THE IMPLICIT BIAS OF STEEPEST DESCENT ON MULTICLASS NONSEPARABLE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

We study steepest descent for linear multiclass classification with softmax cross-entropy on nonseparable data. Unlike binary nonseparable logistic regression, where the data split into separable and nonseparable examples, a multiclass example induces multiple pairwise margin constraints that may differ in separability. We therefore introduce a norm-dependent parameter-space decomposition: one component remains constant throughout the descent, an escaping component that maximizes the margin in the separable subspace, and a bounded component that converges to a finite optimum in the nonseparable subspace. We derive convergence rates of the risk to its infimum and of the normalized margin of the escaping component to the norm-dependent optimum. We also characterize *directional convergence*. When the norm is strictly convex, the max-margin direction is unique and the escaping component converges to it directionally. Without strict convexity, we find that this can *fail*: we construct a linear classification problem and an optimization norm for which the normalized escaping component has multiple distinct max-margin accumulation points, even with separable training data.

1 INTRODUCTION

Logistic regression is a standard approach to classification, in which a linear model produces logits that are mapped to class probabilities and the parameters are chosen to minimize the empirical logistic loss. Even in this basic linear setting, the loss need not attain its infimum at a finite parameter. On linearly separable binary classification data, logistic loss approaches zero only as the parameter norm diverges to infinity, while gradient descent with sufficiently small steps converges in direction to the ℓ_2 max-margin solution (Soudry et al., 2018). This implicit-bias phenomenon has since been studied for other optimization methods in separable linear classification (Nacson et al., 2019b;a; Zhang et al., 2024; Fan et al., 2025).

Nonseparable classification data can exhibit a different behavior: the loss of some training examples can vanish, while that of others converges to a positive value. For logistic regression on binary nonseparable classification data, Ji & Telgarsky (2019) identify a maximal subset of the training examples that is separable. The remaining examples span a subspace where the empirical risk has a unique finite optimum. Its orthogonal complement contains a direction that gives every separable example a positive margin while leaving the margins of the remaining examples unchanged. Gradient descent recovers that finite optimum and, when the separable subset is nonempty, converges in direction to the unique ℓ_2 max-margin direction on the orthogonal complement.

However, the example-wise orthogonal decomposition of the parameter space does not extend directly to the multiclass setting. Each training example contributes one loss term for each incorrect class. For a single example, some of these terms may vanish as the parameter norm diverges, while others converge to a nonzero value. Consequently, partitioning the data only at the level of examples cannot separate the vanishing part of the loss from the part that remains finite.

A separate challenge arises from the optimization geometry. Under gradient descent, the update is orthogonal to every direction along which the loss is unchanged, and therefore the parameter does not move in such directions. Under steepest descent with a general norm, however, a steepest direction need not be orthogonal to these loss-invariant directions and may contain a component along them. Moreover, the steepest direction need not be unique; when there are multiple valid steepest directions, the update requires a *tie-breaking rule* for selecting among them (Murota, 2004).

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.

This dependence of the dynamics on the optimization norm has been studied primarily in separable classification. For separable binary exponential loss, Gunasekar et al. (2018) prove convergence to the normalized max-margin under steepest descent with sufficiently small constant steps. Fan et al. (2025) establish margin convergence rates for normalized steepest descent with entrywise and Schatten norms on separable multiclass data. These results characterize risk and margin behavior under separability. However, when the max-margin direction is not unique, they do not establish whether the normalized parameters converge to a single direction.

Steepest descent on multiclass nonseparable data therefore presents both difficulties simultaneously: the finite and diverging parts of the loss must be identified at the level of class comparisons, while the parameter dynamics depend on the geometry induced by the optimization norm. We ask how these two effects alter the structural and convergence properties known from the separable and binary settings. We study a K-class linear classifier $\boldsymbol{x} \mapsto \boldsymbol{W}\boldsymbol{x}$ with weight matrix $\boldsymbol{W} \in \mathbb{R}^{K\times D}$ under the standard softmax cross-entropy loss, trained by steepest descent with respect to an arbitrary norm on nonseparable data. We summarize our contributions below.

- **Structural decomposition.** We develop a three-way decomposition of the parameter space. Under our decomposition, the weight matrix consists of a component that remains fixed, a bounded component associated with the part of the risk that admits a finite optimum, and an escaping component along which the parameters may diverge. The decomposition operates at the level of class comparisons and depends on the optimization norm (Section 3).
- **Risk and bounded-component convergence.** Under Assumption 4.1 on step sizes, we prove that the empirical risk converges to its infimum. The bounded component of the weight matrix converges to its unique finite optimum. We obtain explicit convergence rates for both (Section 4).
- **Margin and directional convergence.** When the risk infimum is unattained, under an additional condition on step sizes (Assumption 5.1), the escaping component diverges in norm while its normalized margin converges to the norm-dependent max-margin. When the norm restricted to the escaping subspace is strictly convex,[1] the max-margin direction is unique, and both the escaping component and the full weight matrix converge to this direction after normalization (Section 5).
- **Failure of directional convergence.** Without strict convexity, convergence of the normalized margin does not guarantee convergence of the parameter direction. We construct counterexamples both with and without movement along loss-invariant directions (Section 6). In particular, we construct a separable dataset and a norm for which the parameter direction under steepest descent fails to converge for every tie-breaking rule (Theorem 6.3).

Our risk, bounded-component, and margin convergence guarantees hold regardless of the tie-breaking rule used to select a steepest direction.

2 SETUP

We use bold uppercase letters (e.g., $\boldsymbol{X}$) for matrices and bold lowercase letters (e.g., $\boldsymbol{x}$) for vectors. The Frobenius inner product and norm are respectively denoted as $\langle \boldsymbol{A}, \boldsymbol{B}\rangle := \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B})$ and $\|\boldsymbol{A}\|_{\mathrm{F}} := \sqrt{\langle \boldsymbol{A}, \boldsymbol{A}\rangle}$. For a positive integer n, write $[n] := \{1, \dots, n\}$. For a linear subspace $\mathcal{H}$, $\mathcal{H}^\perp$ denotes its Frobenius-orthogonal complement and $\Pi_{\mathcal{H}}\boldsymbol{W} := \arg\min_{\boldsymbol{H}\in\mathcal{H}} \|\boldsymbol{W} - \boldsymbol{H}\|_{\mathrm{F}}^2$ denotes its Frobenius-orthogonal projection. For a set A, $\mathrm{span}(A)$, $\mathrm{conv}(A)$, and $\mathrm{cl}(A)$ denote its linear span, convex hull, and closure, respectively.

(Non)separable classification. Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ be the dataset with $\boldsymbol{x}_i \in \mathbb{R}^D$ and $y_i \in [K]$. Write $\boldsymbol{w}_c^\top$ for the c-th row of $\boldsymbol{W} \in \mathbb{R}^{K\times D}$. The pairwise score gaps and risk are, respectively,

$$\Delta_{ic}(\boldsymbol{W}) := (\boldsymbol{w}_c - \boldsymbol{w}_{y_i})^\top \boldsymbol{x}_i, \qquad R(\boldsymbol{W}) := \frac{1}{N}\sum\nolimits_{i=1}^N \log\left(1 + \sum\nolimits_{c\neq y_i} e^{\Delta_{ic}(\boldsymbol{W})}\right). \tag{1}$$

This is the standard multiclass softmax cross-entropy loss written in terms of pairwise score gaps. The risk is convex and globally smooth for some $\beta > 0$ (Lemma A.2). $\mathcal{D}$ is *separable* if there exists $\boldsymbol{W}$ such that $\Delta_{ic}(\boldsymbol{W}) < 0$ for all i and all $c \neq y_i$. If no such $\boldsymbol{W}$ exists, $\mathcal{D}$ is *nonseparable*.

[1] A norm is strictly convex if the unit ball with respect to that norm is a strictly convex set.

