma 42 gives for every $t \ge t_{\mathrm{ema}}$ that $|\delta_t| \le C_{\mathrm{ema}} \eta_t G_t$. In particular, we have $\delta_t \ge -C_{\mathrm{ema}} \eta_t G_t$ and hence

$$\langle g_t, D_t \rangle_F \ge \langle g_t, \sigma_\epsilon(g_t) \rangle_F - C_{\mathrm{ema}} \eta_t G_t.$$

Multiplying both sides by $-\eta_t$ yields

$$-\eta_t \langle g_t, D_t \rangle_F \le -\eta_t \langle g_t, \sigma_\epsilon(g_t) \rangle_F + C_{\mathrm{ema}} \eta_t^2 G_t.$$

Substituting this estimate into (228), we obtain for every $t \ge t_{\mathrm{ema}}$ that

$$\begin{aligned} L_{t+1} &\le L_t - \eta_t \langle g_t, \sigma_\epsilon(g_t) \rangle_F + C_{\mathrm{ema}} \eta_t^2 G_t + C_H^{\mathrm{mc}} \eta_t^2 G_t \\ &= L_t - \eta_t \langle g_t, \sigma_\epsilon(g_t) \rangle_F + C_{\mathrm{md}}^{\mathrm{mc}} \eta_t^2 G_t, \end{aligned}$$

where $C_{\mathrm{md}}^{\mathrm{mc}} := C_H^{\mathrm{mc}} + C_{\mathrm{ema}}$.

□

I.3 Section F Analogous Lemmas

In this section, we discuss the multiclass versions of the lemmas in Section F. The following lemma is analogous to Lemma 27 in Section F. Recall that

$$m_{ic}(W) = \langle W, Z_{ic} \rangle, \qquad Z_{ic} = (e_{y_i} - e_c) x_i^\top, \qquad \rho(W) = \min_{\substack{i \in [n] \\ c \ne y_i}} m_{ic}(W).$$

Lemma 45. *For all $U, V \in \mathbb{R}^{K \times d}$, we have*

$$|\rho(U) - \rho(V)| \le R_1 \|U - V\|_{\max}, \qquad |\rho(U) - \rho(V)| \le R_2 \|U - V\|_F . \tag{229}$$

Proof. Set $\Delta := U - V$. Fix an example $i \in [n]$ and an incorrect class $c \ne y_i$. By linearity of the pairwise margin, it holds that

$$m_{ic}(U) - m_{ic}(V) = \langle U, Z_{ic} \rangle - \langle V, Z_{ic} \rangle = \left\langle \Delta, (e_{y_i} - e_c) x_i^\top \right\rangle .$$

By duality, it holds that

$$|m_{ic}(U) - m_{ic}(V)| \le \|\Delta\|_{\max} \left\| (e_{y_i} - e_c) x_i^\top \right\|_{\mathrm{sum}} .$$

Because $c \ne y_i$, we have $\|e_{y_i} - e_c\|_1 = 2$. For the rank-one matrix Z_{ic}, we have

$$\begin{aligned} \left\| (e_{y_i} - e_c) x_i^\top \right\|_{\mathrm{sum}} &= \sum_{r=1}^{K} \sum_{j=1}^{d} |(e_{y_i} - e_c)_r (x_i)_j| \\ &= \left(\sum_{r=1}^{K} |(e_{y_i} - e_c)_r| \right) \left(\sum_{j=1}^{d} |(x_i)_j| \right) \\ &= 2 \|x_i\|_1 \le R_1. \end{aligned}$$

Consequently, for every $i \in [n]$ and $c \neq y_i$, it holds that

$$|m_{ic}(U) - m_{ic}(V)| \leq R_1 \|U - V\|_{\max}.$$

Let $\delta_{\max} := R_1\|U - V\|_{\max}$. The preceding inequality implies $m_{ic}(U) \geq m_{ic}(V) - \delta_{\max}$. Since $m_{ic}(V) \geq \rho(V)$, it follows that for every $i \in [n]$ and $c \neq y_i$, $m_{ic}(U) \geq \rho(V) - \delta_{\max}$. Taking the minimum over all pairs (i, c) gives $\rho(U) \geq \rho(V) - \delta_{\max}$. Interchanging U and V gives $\rho(V) \geq \rho(U) - \delta_{\max}$. Combining the two inequalities yields

$$|\rho(U) - \rho(V)| \leq \delta_{\max} = R_1\|U - V\|_{\max}.$$

For the Frobenius-norm, we have

$$|m_{ic}(U) - m_{ic}(V)| \leq \|\Delta\|_F \left\|(e_{y_i} - e_c)x_i^\top\right\|_F.$$