Steepest Descent. Fix a matrix norm $\|\cdot\|$. Its dual norm and norming face are

$$\|\boldsymbol{G}\|_* := \max_{\|\boldsymbol{H}\|\le 1} \langle \boldsymbol{G}, \boldsymbol{H}\rangle, \qquad \mathcal{F}_{\|\cdot\|}(\boldsymbol{G}) := \arg\max_{\|\boldsymbol{H}\|\le 1} \langle \boldsymbol{G}, \boldsymbol{H}\rangle \quad (\boldsymbol{G} \neq \boldsymbol{0}).$$

We omit the norm subscript when it is fixed. The update under steepest descent is as follows:

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta_t \|\nabla R(\boldsymbol{W}_t)\|_* \boldsymbol{S}_t, \qquad \boldsymbol{S}_t \in \mathcal{F}(\nabla R(\boldsymbol{W}_t))^2, \qquad \boldsymbol{W}_0 = 0. \tag{2}$$

Set $\boldsymbol{S}_t = 0$ when the gradient is zero. Otherwise, $\|\boldsymbol{S}_t\| = 1$ and $\langle \nabla R(\boldsymbol{W}_t), \boldsymbol{S}_t\rangle = \|\nabla R(\boldsymbol{W}_t)\|_*$.

Coercivity and convexity. Let $f : \mathcal{H} \to \mathbb{R}$ be differentiable on a linear subspace $\mathcal{H}$, equipped with the norm $\|\cdot\|$. f is *coercive* if $f(\boldsymbol{W}) \to \infty$ as $\|\boldsymbol{W}\| \to \infty$. It is *convex* if $f((1-\theta)\boldsymbol{W} + \theta\boldsymbol{W}') \le (1-\theta)f(\boldsymbol{W}) + \theta f(\boldsymbol{W}')$ for all $\boldsymbol{W}, \boldsymbol{W}'$ and $\theta \in (0,1)$. It is *strictly convex* if this inequality is strict when $\boldsymbol{W} \neq \boldsymbol{W}'$. For $\mu > 0$, it is *μ-strongly convex* on a convex set $\mathcal{A} \subseteq \mathcal{H}$ with respect to $\|\cdot\|$ if $f(\boldsymbol{W}') \ge f(\boldsymbol{W}) + \langle \nabla f(\boldsymbol{W}), \boldsymbol{W}' - \boldsymbol{W}\rangle + \frac{\mu}{2}\|\boldsymbol{W}' - \boldsymbol{W}\|^2$ for $\boldsymbol{W}, \boldsymbol{W}' \in \mathcal{A}$, where μ may depend on $\mathcal{A}$ and the gradient is taken within $\mathcal{H}$. A coercive f attains its minimum on $\mathcal{H}$; if it is also strictly convex, this minimizer is unique.

Maximum margin. Fix a norm $\|\cdot\|$, a linear subspace $\mathcal{H} \subseteq \mathbb{R}^{K\times D}$, and a nonempty set $\mathcal{J} \subseteq \{(i,c) : i \in [N],\ c \in [K] \setminus \{y_i\}\}$ of example–class comparisons. A *max-margin solution* is

$$\boldsymbol{U}_\star \in \arg\max_{\substack{\boldsymbol{U}\in\mathcal{H} \\ \|\boldsymbol{U}\|\le 1}} \min_{(i,c)\in\mathcal{J}} \{-\Delta_{ic}(\boldsymbol{U})\}.$$

Each $-\Delta_{ic}(\boldsymbol{U})$ measures the correct class's advantage over class c. The minimum selects the least favorable comparison, and the norm constraint fixes the scale. If the optimal value is positive, every solution has norm one.

3 Decomposition of parameter space

3.1 Subspace decomposition

In this section, we develop a direct-sum decomposition of our model parameter space $\mathbb{R}^{K\times D}$:

$$\mathbb{R}^{K\times D} = \mathcal{N} \oplus \mathcal{M} \oplus \mathcal{E}. \tag{3}$$

This decomposition is dependent on both the steepest descent norm and the (nonseparable) dataset, and the subspaces are not necessarily orthogonal to each other. The component in $\mathcal{N}$ remains fixed under steepest descent. The space $\mathcal{M}$ contains the "escape" directions along which the parameters can diverge while the risk decreases; it can also contain directions that leave the loss unchanged. The space $\mathcal{E}$ carries the remaining risk, which has a unique finite minimizer.

For binary nonseparable classification, Ji & Telgarsky (2019) construct an orthogonal decomposition by partitioning the training examples into a maximal linearly separable subset and its complement. The examples in the complementary subset span a subspace on which the empirical risk has a unique finite minimizer. Its orthogonal complement contains the separating directions for the maximal separable subset, and gradient descent approaches the risk infimum by combining the finite minimizer in the former subspace with a diverging max-margin component in the latter when the separable subset is nonempty. In multiclass classification, however, a single example contributes several class comparisons that need not share the same separability behavior; consequently, the same example-wise decomposition does not apply. Section C gives an explicit example.

First, we identify the directions that leave every pairwise score gap unchanged. For $\mathcal{I} := \{(i,c) : i \in [N],\ c \in [K] \setminus \{y_i\}\}$, define the space of directions that do not change the loss by

$$\mathcal{S} := \{\boldsymbol{W} : \Delta_{ic}(\boldsymbol{W}) = 0, \quad \forall (i,c) \in \mathcal{I}\}.$$

Thus, $\mathcal{S}$ is a subspace and $R(\boldsymbol{W} + t\boldsymbol{H}) = R(\boldsymbol{W})$ for $\boldsymbol{H} \in \mathcal{S}$ and all t. Moreover, $\nabla R(\boldsymbol{W}) \in \mathcal{S}^\perp$ (Lemma A.1). Take the span of all norming faces in the orthogonal complement of $\mathcal{S}$:

$$\mathcal{Z} := \operatorname{span}\left(\bigcup_{\boldsymbol{G}\in\mathcal{S}^\perp\setminus\{\boldsymbol{0}\}} \mathcal{F}(\boldsymbol{G})\right).$$

[2]The maximizer need not be unique if the norm is not strictly convex (Fan et al., 2025). Our Theorems 4.3, 5.2 and 6.3 hold for any maximizer, while other results could depend on the tie-breaking rule.

Every steepest descent update lies in $\mathcal{Z}$, although $\mathcal{Z}$ need not be orthogonal to $\mathcal{S}$. Its intersection with $\mathcal{S}$ identifies the directions of constant loss. Define

$$\mathcal{L} := \mathcal{S} \cap \mathcal{Z}, \qquad \mathcal{N} := \mathcal{S} \cap \mathcal{L}^{\perp}.$$

Then $\mathcal{S} = \mathcal{N} \oplus \mathcal{L}$ and $\mathcal{N} \perp \mathcal{L}$. Moreover, Proposition B.1 gives $\mathbb{R}^{K\times D} = \mathcal{N} \oplus \mathcal{Z}$. The $\mathcal{N}$-coordinate in this direct sum is invariant under the steepest descent. For the Frobenius norm, $\mathcal{Z} = \mathcal{S}^{\perp}$ and $\mathcal{N} = \mathcal{S}$ (Corollary B.2). We then split $\mathcal{Z}$ according to the directions along whose rays the risk does not increase. The *recession cone* is

$$\mathcal{C} := \{\boldsymbol{U} \in \mathcal{Z} : \Delta_{ic}(\boldsymbol{U}) \le 0 \; \forall (i,c) \in \mathcal{I}\}.$$

It is a closed convex polyhedral cone. Its elements are the directions in $\mathcal{Z}$ where $r \mapsto R(\boldsymbol{W} + r\boldsymbol{U})$ is nonincreasing for every $\boldsymbol{W}$ and $r \ge 0$ (Lemma D.1; see also Bertsekas (2009)). Define its span and the remaining orthogonal component by

$$\mathcal{M} := \operatorname{span}(\mathcal{C}), \qquad \mathcal{E} := \mathcal{Z} \cap \mathcal{M}^{\perp}.$$

The largest subspace in $\mathcal{C}$, called its lineality space, is $\mathcal{C} \cap (-\mathcal{C}) = \mathcal{L} = \mathcal{S} \cap \mathcal{Z}$. Hence, $\mathcal{L} \subseteq \mathcal{M} \subseteq \mathcal{Z}$ and $\mathcal{Z} = \mathcal{M} \oplus \mathcal{E}$. Therefore, Equation 3 follows from

$$\mathbb{R}^{K\times D} = \mathcal{N} \oplus \mathcal{Z} = \mathcal{N} \oplus (\mathcal{M} \oplus \mathcal{E}) = \mathcal{N} \oplus (\operatorname{span}(\mathcal{C}) \oplus \mathcal{E}),$$

with $\mathcal{L} \subseteq \mathcal{M}$; see Proposition B.3 for details. We can then decompose $\boldsymbol{W} \in \mathbb{R}^{K\times D}$ as

$$\boldsymbol{W} = \boldsymbol{W}^{\mathrm{n}} + \boldsymbol{U} + \boldsymbol{V}, \quad \boldsymbol{W}^{\mathrm{n}} \in \mathcal{N}, \quad \boldsymbol{U} \in \mathcal{M}, \quad \boldsymbol{V} \in \mathcal{E}. \tag{4}$$

Since we set $\boldsymbol{W}_0 = \boldsymbol{0}$ in Equation 2, we have $\boldsymbol{W}_t^{\mathrm{n}} \equiv \boldsymbol{0}$.

3.2 Active pairs and reduced risk

Next, we identify the optimization problem on $\mathcal{E}$ and the comparisons whose score gaps can diverge along $\mathcal{M}$. Partition the example–class pairs according to whether their score gaps can decrease along recession directions:

$$\mathcal{I}_\star := \{(i,c) \in \mathcal{I} : \Delta_{ic}(\boldsymbol{U}) = 0, \quad \forall \boldsymbol{U} \in \mathcal{C}\}, \qquad \mathcal{J}_\star := \mathcal{I} \setminus \mathcal{I}_\star.$$

We call the pairs in $\mathcal{I}_\star$ active and $\mathcal{J}_\star$ inactive. The score gaps on $\mathcal{I}_\star$ vanish on $\mathcal{M} = \operatorname{span}(\mathcal{C})$. If $\mathcal{J}_\star \neq \varnothing$, there exists $\widehat{\boldsymbol{U}} \in \mathcal{C}$ such that $\Delta_{ic}(\widehat{\boldsymbol{U}}) < 0$ for all $(i,c) \in \mathcal{J}_\star$ (Lemma D.2). When $\mathcal{J}_\star = \varnothing$, set $\widehat{\boldsymbol{U}} = \boldsymbol{0}$. Thus, $\mathcal{I}_\star = \varnothing$ is equivalent to multiclass separability: there exists a parameter $\boldsymbol{W}$ with $\Delta_{ic}(\boldsymbol{W}) < 0$ for every pair. Since the score gaps on active pairs vanish on $\mathcal{N} \oplus \mathcal{M}$, they depend only on $\boldsymbol{V} \in \mathcal{E}$. This defines the reduced risk by

$$R_0(\boldsymbol{V}) := \frac{1}{N} \sum\nolimits_{i=1}^{N} \log\left(1 + \sum\nolimits_{c:(i,c)\in\mathcal{I}_\star} e^{\Delta_{ic}(\boldsymbol{V})}\right), \qquad \boldsymbol{V} \in \mathcal{E}. \tag{5}$$

Unlike an example-wise split from Ji & Telgarsky (2019), the reduced risk retains the active comparisons of an example even when its inactive terms vanish, since $R_0(\boldsymbol{V}) = \lim_{r\to\infty} R(\boldsymbol{V} + r\widehat{\boldsymbol{U}})$.

Theorem 3.1 (Reduced risk and attainment)**.** *The reduced risk R_0 is coercive and strictly convex on $\mathcal{E}$, and strongly convex on every compact convex subset of $\mathcal{E}$. It has a unique minimizer $\boldsymbol{V}_\star$, and*

$$R_\star := \inf_{\boldsymbol{W}} R(\boldsymbol{W}) = R_0(\boldsymbol{V}_\star).$$

The infimum of the risk on the full parameter space is attained if and only if $\mathcal{J}_\star = \varnothing$.

Proof sketch. When $\mathcal{E} = \{\boldsymbol{0}\}$, the coercivity and strict/strong convexity of R_0 are vacuously true; assume $\mathcal{E} \neq \{\boldsymbol{0}\}$ henceforth. The recession function of R_0 is

$$\rho_0(\boldsymbol{V}) := \frac{1}{N} \sum\nolimits_{i=1}^{N} \max\left\{0, \max_{c:(i,c)\in\mathcal{I}_\star} \Delta_{ic}(\boldsymbol{V})\right\}.$$

It is continuous and coercive; the coercivity of R_0 follows from $R_0 \ge \rho_0$. If $\mathcal{J}_\star \neq \varnothing$, any finite $\boldsymbol{W} = \boldsymbol{W}^{\mathrm{n}} + \boldsymbol{U} + \boldsymbol{V}$ following Equation 4 has $R(\boldsymbol{W}) > R_0(\boldsymbol{V})$, while $R(\boldsymbol{V}_\star + s\widehat{\boldsymbol{U}}) \to R_0(\boldsymbol{V}_\star)$ as $s \to \infty$. If $\mathcal{J}_\star = \varnothing$, then $\mathcal{C} = \mathcal{M} = \mathcal{L}$ and $R(\boldsymbol{W}) = R_0(\boldsymbol{V})$; see Theorem D.3 for details. □

We first study whether the risk converges to its infimum and $\boldsymbol{V}_t$ converges to $\boldsymbol{V}_\star$, then ask whether the normalized margin of $\boldsymbol{U}_t$ converges to its maximum.

4 RISK CONVERGENCE

We first handle whether the risk can be minimized and whether $\boldsymbol{V}_t \to \boldsymbol{V}_\star$. We first state an assumption on step sizes, which is common in the literature (Lyu & Li, 2020; Fan et al., 2025).

Assumption 4.1. $0 < \eta_t \le \frac{1}{\beta}$ *and* $\sum_{t=0}^{\infty} \eta_t = \infty$.

This includes constant step sizes and $\eta_t \asymp t^{-a}$ for $a \in (0, 1]$. When $\mathcal{J}_\star \neq \varnothing$, define the max-margin

$$\gamma_\star := \max_{\substack{\boldsymbol{U} \in \mathcal{M} \\ \|\boldsymbol{U}\| \le 1}} \min_{(i,c) \in \mathcal{J}_\star} \{-\Delta_{ic}(\boldsymbol{U})\} > 0. \tag{6}$$

The max margin $\gamma_\star$ is strictly positive from Lemma D.2 and is maximized at $\boldsymbol{U}_\star$ from the compactness of the unit ball in $\mathcal{M}$. The proof controls both R_0, which retains the active pairs, and the residual $\tilde{R}(\boldsymbol{U}; \boldsymbol{V}) := R(\boldsymbol{W}) - R_0(\boldsymbol{V})$ arising from inactive pairs through the gradient norm.

Lemma 4.2 (Sublevel gradient bound)**.** *For every $b > R_\star$, there is $\kappa_b > 0$ such that*

$$\|\nabla R(\boldsymbol{W})\|_* \ge \kappa_b (R(\boldsymbol{W}) - R_\star) \qquad \textit{whenever } R(\boldsymbol{W}) \le b. \tag{7}$$

Proof sketch. When $\mathcal{J}_\star = \varnothing$, the claim is immediate from the strong convexity of R_0 on its sublevel set. Assume $\mathcal{J}_\star \neq \varnothing$. For $\mathcal{E} \neq \{\boldsymbol{0}\}$, we have

$$\|\nabla R(\boldsymbol{W})\|_* \ge \langle -\nabla R(\boldsymbol{W}), \boldsymbol{U}_\star \rangle \ge \gamma_\star C_1 \widetilde{R}(\boldsymbol{U}; \boldsymbol{V}),$$
$$\|\nabla R(\boldsymbol{W})\|_* \ge C_2 (R_0(\boldsymbol{V}) - R_\star) - C_3 \widetilde{R}(\boldsymbol{U}; \boldsymbol{V}),$$

where $C_1, C_2 > 0$ and $C_3 \ge 0$. The first bound follows as $-\Delta_{ic}(\boldsymbol{U}_\star) \ge \gamma_\star$ for inactive pairs. Projecting the gradient onto $\mathcal{E}$ gives the second bound. A weighted sum of the two inequalities gives Equation 7. If $\mathcal{E} = \{\boldsymbol{0}\}$, then the first bound alone suffices; see Section E.1 for details. □

We now can claim the convergence of risk and $\boldsymbol{V}_t$ with explicit rates. Write $\delta_t := R(\boldsymbol{W}_t) - R_\star$.