For every rank-one matrix $ab^\top$, it holds that $\|ab^\top\|_F = \|a\|_2\|b\|_2$. Consequently, we have

$$\left\|(e_{y_i} - e_c)x_i^\top\right\|_F = \|e_{y_i} - e_c\|_2\|x_i\|_2 = \sqrt{2}\,\|x_i\|_2 \leq R_2.$$

Therefore, it holds that $|m_{ic}(U) - m_{ic}(V)| \leq R_2\|U - V\|_F$. The rest of the proof follows the same steps as the max-norm case and we obtain

$$|\rho(U) - \rho(V)| \leq R_2\|U - V\|_F.$$

□

Based on Lemma 45, we can derive the following lemma which is analogous to Lemma 28 in Section F.

Lemma 46. *For any nonzero matrices $U, V \in \mathbb{R}^{K\times d}$, it holds*

$$\left|\frac{\rho(U)}{\|U\|_{\max}} - \frac{\rho(V)}{\|V\|_{\max}}\right| \leq \frac{2R_1\|U - V\|_{\max}}{\|V\|_{\max}},$$

and

$$\left|\frac{\rho(U)}{\|U\|_F} - \frac{\rho(V)}{\|V\|_F}\right| \leq \frac{2R_2\|U - V\|_F}{\|V\|_F}.$$

Proof. The proof follows the same steps as Lemma 28 by replacing the vector ℓ_∞- and ℓ_2-norm with the matrix max- and Frobenius-norms respectively, and using Lemma 45 instead of Lemma 27. □

Next, we extend Lemma 26 to the multiclass setting. We define the max-norm solution set in this case as

$$\mathcal{M}_{\max} := \{U \in \mathbb{R}^{K\times d} : \|U\|_{\max} \leq 1,\ \rho(U) = \gamma_{\max}\}.$$

Lemma 47 (Multiclass Hoffman system)**.** *Let $z_{ic} := \operatorname{vec}(Z_{ic}) \in \mathbb{R}^{Kd}$, and let $Z_{\mathrm{mc}} \in \mathbb{R}^{n(K-1)\times Kd}$ be the matrix containing the row vectors $z_{ic}^\top$, and define*

$$A_H := \begin{bmatrix} I_{Kd} \\ -I_{Kd} \\ -Z_{\mathrm{mc}} \end{bmatrix}, \qquad c_H := \begin{bmatrix} \mathbf{1}_{Kd} \\ \mathbf{1}_{Kd} \\ -\gamma_{\max}\mathbf{1}_{n(K-1)} \end{bmatrix}.$$

There is a finite constant $H_{2,\infty}(A_H)$, depending only on A_H, such that

$$\operatorname{dist}_F(W, \mathcal{M}_{\max}) \leq H_{2,\infty}(A_H) \left\|\left[A_H \operatorname{vec}(W) - c_H\right]_+\right\|_\infty, \tag{230}$$

where $[\cdot]_+$ is entrywise and $\operatorname{dist}_F(W, \mathcal{M}_{\max}) := \inf_{U\in\mathcal{M}_{\max}} \|W - U\|_F$.

Proof. As in the binary case, it is enough to verify that $\mathcal{M}_{\max} = \{W : A_H \operatorname{vec}(W) \leq c_H\}$. The first two blocks of $A_H \operatorname{vec}(W) \leq c_H$ are equivalent to $\|W\|_{\max} \leq 1$, and the last block is equivalent to $m_{ic}(W) \geq \gamma_{\max}$ for all (i, c) with $c \neq y_i$. Consequently, we have $\rho(W) \geq \gamma_{\max}$. The reverse inequality $\rho(W) \leq \gamma_{\max}$ follows directly from the definition of $\gamma_{\max}$. Putting things together, we obtain

$$\{W : A_H \operatorname{vec}(W) \leq c_H\} = \mathcal{M}_{\max}.$$

The remaining arguments are the same as those in Lemma 26. □

I.4 Main Results

Throughout this section, we assume multiclass separability, $0 \le \beta_1 \le \beta_2 < 1$, and $\epsilon \in (0, 1]$. The step size schedule is $\eta_t = \eta_0(t+1)^{-a}$ with $\eta_0 > 0$ and $a \in (1/3, 1]$, satisfying Proposition 9. The loss is the cross-entropy loss defined in (183). Recall that $R_1 = 2B_1$ and $R_2 = \sqrt{2}B_2$ where $B_1 = \max_{i\in[n]}\|x_i\|_1$ and $B_2 = \max_{i\in[n]}\|x_i\|_2$. Similar to the binary case, we define the sign and gradient residuals in the multiclass setting by

$$\lambda_t := \frac{G_t}{\epsilon}, \qquad r_t^{\mathrm{sgn}} := \frac{\big\|g_t \odot \big(D_t - \mathrm{sign}(g_t)\big)\big\|_{\mathrm{sum}}}{\|g_t\|_{\mathrm{sum}}}, \qquad r_t^{\mathrm{grad}} := \frac{\|\epsilon D_t - g_t\|_F}{\|g_t\|_F}.$$

Throughout the multiclass extension, we make the following substitutions:

$$\big(\|\cdot\|_\infty, \|\cdot\|_1, \|\cdot\|_2; \gamma_\infty, \gamma_2\big) \longrightarrow \big(\|\cdot\|_{\max}, \|\cdot\|_{\mathrm{sum}}, \|\cdot\|_F; \gamma_{\max}, \gamma_F\big).$$

Furthermore, we replace the constant d with Kd.

The following lemma is the multiclass version of Lemma 1. It can be directly obtained by applying the coordinatewise proof of Lemma 1 to the Kd entries.

Lemma 48 (Multiclass accumulated radius bound)**.** *Suppose that $\|\nabla L(W)\|_{\max} \le B_g$ for every $W \in \mathbb{R}^{K\times d}$. There exists a constant $C_{\mathrm{rad}} \ge 0$ (depending only on β_1 and β_2) such that for every $\epsilon > 0$ and integers $0 \le r < T$,*

$$\left\|\sum_{t=r}^{T-1} \eta_t D_t\right\|_{\max} \le S_T - S_r + C_{\mathrm{rad}}\eta_r\left[1 + \log\frac{B_g^2 + \epsilon^2}{(1-\beta_2)\epsilon^2}\right]. \tag{231}$$

In the multiclass setting, we can take $B_g = R_1$ since $\|\nabla L(W)\|_{\max} \le R_1 G(W) \le R_1$.

The following proposition is the multiclass counterpart of Proposition 2.1. Its proof follows the same steps, with Lemma 7, Lemma 8, Lemma 11, Lemma 12, Lemma 13, Lemma 17, Lemma 18, and Proposition 19 replaced by Lemma 33, Lemma 34, Lemma 38, Lemma 35, Lemma 37, Lemma 42, Lemma 43, and Proposition 44, respectively.

Proposition 49. *For every fixed $\epsilon \in (0, 1]$, the corresponding Adam trajectory satisfies $L_t \to 0$ and $G_t \to 0$ as $t \to \infty$. Moreover, there exist an integer $T_* > t_{\mathrm{ema}}$ and a constant $g_* > 0$ (both independent of ϵ) such that for every $\epsilon \in (0, 1]$,*

$$G_{T_*} \ge g_*, \qquad L_t \le \frac{\log 2}{n}, \qquad \frac{1}{2}L_t \le G_t \le L_t \quad \textit{for every } t \ge T_*.$$