Theorem 4.3 (Risk and bounded-component convergence)**.** *Assume Assumption 4.1. Set $b = R(\boldsymbol{W}_0)$. Let $\lambda_b > 0$ be a strong-convexity modulus of R_0 on $\{\boldsymbol{V} \in \mathcal{E} : R_0(\boldsymbol{V}) \le b\}$, with respect to the norm restricted to $\mathcal{E}$; when $\mathcal{E} = \{\boldsymbol{0}\}$, set $\lambda_b = \beta$. If $\mathcal{J}_\star \neq \varnothing$, then*

$$\delta_t \le \left(\delta_0^{-1} + \frac{\kappa_b^2}{2} \sum\nolimits_{j=0}^{t-1} \eta_j \right)^{-1}. \tag{8}$$

If $\mathcal{J}_\star = \varnothing$, then $\delta_t \le \delta_0 \exp(-\lambda_b \sum_{j<t} \eta_j)$. In both cases, $\boldsymbol{V}_t \to \boldsymbol{V}_\star$ and

$$\|\boldsymbol{V}_t - \boldsymbol{V}_\star\|^2 \le \frac{2}{\lambda_b} \delta_t. \tag{9}$$

If $\delta_0 = 0$, the trajectory is stationary and the risk is identically constant. Theorem 4.3 implies that the risk monotonically decreases at a rate inversely proportional to the sum of step sizes. When $\mathcal{J}_\star = \varnothing$, R_0 is strongly convex on its sublevel set and thus exhibits an exponential rate. Moreover, the rate of risk directly transfers to the convergence rate of $\boldsymbol{V}_t$ to $\boldsymbol{V}_\star$.

For gradient descent on binary nonseparable logistic regression, Ji & Telgarsky (2019) obtain

$$\delta_t = O\left(\frac{1}{t} + \frac{\log^2 t}{\sum_{j<t} \eta_j} \right).$$

For $\eta_t \asymp t^{-a}$ for $a \in [0, 1)$, our rate is faster by a polylogarithmic factor; the same improvement holds for the convergence rate of $\boldsymbol{V}_t$. The improvement is in the upper bound of δ_t: Ji & Telgarsky (2019) used a comparison point of order $O(\log t)$, whereas our Lemma 4.2 gives a recurrence for δ_t.

Proof sketch. Smoothness and steepest descent update give

$$\delta_{t+1} \le \delta_t - \frac{\eta_t}{2} \|\nabla R(\boldsymbol{W}_t)\|_*^2.$$

If $\mathcal{J}_\star \neq \varnothing$, nonattainment ensures $\delta_t > 0$. Apply Lemma 4.2 and telescope $1/\delta_t$. If $\mathcal{J}_\star = \varnothing$ and $\delta_0 > 0$, the strong convexity of R_0 gives $\|\nabla R(\boldsymbol{W}_t)\|_*^2 \ge 2\lambda_b \delta_t$. Since $R|_{\mathcal{E}} = R_0$, one has $\lambda_b \le \beta$; hence the factor $1 - \lambda_b \eta_t$ is nonnegative, giving the exponential bound. Finally, strong convexity of R_0 and $R_0(\boldsymbol{V}_t) - R_\star \le \delta_t$ prove Equation 9; see Section E for details. □

5 MARGIN CONVERGENCE

We next analyze whether the normalized margin at $\boldsymbol{U}_t$ approaches its maximum as $t \to \infty$. Throughout this section, assume $\mathcal{J}_\star \neq \varnothing$; otherwise, $\mathcal{L} = \mathcal{M}$ and there are no inactive margin constraints. Recall the positive maximum margin $\gamma_\star$ from Equation 6. The max-margin set is

$$\mathcal{U} := \{\boldsymbol{U} \in \mathcal{M} : \|\boldsymbol{U}\| \leq 1,\ -\Delta_{ic}(\boldsymbol{U}) \geq \gamma_\star \text{ for every } (i,c) \in \mathcal{J}_\star\}.$$

It is nonempty, compact, and convex. We require an additional assumption on step size to prove margin convergence.

Assumption 5.1. *Assume Assumption 4.1 and* $\sum_{t=1}^{\infty} \frac{\eta_t^2}{(\sum_{j=0}^{t-1} \eta_j)^{1/2}} < \infty$.

In particular, $\eta_t \asymp t^{-a}$ satisfies Assumption 5.1 for $1/3 < a \leq 1$.

Theorem 5.2 (Margin convergence)**.** *Under Assumption 5.1,* $\|\boldsymbol{U}_t\| \to \infty$ *and*

$$0 \leq \gamma_\star - \frac{\min_{(i,c)\in\mathcal{J}_\star}\{-\Delta_{ic}(\boldsymbol{U}_t)\}}{\|\boldsymbol{U}_t\|} = O\left(\frac{1}{\log(\sum_{j=0}^{t-1}\eta_j)}\right). \tag{10}$$

For separable binary classification, Nacson et al. (2019a) showed the optimal rate $O(\frac{1}{\log t})$ for separable datasets is achieved by steepest descent with exponential loss and constant steps. Our rate matches theirs for $\eta_t \asymp t^{-a}$ for $a \in (\frac{1}{3}, 1)$ and applies to multiclass nonseparable datasets and cross-entropy loss.

Proof sketch. Since $\mathcal{J}_\star \neq \varnothing$, $\delta_t > 0$. The constants C_4, C_5, C_6, C_7 below are positive. By Theorem 4.3, $\boldsymbol{V}_t$ is bounded, thus

$$\|\boldsymbol{U}_t\| \leq \textstyle\sum_{j<t}\eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_* + C_4. \tag{11}$$

A lower bound of $\left\|\nabla R(\boldsymbol{W}_t)\right\|_*$ as a quadratic in δ_t, smoothness in R, and Assumption 5.1 give

$$-\log\delta_t \geq \gamma_\star \textstyle\sum_{j<t}\eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_* - C_5. \tag{12}$$

Next, $R(\boldsymbol{W}_t) - R_0(\boldsymbol{V}_t) \leq \delta_t$ implies, for each inactive pair,

$$e^{\Delta_{ic}(\boldsymbol{W}_t)} \leq \left(1 + \textstyle\sum_{c':(i,c')\in\mathcal{I}_\star} e^{\Delta_{ic'}(\boldsymbol{V}_t)}\right)(e^{N\delta_t} - 1).$$

Since $\boldsymbol{V}_t$ is bounded and $\delta_t \to 0$, for every inactive pair and all sufficiently large t,

$$e^{\Delta_{ic}(\boldsymbol{W}_t)} \leq C_6\delta_t, \qquad -\Delta_{ic}(\boldsymbol{U}_t) = -\Delta_{ic}(\boldsymbol{W}_t) + \Delta_{ic}(\boldsymbol{V}_t) \geq -\log\delta_t - C_7. \tag{13}$$

Combining Equations 11, 12, and 13, and taking the minimum over inactive pairs gives

$$\min_{(i,c)\in\mathcal{J}_\star}\{-\Delta_{ic}(\boldsymbol{U}_t)\} \geq \gamma_\star\textstyle\sum_{j<t}\eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_* - C_5 - C_7 \geq \gamma_\star\|\boldsymbol{U}_t\| - \gamma_\star C_4 - C_5 - C_7. \tag{14}$$

In addition, Equation 13 and the risk rate in Equation 8 yield

$$\gamma_\star\|\boldsymbol{U}_t\| \geq \min_{(i,c)\in\mathcal{J}_\star}\{-\Delta_{ic}(\boldsymbol{U}_t)\} \geq -\log\delta_t - C_7 \geq \log(\textstyle\sum_{j<t}\eta_j) - O(1). \tag{15}$$

Hence $\|\boldsymbol{U}_t\| \to \infty$. Finally, Equations 14 and 15 imply

$$0 \leq \gamma_\star - \frac{\min_{(i,c)\in\mathcal{J}_\star}\{-\Delta_{ic}(\boldsymbol{U}_t)\}}{\|\boldsymbol{U}_t\|} \leq \frac{\gamma_\star C_4 + C_5 + C_7}{\|\boldsymbol{U}_t\|} = O\left(\frac{1}{\log(\sum_{j<t}\eta_j)}\right).$$

See Lemmas F.1 to F.4 for details. □

Constant steps. Assumption 5.1 excludes constant steps, for which the margin convergence of steepest descent in separable binary classification with exponential loss is guaranteed (Nacson et al., 2019a). In contrast, there is a nonseparable binary dataset and a norm such that steepest descent with respect to the norm and tie-breaking rule of minimum Euclidean norm cannot attain max-margin for any $0 < \eta \leq 1$; see Section G for details. For entrywise ℓ_∞ norm, this tie-breaking rule gives the usual convention in signed gradient descent: it chooses zero at zero coordinates of the gradient.