Denote $\ell_\epsilon = \log\frac{1}{\epsilon}$. Recall that $G_t = G(W_t) = \frac{1}{n}\sum_{i=1}^n \sum_{c\neq y_i} p_{ic}(W_t)$ where $p_{ic}(W_t)$ is the c-th entry of the vector $p_i(W_t) = \mathrm{softmax}(W_t x_i)$. Let T_* be the common low-loss time in Proposition 49. The stopping times in the multiclass setting are defined in the same way as the binary setting using the multiclass G_t:

$$T_1^-(\epsilon; b) := \inf\{t \ge T_* : G_t \le \epsilon\ell_\epsilon^b\}, \qquad T_1^\delta(\epsilon) := \inf\{t \ge T_* : G_t \le \delta\epsilon\},$$
$$T_1^+(\epsilon; b) := \inf\{t \ge T_* : G_t \le \epsilon\ell_\epsilon^{-b}\}.$$

The following proposition is the multiclass counterpart of Proposition C.1. Its proof follows the same steps, with Lemma 7 and Lemma 17 replaced by Lemma 33 and Lemma 42 respectively.

Proposition 50. *Denote $C_{\mathrm{loc}}^{\mathrm{mc}} := C_{\mathrm{ema}} \max\left\{\frac{1}{\gamma_{\max}}, \frac{1}{\gamma_F}\right\}$. For every $\epsilon > 0$ and $t \ge t_{\mathrm{ema}}$, we have*

$$\left[\frac{1}{1 + R_1\lambda_t} - C_{\mathrm{loc}}^{\mathrm{mc}}\eta_t\right]_+ \le r_t^{\mathrm{sgn}} \le \min\left\{1, \frac{Kd}{\gamma_{\max}\lambda_t}\right\} + C_{\mathrm{loc}}^{\mathrm{mc}}\eta_t,$$

and

$$\left[1 - \frac{\sqrt{Kd}}{\gamma_F\lambda_t} - C_{\mathrm{loc}}^{\mathrm{mc}}\eta_t\right]_+ \le r_t^{\mathrm{grad}} \le \frac{R_1\lambda_t}{1 + R_1\lambda_t} + C_{\mathrm{loc}}^{\mathrm{mc}}\eta_t,$$

where $[x]_+ := \max\{x, 0\}$.

The following proposition is the multiclass counterpart of Proposition C.2. Its proof follows the same steps, with Lemma 12 and Proposition 2.1 replaced by Lemma 35 and Proposition 49 respectively.

Proposition 51. *For any $b, \delta \in (0,1)$, there exists $\epsilon_0(b,\delta) > 0$ such that for every $0 < \epsilon \le \epsilon_0(b,\delta)$, we have*

$$T_1^-(\epsilon; b) < T_1^\delta(\epsilon) < T_1^+(\epsilon; b).$$

The following theorem is the multiclass counterpart of Theorem 2. Its proof follows the same steps, replacing Proposition 2.1, Proposition C.1, Proposition C.2, and Lemma 12 with Proposition 49, Proposition 50, Proposition 51, and Lemma 35, respectively.

Theorem 52. *For any $b, \delta \in (0,1)$, there exists a constant $\tilde{\epsilon}_{\mathrm{res}}(b,\delta)$ such that for every $0 < \epsilon \le \tilde{\epsilon}_{\mathrm{res}}(b,\delta)$, we have*

$$\max\left\{r^{\mathrm{sgn}}_{T_1^-(\epsilon;b)}, r^{\mathrm{grad}}_{T_1^+(\epsilon;b)}\right\} = \begin{cases} O\left(\ell_\epsilon^{-b} + \ell_\epsilon^{-a/(1-a)}\right), & a \in (1/3, 1), \\ O\left(\ell_\epsilon^{-b}\right), & a = 1. \end{cases}$$

Denote $K_0 := R_1 C_D$. We also have

$$r^{\mathrm{grad}}_{T_1^\delta(\epsilon)} \le \frac{R_1\delta}{1+R_1\delta} + \begin{cases} O\left(\ell_\epsilon^{-a/(1-a)}\right), & a \in (1/3, 1), \\ O\left(\epsilon^{1/(2K_0\eta_0)}\right), & a = 1. \end{cases}$$

Let $\widehat{\gamma}_p(W) := \frac{\rho(W)}{\|W\|_p}, \quad p \in \{\max, F\}$. The max-norm optimal solution set is defined as

$$\mathcal{M}_{\max} := \{U \in \mathbb{R}^{K\times d} : \|U\|_{\max} \le 1,\ \rho(U) = \gamma_{\max}\}.$$