Nonseparability creates opposing gradients that drive $\boldsymbol{V}_t$ towards $\boldsymbol{V}_\star$. The optimization norm can couple these gradients to movement in $\boldsymbol{U}_t$. With constant steps, the gradient on $\boldsymbol{V}_t$ can remain significant relative to the decaying gradient on $\boldsymbol{U}_t$, even though $\boldsymbol{V}_t$ converges. The updates are repeatedly redirected, preventing the normalized margin from converging to its maximum.

5.1 Strict convexity and directional convergence

Margin convergence confines all limiting directions to the max-margin directions but does not necessarily choose one. Strict convexity of the norm forces the max-margin direction to be unique; we further characterize the rate of convergence to that direction via modulus of convexity.

For the norm $\|\cdot\|$ restricted to $\mathcal{M}$, define its modulus of convexity and its generalized inverse by

$$\begin{aligned}\omega_{\mathcal{M}}(r) &:= \min_{\substack{\boldsymbol{H},\boldsymbol{H}'\in\mathcal{M}\\ \|\boldsymbol{H}\|=\|\boldsymbol{H}'\|=1}} \left(1-\left\|\frac{\boldsymbol{H}+\boldsymbol{H}'}{2}\right\| : \|\boldsymbol{H}-\boldsymbol{H}'\|\geq r\right), \qquad 0\leq r\leq 2,\\ \omega^{\dagger}_{\mathcal{M}}(s) &:= \sup\{r\in[0,2]:\omega_{\mathcal{M}}(r)\leq s\}, \qquad s\geq 0.\end{aligned} \tag{16}$$

If the norm is strictly convex, then $\omega_{\mathcal{M}}(r)>0$ for $r>0$ and $\omega^{\dagger}_{\mathcal{M}}(s)\to 0$ as $s\downarrow 0$ (Lemma F.5).

Theorem 5.3 (Directional convergence)**.** *Assume that $\|\cdot\|$ is strictly convex (restricted to $\mathcal{M}$). Then the max-margin set $\mathcal{U}=\{\boldsymbol{U}_\star\}$. Moreover, under Assumption 5.1, for all sufficiently large t,*

$$\left\|\frac{\boldsymbol{U}_t}{\|\boldsymbol{U}_t\|}-\boldsymbol{U}_\star\right\| \leq \omega^{\dagger}_{\mathcal{M}}\left(\frac{\gamma_\star-\min_{(i,c)\in\mathcal{J}_\star}\{-\Delta_{ic}(\boldsymbol{U}_t)\}/\|\boldsymbol{U}_t\|}{2\gamma_\star}\right)\longrightarrow 0. \tag{17}$$

In addition, $\left\|\frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}-\boldsymbol{U}_\star\right\|\to 0$. If $\omega_{\mathcal{M}}(r)\geq c_\omega r^q$ for some $c_\omega>0$ and $q\geq 2$, both errors have rate of $O((\log(\sum_{j<t}\eta_j))^{-1/q})$.

For entrywise matrix ℓ_p norms, $1<p<\infty$, the modulus admits a power-type lower bound with exponent $\max\{2,p\}$ (Lindenstrauss & Tzafriri, 1979). In particular, for gradient descent ($p=2$) and $\eta_j\asymp\frac{1}{\sqrt{j+1}}$, the convergence rate of $\left\|\frac{\boldsymbol{U}_t}{\|\boldsymbol{U}_t\|}-\boldsymbol{U}_\star\right\|^2$ is $O((\log t)^{-1})$, compared with $O(\log\log t/\log t)$ in Ji & Telgarsky (2019).

Proof sketch. If there are at least two max-margin directions, then their midpoint contradicts the maximality of the max-margin. By definition, $\omega_{\mathcal{M}}\left(\left\|\frac{\boldsymbol{U}_t}{\|\boldsymbol{U}_t\|}-\boldsymbol{U}_\star\right\|\right)\leq 1-\left\|\frac{\boldsymbol{U}_t/\|\boldsymbol{U}_t\|+\boldsymbol{U}_\star}{2}\right\|$. The right-hand side is bounded by Equation 10 and thus vanishes to 0; taking the generalized inverse proves directional convergence of $\boldsymbol{U}_t$. The directional convergence of $\boldsymbol{W}_t$ follows from

$$\left\|\frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}-\frac{\boldsymbol{U}_t}{\|\boldsymbol{U}_t\|}\right\|\leq\frac{2\|\boldsymbol{V}_t\|}{\|\boldsymbol{U}_t\|}\to 0.$$

See Section F.4 for details. For details when $\omega_{\mathcal{M}}$ has a power-type lower bound, see Lemma F.5. □

6 Directional convergence may not always happen

Margin convergence need not select a single point when the max-margin set $\mathcal{U}$ contains multiple points. We now study whether the normalized parameters still converge to one of them. The *accumulation set* of a sequence is the set of all limits of its convergent subsequences. We first characterize the accumulation set of normalized $\boldsymbol{U}_t$.

Proposition 6.1. *If $\mathcal{J}_\star\neq\varnothing$ and Assumption 5.1 holds, the accumulation set of $\boldsymbol{U}_t/\|\boldsymbol{U}_t\|$ is a nonempty, compact, connected subset of $\mathcal{U}$.*

Without strict convexity, the accumulation set need not be a singleton. We now show that this accumulation set can contain **more than one point**.

6.1 Counterexample inside lineality

Recall the lineality $\mathcal{L}=\mathcal{S}\cap\mathcal{Z}$ from Section 3.1: it consists of directions in the update space along which all score gaps remain unchanged, thus the loss and unnormalized margin are unchanged. We design a deterministic tie-breaking rule such that movements in $\mathcal{L}$ prevent directional convergence.

Proposition 6.2 (Nonconvergence by lineality)**.** *There exists a dataset with $\mathcal{J}_\star \neq \varnothing$, a norm, a step size schedule satisfying Assumption 5.1, and a tie-breaking rule of steepest direction such that neither $\boldsymbol{U}_t$ nor $\boldsymbol{W}_t$ under steepest descent converges in direction, although $R(\boldsymbol{W}_t) \to R_\star$ and the normalized margin of $\boldsymbol{U}_t$ converges to $\gamma_\star$.*

Proof sketch. The tie-breaking rule keeps $\Pi_{\mathcal{L}}\boldsymbol{W}_t$ in a fixed two-dimensional subspace. Represent its direction by an angle $\theta_t \in \mathbb{R}$; θ_t can have values out of $[0, 2\pi]$. The construction sets $\theta_1 = \frac{\pi}{4}$ and gives $\theta_t \to \infty$ but $\theta_{t+1} - \theta_t \to 0$. Hence, $\Pi_{\mathcal{L}}\boldsymbol{W}_t$ rotates forever in that two-dimensional subspace, and the accumulation set of the direction of $\Pi_{\mathcal{L}}\boldsymbol{W}_t$ is the entire unit circle. Moreover, the norm ratio of $\Pi_{\mathcal{L}}\boldsymbol{W}_t$ to $\boldsymbol{W}_t$ is bounded below by a positive constant, hence neither $\boldsymbol{W}_t$ nor $\boldsymbol{U}_t$ can converge in direction; see Section H for details. □

6.2 Counterexample outside lineality

The previous section indicates the lineality as a source of failure of directional convergence. But is this the only cause of such a failure? We claim that the directional convergence could fail even in the absence of lineality by giving an example with a separable binary dataset. We use $\boldsymbol{w} \in \mathbb{R}^D$, $y_i \in \{\pm 1\}$, and the score $-y_i \boldsymbol{w}^\top \boldsymbol{x}_i$.

Fix $m \geq 7$ and consider the dataset in $\mathbb{R}^3$:

$$\mathcal{D}_{\text{sep}} = \{(\boldsymbol{e}_1, +1), (2\boldsymbol{e}_1 + \boldsymbol{e}_2, +1), (m\boldsymbol{e}_1 + \boldsymbol{e}_3, +1), (-m\boldsymbol{e}_1 + \boldsymbol{e}_3, -1)\}.$$

The vector $\boldsymbol{e}_1$ separates $\mathcal{D}_{\text{sep}}$. $\boldsymbol{x}_i$'s span $\mathbb{R}^3$, thus $\mathcal{S} = \mathcal{L} = \{\boldsymbol{0}\}$ and $\mathcal{M} = \mathcal{Z} = \mathbb{R}^3$.

Theorem 6.3 (Directional nonconvergence on separable data)**.** *There is a norm $\|\cdot\|_{\mathbb{B}}$ such that under $\eta_t = \frac{\eta_0}{t+1}$ with sufficiently small η_0, the accumulation set of $\frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|_{\mathbb{B}}}$ under the steepest descent is*

$$\{(1, 1, s)^\top : |s| \leq 1/4\}.$$

*Moreover, this set is exactly the max-margin set. The risk vanishes and the normalized margin converges to max-margin. This result holds for **any** tie-breaking rule for steepest directions.*

Proof sketch. Set $b_1 = 10$, $b_j = (1 + j^2)b_{j-1}$ for $j \geq 2$. Define a vector sequence $\boldsymbol{q}_j = \left(1 - \sum_{l=j}^{\infty} 2^{-l-1} e^{-b_l},\ 1 + 2^{-j},\ \frac{(-1)^j}{4}\right)^\top$. Let $\mathbb{B} = \operatorname{cl}\operatorname{conv}\left(\frac{1}{10}\mathbb{B}_2 \cup \{\pm \boldsymbol{q}_j : j \geq 1\}\right)$, where $\mathbb{B}_2$ is the Euclidean unit ball. The Minkowski functional of $\mathbb{B}$, $\|\boldsymbol{w}\|_{\mathbb{B}} = \inf\{r > 0 : \boldsymbol{w} \in r\mathbb{B}\}$, is a norm. We have that $\|\boldsymbol{q}_j\|_{\mathbb{B}} = 1$, the max-margin is 1, and the max-margin set is $\{(1, 1, s)^\top : |s| \leq 1/4\}$. Moreover, Theorems 4.3 and 5.2 yield that $\|\boldsymbol{w}_t\|_{\mathbb{B}} \to \infty$ and $\sum_{j<t} \eta_j \|\nabla R(\boldsymbol{w}_j)\|_{\mathbb{B},*} \to \infty$.

Let $\boldsymbol{w}_t = (x_t, y_t, z_t)$. Every unit update direction lies in $\operatorname{cl}\operatorname{conv}\{\boldsymbol{q}_j : j \geq 1\}$. The bounds on x_t and y_t imply that $x_t + y_t$ increases to infinity with increments vanishing to 0. For $j \geq 2$, define $t_j^- := \min\{t : x_t + y_t \geq b_{j-1} + \log 8\}$ and $t_j^+ := \min\{t : x_t + y_t \geq b_j - \log 8\}$. For sufficiently large j, $t_j^- < t_j^+$ and $\boldsymbol{s}_t = -\boldsymbol{q}_j$ (Equation 2) throughout $t_j^- \leq t < t_j^+$. The steepest directions outside these intervals need not be unique and we allow any tie-breaking rule. From $b_j - b_{j-1} = j^2 b_{j-1}$, the accumulated update between t_j^- and t_j^+ eventually dominates all preceding updates:

$$\frac{\sum_{t=0}^{t_j^- - 1} \eta_t \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{B},*}}{\sum_{t=t_j^-}^{t_j^+ - 1} \eta_t \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{B},*}} = O\left(j^{-2}\right) \to 0 \qquad (\text{as } j \to \infty).$$

Since $\boldsymbol{w}_{t_j^+} = \boldsymbol{w}_{t_j^-} + \boldsymbol{q}_j \sum_{t=t_j^-}^{t_j^+ - 1} \eta_t \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{B},*}$ and $\|\boldsymbol{q}_j\|_{\mathbb{B}} = 1$,

$$\left\| \frac{\boldsymbol{w}_{t_j^+}}{\sum_{t=0}^{t_j^+ - 1} \eta_t \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{B},*}} - \boldsymbol{q}_j \right\|_{\mathbb{B}} \to 0, \qquad \frac{\left\|\boldsymbol{w}_{t_j^+}\right\|_{\mathbb{B}}}{\sum_{t=0}^{t_j^+ - 1} \eta_t \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{B},*}} \to 1 \qquad (\text{as } j \to \infty).$$

But $\boldsymbol{q}_{2j} \to (1, 1, 1/4)$ while $\boldsymbol{q}_{2j+1} \to (1, 1, -1/4)$. The max-margin set has these limit points as its endpoints, and by Proposition 6.1, agrees with the accumulation set; see Section I for details. □

The construction fixes a sequence of unit directions with two distinct limits. The directional convergence fails if the updates align for a sufficiently long time to dominate all preceding updates.

Extension to nonseparable dataset The example in Theorem 6.3 naturally extends to nonseparable classification. Embed the dataset $\mathcal{D}_{\mathrm{sep}}$ in $\mathbb{R}^4$ by appending a zero and add $(\boldsymbol{e}_4, +1), (\boldsymbol{e}_4, -1)$. Define

$$\left\|(x, y, z, v)^\top\right\|_{\mathrm{ns}} := \max\{\left\|(x, y, z)^\top\right\|_{\mathbb{B}}, |v|\}.$$

The two new examples force every recession direction to have fourth coordinate zero. Thus $\mathcal{L}_{\mathrm{ns}} = \{\mathbf{0}\}$, $\mathcal{M}_{\mathrm{ns}} = \mathbb{R}^3 \times \{\mathbf{0}\}$, and $\mathcal{E}_{\mathrm{ns}} = \mathrm{span}\{\boldsymbol{e}_4\}$. See Section J for the proof of Corollary 6.4.

Corollary 6.4. *Under the minimum Euclidean norm tie-breaking rule and $\eta_t = \frac{\eta_0}{t+1}$ with sufficiently small η_0, the accumulation set of $\frac{\boldsymbol{w}_t^{\mathrm{ns}}}{\|\boldsymbol{w}_t^{\mathrm{ns}}\|_{\mathrm{ns}}}$ under the steepest descent is $\{(1, 1, s, 0)^\top : |s| \le 1/4\}$.*

7 Related work

Separable linear classification. For separable binary linear classification, Soudry et al. (2018) prove ℓ_2 max-margin directional convergence of small fixed-step GD for smooth losses with an exponential tail property, including logistic loss. Nacson et al. (2019b) extend the results for small fixed-step SGD. For separable exponential loss, Gunasekar et al. (2018) establish optimal normalized margin under steepest descent with bounded step sizes. Nacson et al. (2019a) further prove that the steepest descent exhibits margin-convergence rates of $O(1/\log t)$ on exponential loss.

For multiclass separable classification, Soudry et al. (2018) give the Frobenius max-margin direction of gradient descent for cross-entropy loss with small constant steps. Ravi et al. (2024) extend the analysis to a broader family of convex smooth multiclass losses satisfying their exponential-tail condition. Fan et al. (2025) analyze normalized steepest descent and its momentum variant with respect to entrywise and Schatten p-norms with cross-entropy loss.

The implicit bias literature further extends to adaptive optimizers. AdaGrad induces a trajectory-dependent weighted-Euclidean max-margin geometry (Qian & Qian, 2019), whereas full-batch Adam with $\varepsilon = 0$ in the denominator yields the normalized ℓ_∞ max-margin for separable binary linear classification with exponential and logistic loss under vanishing step sizes (Zhang et al., 2024).