We further define a few geometric quantities:

$$\gamma_{F|\max} := \max_{U\in\mathcal{M}_{\max}} \frac{\rho(U)}{\|U\|_F}, \qquad \Delta^{\mathrm{mc}}_{\mathrm{geom}} := \gamma_F - \gamma_{F|\max}, \qquad r_\epsilon := \begin{cases} \ell_\epsilon^{-a}, & a \in (1/3,1), \\ (\log \ell_\epsilon)/\ell_\epsilon, & a = 1. \end{cases}$$

The multiclass geometric separation $\triangle^{\mathrm{mc}}_{\mathrm{geom}}$ is the difference between the Frobenius data margin and the best attainable Frobenius margin by the classifiers in $\mathcal{M}_{\max}$.

Denote $T_1 := T_1^\delta(\epsilon)$. Define $u_t := \frac{g_t}{\|g_t\|_F}$, $\theta_t := \frac{\eta_t\|g_t\|_F}{\epsilon}$, and $e_t := \frac{\epsilon D_t - g_t}{\|g_t\|_F}$. The post-crossover update of Adam can be written as $W_{t+1} = W_t - \theta_t(u_t + e_t)$. As in the binary case, we further define the accumulated original and effective step sizes as:

$$V_t^\delta(\epsilon) := \sum_{s=T_1}^{t-1} \eta_s, \qquad H_t^\delta(\epsilon) := \sum_{s=T_1}^{t-1} \theta_s.$$

The following proposition is the multiclass counterpart of Proposition 3.1. Its proof uses Proposition 49 in place of Proposition 2.1, Lemma 33 and Lemma 42 in place of Lemma 7 and Lemma 17, and Lemma 35 for the trajectory bound from Lemma 12. The scalar stepsize estimate in Lemma 24 applies unchanged. The required two-sided post-crossover loss estimate follows by adapting Part 3 of the proof of Theorem 4 to the multiclass setting.

Proposition 53. *For every $\delta \in (0,1)$, there exists a constant $\tilde{\epsilon}_1^{\mathrm{mc}}(\delta)$ such that, for every $0 < \epsilon \le \tilde{\epsilon}_1^{\mathrm{mc}}(\delta)$ and every $t \ge T_1^\delta(\epsilon)$, we have $L_t \asymp G_t \asymp \frac{\epsilon}{1+V_t^\delta(\epsilon)}$, $H_t^\delta(\epsilon) \asymp \log(1+V_t^\delta(\epsilon))$, and $\sum_{s=T_1^\delta(\epsilon)}^{\infty} \theta_s\|e_s\|_F \le E_0^{\mathrm{mc}}$, where E_0^{mc} is a constant that depends on the data and the step-size schedule.*

The following theorem is the multiclass counterpart of Theorem 29. Its proof follows the same steps, replacing Lemma 12, Lemma 1, Lemma 7, Lemma 8, Lemma 11, Lemma 18, Lemma 26, and Lemma 27 with Lemma 35, Lemma 48, Lemma 33, Lemma 34, Lemma 38, Lemma 43, Lemma 47, and Lemma 45, respectively, and Proposition 19 and Proposition 2.1 with Proposition 44 and Proposition 49. The inequality in Lemma 25 applies to the Frobenius norm after vectorization.

Theorem 54. *For $\delta \in (0,1)$, there exists a constant $\tilde{\epsilon}_{2,\mathrm{mc}}(\delta)$ such that for every $0 < \epsilon \le \tilde{\epsilon}_{2,\mathrm{mc}}(\delta)$, we have*

$$\gamma_{\max} - \widehat{\gamma}_{\max}(W_{T_1^\delta(\epsilon)}) \le O(r_\epsilon), \qquad \gamma_F - \widehat{\gamma}_F(W_{T_1^\delta(\epsilon)}) \ge \Delta^{\mathrm{mc}}_{\mathrm{geom}} - O(r_\epsilon).$$

The following theorem is the multiclass counterpart of Theorem 31. Its proof follows the same steps, and uses Propositions 44, 49 and 51 in place of Proposition 19 and Propositions 2.1 and C.2, Lemmas 33, 34, 38, 43 and 46 in place of Lemmas 7, 8, 11, 18 and 28, Lemma 35 in place of Lemma 12, and Theorem 54 in place of Theorem 29, respectively.