Nonseparable classification. The closest precedent is the binary logistic analysis of Ji & Telgarsky (2019). Their analysis combines a logarithmically growing comparator with a telescoping argument, then uses the Fenchel–Young inequality to control the parameter direction. Ji et al. (2020) also relate GD and regularization-path directions on general binary datasets, showing that the two paths have the same directional limit whenever either limit exists and the risk infimum is unattained.

Directional convergence. Beyond linear models, directional convergence and margin-based characterizations have been studied for homogeneous models under gradient flow and steepest flow (Lyu & Li, 2020; Ji & Telgarsky, 2020; Tsilivis et al., 2025). In particular, Ji & Telgarsky (2020) show that the negative gradient can asymptotically align with the parameter and converge to the same direction. Appendix K complements these results with an example in which the parameter converges in direction although the steepest directions do not.

8 Discussion

A split of example–class pairs, rather than a split by examples, identifies the risk that remains at infinity. From this observation, we identified a norm-dependent decomposition of the parameter space using the recession cone: a stationary component, a bounded component, and, when the infimum risk is unattained, an escaping component. We then obtain convergence rates of risk, bounded component, and normalized margin. Strict convexity gives a unique max-margin direction and a convergence rate through the modulus of convexity. Without strict convexity, the max-margin directions need not be unique and directional convergence can fail through rotations in lineality or, even without lineality, as the trajectory is successively dominated by updates toward different max-margin directions. The construction leaves open which additional properties of familiar non-strictly convex norms force directional convergence when their max-margin sets are not singletons. Extending the nonseparable risk and margin analysis to other optimizers is another direction.

AI USE STATEMENT

In this work, we used generative AI tools for refining mathematical formulations and claims, suggesting proof strategies, assisting in the writing of proofs, and verifying the proofs. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, interpreting results, and assisting with translation. Generating synthetic datasets, cleaning and reformatting datasets, supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for transcribing recordings of research material, drafting parts of a research paper, and editing a research paper to improve readability. We have reviewed all AI-assisted work. All AI-generated content in the main text and appendix is verified and revised by authors. Every mathematical claim and its proof that AI has improved and polished were checked line-by-line. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Dimitri Bertsekas. *Convex optimization theory*. Athena Scientific, 2009.

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

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on learning theory*, pp. 1772–1798. PMLR, 2019.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pp. 2109–2136. PMLR, 2020.

Joram Lindenstrauss and Lior Tzafriri. *Classical Banach spaces II: Function spaces*. Springer, 1979.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Kazuo Murota. On steepest descent algorithms for discrete convex functions. *SIAM Journal on Optimization*, 14(3):699–707, 2004.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428. PMLR, 2019a.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3051–3059. PMLR, 2019b.

Qian Qian and Xiaoyuan Qian. The implicit bias of adagrad on separable data. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Walter Rudin. *Functional Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, 2nd edition, 1991.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171 – 176, 1958.

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.

Nikolaos Tsilivis, Gal Vardi, and Julia Kempe. Flavors of margin: Implicit bias of steepest descent in homogeneous neural networks. In *International Conference on Learning Representations*, 2025.

Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data. *Advances in neural information processing systems*, 37:23988–24021, 2024.

CONTENTS

1 Introduction **1**

2 Setup **2**

3 Decomposition of parameter space **3**

3.1 Subspace decomposition 3

3.2 Active pairs and reduced risk 4

4 Risk convergence **5**

5 Margin convergence **6**

5.1 Strict convexity and directional convergence 7

6 Directional convergence may not always happen **7**

6.1 Counterexample inside lineality 7

6.2 Counterexample outside lineality 8

7 Related work **9**

8 Discussion **9**

A Preliminaries **14**

B Decomposition of the parameter space **15**

C Failure of an example-wise split in multiclass classification **17**

D Reduced risk **17**

E Risk convergence **19**

E.1 Nonempty $\mathcal{J}_\star$ 20

E.2 Empty $\mathcal{J}_\star$ 22

F Margin and directional convergence **22**

F.1 Gradient and excess risk 23

F.2 Summability of error terms 23

F.3 Inactive score gaps and excess risk 25

F.4 Directional convergence rate from modulus of convexity 26

G Counterexample with constant steps **28**

G.1 Setup 28

G.2 Steepest direction and update control 29

H Counterexample inside lineality **33**

H.1 Setup 34

H.2 Radial growth and rotation 36

I Counterexample outside lineality **37**

I.1 Setup 37

I.2 Growth of parameter 39

I.3 Gradient ratios 40

I.4 Dominant updates and directional nonconvergence 41

J Nonseparable counterexample outside lineality **41**

J.1 Setup 42

J.2 Trajectory embedding 42

K Directional convergence without convergence of steepest directions **43**

L Norm-dependent max-margin duality **45**

A PRELIMINARIES

For $(i, c) \in \mathcal{I}$, let $\boldsymbol{M}_{(i,c)} \in \mathbb{R}^{K\times D}$ be defined by

$$(\boldsymbol{M}_{(i,c)})_c = \boldsymbol{x}_i^\top, \qquad (\boldsymbol{M}_{(i,c)})_{y_i} = -\boldsymbol{x}_i^\top,$$

and all the other rows as zero. Then $\Delta_{ic}(\boldsymbol{W}) = (\boldsymbol{w}_c - \boldsymbol{w}_{y_i})^\top \boldsymbol{x}_i = \langle \boldsymbol{M}_{(i,c)}, \boldsymbol{W}\rangle$.

Lemma A.1. *For every $c \in [K]$, with $\Delta_{iy_i}(\boldsymbol{W}) = 0$, let*

$$p_{ic}(\boldsymbol{W}) := \frac{e^{\Delta_{ic}(\boldsymbol{W})}}{1 + \sum_{c'\neq y_i} e^{\Delta_{ic'}(\boldsymbol{W})}}.$$

Then

$$\nabla R(\boldsymbol{W}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{c\neq y_i} p_{ic}(\boldsymbol{W})\boldsymbol{M}_{(i,c)}. \tag{18}$$

The sum of the rows of $\nabla R(\boldsymbol{W})$ is zero. Moreover, $\nabla R(\boldsymbol{W}) \in \mathcal{S}^\perp$.

Proof. Equation 18 is immediate from

$$\nabla R(\boldsymbol{W}) = \frac{1}{N}\sum_i (\boldsymbol{p}_i(\boldsymbol{W}) - \boldsymbol{e}_{y_i})\boldsymbol{x}_i^\top = \frac{1}{N}\sum_{i=1}^{N}\sum_{c\neq y_i} p_{ic}(\boldsymbol{W})\boldsymbol{M}_{(i,c)},$$

where $\boldsymbol{p}_i(\boldsymbol{W})$ is the column vector obtained by stacking $p_{ic}(\boldsymbol{W})$ for all c. Each matrix $\boldsymbol{M}_{(i,c)}$ has row sum zero by its construction, thus the sum of the rows of ∇R is zero.

Recall that $\mathcal{S} = \{\boldsymbol{W} : \Delta_{ic}(\boldsymbol{W}) = 0, \forall i,\ \forall c \neq y_i\}$. If $\boldsymbol{S} \in \mathcal{S}$, then $R(\boldsymbol{W} + t\boldsymbol{S}) = R(\boldsymbol{W})$ for every t, and therefore

$$\langle \nabla R(\boldsymbol{W}), \boldsymbol{S}\rangle = \frac{d}{dt} R(\boldsymbol{W} + t\boldsymbol{S})\Big|_{t=0} = 0.$$

□

Lemma A.2 (Convexity and smoothness)**.** *The risk R is convex and globally β-smooth with respect to $\|\cdot\|$ for any $\beta > 0$ satisfying*

$$\beta \geq \frac{1}{4N}\sum_{i=1}^{N}\left(\max_{1\leq c,c'\leq K}\sup_{\|\boldsymbol{H}\|\leq 1} |(\boldsymbol{h}_c - \boldsymbol{h}_{c'})^\top \boldsymbol{x}_i|\right)^2. \tag{19}$$

That is,

$$R(\boldsymbol{W} + \boldsymbol{H}) \leq R(\boldsymbol{W}) + \langle \nabla R(\boldsymbol{W}), \boldsymbol{H}\rangle + \frac{\beta}{2}\|\boldsymbol{H}\|^2.$$

Consequently,

$$\|\nabla R(\boldsymbol{W})\|_*^2 \leq 2\beta\big(R(\boldsymbol{W}) - R_\star\big). \tag{20}$$

Proof. Recall the risk

$$R(\boldsymbol{W}) = \frac{1}{N}\sum_i R_i(\boldsymbol{W}) := \frac{1}{N}\sum_i \log\left(1 + \sum_{c\neq y_i} e^{\Delta_{ic}(\boldsymbol{W})}\right).$$

Each summand is a log-sum-exp of linear functions Δ_{ic}, hence it is convex.