Theorem 55. *For any $\delta, b \in (0,1)$, there exists a constant $\epsilon'_{2,\mathrm{mc}}(\delta, b)$ such that for every $0 < \epsilon \le \epsilon'_{2,\mathrm{mc}}(\delta, b)$, we have*

$$\gamma_{\max} - \widehat{\gamma}_{\max}(W_{T_1^-(\epsilon;b)}) \le O(r_\epsilon), \qquad \gamma_F - \widehat{\gamma}_F(W_{T_1^-(\epsilon;b)}) \ge \Delta_{\mathrm{geom}} - O(r_\epsilon).$$

The following proposition is the multiclass counterpart of Proposition 3.2. Its proof follows the same steps, using Proposition 53, Theorem 54, Proposition 49, Lemma 34, Lemma 46 in place of Proposition 3.1, Theorem 29, Proposition 2.1, Lemma 8, Lemma 28, respectively.

Proposition 56. *For every $\delta \in (0,1)$, there exists $\tilde{\epsilon}_3^{\mathrm{mc}}(\delta) > 0$ such that for $0 < \epsilon \le \tilde{\epsilon}_3^{\mathrm{mc}}(\delta)$ and $t \ge T_1^\delta(\epsilon)$, we have*

$$\gamma_{\max} - \widehat{\gamma}_{\max}(W_t) \le O\left(r_\epsilon + \frac{H_t^\delta(\epsilon)}{\ell_\epsilon}\right), \qquad \gamma_F - \widehat{\gamma}_F(W_t) \ge \Delta_{\mathrm{geom}}^{\mathrm{mc}} - O\left(r_\epsilon + \frac{H_t^\delta(\epsilon)}{\ell_\epsilon}\right).$$

The following corollary is the multiclass counterpart of Corollary 32. Its proof follows the same steps, and uses Proposition 51, Proposition 53, Lemma 35, and Proposition 56 in place of Proposition C.2, Proposition 3.1, Lemma 12, and Proposition 3.2, respectively.

Corollary 57. *For any $\delta, b \in (0,1)$, there exists a constant $\epsilon''_{2,\mathrm{mc}}(\delta, b)$ such that for every $0 < \epsilon \le \epsilon''_{2,\mathrm{mc}}(\delta, b)$, we have*

$$\gamma_{\max} - \widehat{\gamma}_{\max}(W_{T_1^+(\epsilon;b)}) \le O(r_\epsilon), \qquad \gamma_F - \widehat{\gamma}_F(W_{T_1^+(\epsilon;b)}) \ge \Delta_{\mathrm{geom}} - O(r_\epsilon).$$

We state two additional lemmas that are needed for the proof of Theorem 60. The following lemma is the multiclass counterpart of Lemma 20. Its proof follows the same steps, and uses Proposition 49, Lemma 33, Lemma 43, Proposition 44, Lemma 38 in place of Proposition 2.1, Lemma 7, Lemma 18, Proposition 19, and Lemma 11, respectively.

Lemma 58. *For every $\zeta > 0$, there exists an integer t_ζ, (independent of ϵ) such that for every $\epsilon \in (0,1]$, we have*

$$L_t \le \zeta, \qquad t \ge t_\zeta.$$

The following lemma is the multiclass counterpart of Lemma 21. Its proof follows the same steps, using Proposition 49, Lemma 35, and Lemma 34 in place of Proposition 2.1, Lemma 12, and Lemma 8, respectively.

Lemma 59. *Let $h_\epsilon > 0$ satisfy $h_\epsilon \to 0$ as $\epsilon \downarrow 0$. Define $\tau_\epsilon := \inf\{t \ge T_* : G_t \le h_\epsilon\}$ where T_* is the time in Proposition 49. Then, we have $\tau_\epsilon < \infty$ for every $\epsilon \in (0,1]$, and*

$$\tau_\epsilon \to \infty, \qquad \frac{G_{\tau_\epsilon}}{h_\epsilon} \to 1, \qquad \frac{L_{\tau_\epsilon}}{h_\epsilon} \to 1, \qquad \text{as } \epsilon \downarrow 0.$$

The following theorem is the multiclass counterpart of Theorem 4. Its proof follows the same steps, using Lemma 48, Lemma 33, Lemma 34, Lemma 38, Lemma 43, Lemma 58, Lemma 59, Proposition 44, Proposition 49, Lemma 35, Lemma 37, Lemma 42 in place of Lemma 1, Lemma 7, Lemma 8, Lemma 11, Lemma 18, Lemma 20, Lemma 21, Proposition 19, Proposition 2.1, Lemma 10, Lemma 13, and Lemma 17, respectively.

Theorem 60. *For every fixed $b, \delta \in (0,1)$, as $\epsilon \downarrow 0$, we have $S_{T_1^\delta(\epsilon)} - S_{T_1^-(\epsilon;b)} = O(\log \ell_\epsilon)$, $S_{T_1^+(\epsilon;b)} - S_{T_1^\delta(\epsilon)} = \Theta(\ell_\epsilon^b)$, and*

$$S_{T_1^\delta(\epsilon)} - S_{T_*} = \frac{\ell_\epsilon}{\gamma_{\max}} + o(\ell_\epsilon).$$

Finally, we discuss Euclidean optimality in the multiclass case. For $\zeta \in (0, \Delta_{\mathrm{geom}}^{\mathrm{mc}})$, define

$$\tau_\zeta^\delta(\epsilon) := \inf\left\{t \ge T_1^\delta(\epsilon) : \gamma_F - \widehat{\gamma}_F(W_t) \le \zeta\right\}.$$

The following theorem is the multiclass counterpart of Theorem 5. Its proof follows the same steps, using Proposition 53, Proposition 49, Lemma 33, Lemma 34, Lemma 35, Lemma 37, Lemma 46, and Theorem 54 in place of Proposition 3.1, Proposition 2.1, Lemma 7, Lemma 8, Lemma 10, Lemma 13, Lemma 28, Theorem 29, respectively.

Theorem 61. *For any $\delta \in (0,1)$ and $\zeta \in (0, \Delta^{\mathrm{mc}}_{\mathrm{geom}})$, there exists a constant $\tilde{\epsilon}_{4,\mathrm{mc}}(\delta,\zeta)$ such that the following statements hold for every $0 < \epsilon \le \tilde{\epsilon}_{4,\mathrm{mc}}(\delta,\zeta)$:*

(i) For every $t \ge T_1^\delta(\epsilon)$, we have $\gamma_F - \hat{\gamma}_F(W_t) \le O\big(\frac{1+\|W_{T_1^\delta(\epsilon)}\|_F}{1+H_t^\delta(\epsilon)}\big)$.

(ii) There exist constants $0 < c_\zeta \le C_\zeta < \infty$ such that $c_\zeta \|W_{T_1^\delta(\epsilon)}\|_F \le H^\delta_{\tau_\zeta^\delta(\epsilon)}(\epsilon) \le C_\zeta(1 + \|W_{T_1^\delta(\epsilon)}\|_F)$. Consequently, it holds that $H^\delta_{\tau_\zeta^\delta(\epsilon)}(\epsilon) = \Theta(\log \frac{1}{\epsilon})$ and $V_{\tau_\zeta^\delta(\epsilon)} = \epsilon^{-\Theta(1)}$.

The proof of Corollary 62 follows the same steps as Corollary 6.

Corollary 62. *Suppose $\Delta^{\mathrm{mc}}_{\mathrm{geom}} > 0$. Fix $\delta \in (0,1)$ and $\zeta \in (0, \Delta^{\mathrm{mc}}_{\mathrm{geom}})$. For $\eta_t = \eta_0(t+1)^{-a}$, as $\epsilon \downarrow 0$,*

$$T_1^\delta(\epsilon) = \Theta\Big(\ell_\epsilon^{1/(1-a)}\Big), \qquad \tau_\zeta^\delta(\epsilon) = \epsilon^{-\Theta(1)/(1-a)}, \qquad a \in (1/3, 1),$$
$$T_1^\delta(\epsilon) = \epsilon^{-1/(\eta_0\gamma_{\max})+o(1)}, \quad \tau_\zeta^\delta(\epsilon) = \exp\big(\epsilon^{-\Theta(1)}\big), \qquad a = 1.$$