For $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K]^\top$, we have that

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} R_i(\boldsymbol{W} + t\boldsymbol{H})\Big|_{t=0} = \sum_c p_{ic}(\boldsymbol{W})(\boldsymbol{h}_c^\top \boldsymbol{x}_i)^2 - \left(\sum_c p_{ic}(\boldsymbol{W})(\boldsymbol{h}_c^\top \boldsymbol{x}_i)\right)^2,$$

upper bounded by $\frac{1}{4}\left(\max_c \boldsymbol{h}_c^\top \boldsymbol{x}_i - \min_c \boldsymbol{h}_c^\top \boldsymbol{x}_i\right)^2$. Hence, $\left\langle \boldsymbol{H}, \nabla^2 R(\boldsymbol{W})\boldsymbol{H}\right\rangle \le \beta \|\boldsymbol{H}\|^2$, which implies the β-smoothness of R. If $\nabla R(\boldsymbol{W}) \neq 0$, choose $\boldsymbol{S} \in \mathcal{F}(\nabla R(\boldsymbol{W}))$. Substituting $\boldsymbol{H} = -\beta^{-1}\|\nabla R(\boldsymbol{W})\|_* \boldsymbol{S}$ into the smoothness bound gives

$$R_\star \le R\left(\boldsymbol{W} - \frac{\|\nabla R(\boldsymbol{W})\|_*}{\beta}\boldsymbol{S}\right) \le R(\boldsymbol{W}) - \frac{\|\nabla R(\boldsymbol{W})\|_*^2}{2\beta}.$$

This proves Equation 20; if the gradient vanishes, the inequality holds directly.

□

B Decomposition of the parameter space

Proposition B.1. *$\mathbb{R}^{K\times D} = \mathcal{S} + \mathcal{Z}$. Consequently, if $\mathcal{N}$ is the (Frobenius-)orthogonal complement of $\mathcal{L} = \mathcal{S} \cap \mathcal{Z}$ in $\mathcal{S}$, then*

$$\mathbb{R}^{K\times D} = \mathcal{N} \oplus \mathcal{Z}.$$

Proof. Suppose not. Then there exists $\boldsymbol{G} \neq 0$ such that $\langle \boldsymbol{G}, \boldsymbol{W}\rangle = 0$ for all $\boldsymbol{W} \in \mathcal{S} + \mathcal{Z}$. In particular, $\boldsymbol{G} \in \mathcal{S}^\perp \setminus \{\boldsymbol{0}\}$, so $\mathcal{F}_{\|\cdot\|}(\boldsymbol{G}) \subseteq \mathcal{Z}$. Thus, $\langle \boldsymbol{G}, \boldsymbol{U}\rangle = 0$ for every $\boldsymbol{U} \in \mathcal{F}_{\|\cdot\|}(\boldsymbol{G})$, whereas the definition of the dual norm gives $\langle \boldsymbol{G}, \boldsymbol{U}\rangle = \|\boldsymbol{G}\|_* > 0$ for every $\boldsymbol{U} \in \mathcal{F}_{\|\cdot\|}(\boldsymbol{G})$, a contradiction.

Since $\mathcal{S} = \mathcal{N} \oplus \mathcal{L}$ and $\mathcal{L} \subseteq \mathcal{Z}$, the sum $\mathcal{N} + \mathcal{Z}$ is all of $\mathbb{R}^{K\times D}$. If $\boldsymbol{X} \in \mathcal{N} \cap \mathcal{Z}$, then $\boldsymbol{X} \in \mathcal{S} \cap \mathcal{Z} = \mathcal{L}$, thus $\boldsymbol{X} = 0$. □

Then, Lemma A.1 gives $\nabla R(\boldsymbol{W}_t) \in \mathcal{S}^\perp$, thus every SD step of Equation 2 lies in $\mathcal{Z}$. Hence, the $\mathcal{N}$-component of the trajectory is constant.

Corollary B.2 (Frobenius specialization)**.** *For the Frobenius norm, $\mathcal{N}_{\|\cdot\|_F} = \mathcal{S}$ and $\mathcal{Z}_{\|\cdot\|_F} = \mathcal{S}^\perp$.*

Proof. For every nonzero $\boldsymbol{G} \in \mathbb{R}^{K\times D}$,

$$\mathcal{F}_{\|\cdot\|_F}(\boldsymbol{G}) = \left\{\frac{\boldsymbol{G}}{\|\boldsymbol{G}\|_{\mathrm{F}}}\right\}. \tag{21}$$

Therefore,

$$\mathcal{Z}_{\|\cdot\|_F} = \operatorname{span}\left\{\frac{\boldsymbol{G}}{\|\boldsymbol{G}\|_{\mathrm{F}}} : \boldsymbol{G} \in \mathcal{S}^\perp \setminus \{\boldsymbol{0}\}\right\} = \mathcal{S}^\perp, \tag{22}$$

thus $\mathcal{N}_{\|\cdot\|_F} = \mathcal{S}$. □

Proposition B.3 (Recession cone and decomposition)**.** *The cone $\mathcal{C}$ is closed, convex, and polyhedral. Its lineality is*

$$\mathcal{C} \cap (-\mathcal{C}) = \mathcal{L}.$$

With $\mathcal{M} = \operatorname{span}(\mathcal{C})$ and $\mathcal{E} = \mathcal{Z} \cap \mathcal{M}^\perp$,

$$\mathbb{R}^{K\times D} = \mathcal{N} \oplus \mathcal{M} \oplus \mathcal{E}.$$

Proof. The cone $\mathcal{C}$ is polyhedral as it is an intersection of finitely many closed halfspaces in $\mathcal{Z}$; closedness and convexity follow. $\boldsymbol{U}$ belongs to both $\mathcal{C}$ and $-\mathcal{C}$ exactly when all its score gaps $\Delta_{ic}(\boldsymbol{U})$ vanish, which is equivalent to membership in $\mathcal{Z} \cap \mathcal{S} = \mathcal{L}$. The decomposition of the parameter space follows from Proposition B.1. □

Lemma B.4. *Assume $\boldsymbol{x}_i \neq 0$ for every i. Then the following are equivalent:*

(i) $\mathcal{E} = \{\boldsymbol{0}\}$.

(ii) $\mathcal{I}_\star = \varnothing$.

(iii) the training dataset is multiclass separable, i.e., there exists $\widehat{\boldsymbol{U}} \in \mathcal{Z}$ such that

$$\Delta_{ic}(\widehat{\boldsymbol{U}}) < 0 \qquad \text{for all } i \text{ and } c \neq y_i.$$

Proof. Since $\mathcal{Z} = \mathcal{M} \oplus \mathcal{E}$,

$$\mathcal{E} = \{\mathbf{0}\} \iff \mathcal{M} = \mathcal{Z}.$$

Assume $\mathcal{M} = \mathcal{Z}$ and let $(i, c) \in \mathcal{I}_\star$. By linearity of Δ_{ic},

$$\Delta_{ic}(\boldsymbol{U}) = 0 \qquad \forall \boldsymbol{U} \in \operatorname{span}(\mathcal{C}) = \mathcal{M} = \mathcal{Z}.$$

Every score gap vanishes on $\mathcal{N} \subseteq \mathcal{S}$, hence

$$\Delta_{ic}(\boldsymbol{N}) = 0 \qquad \forall \boldsymbol{N} \in \mathcal{N}.$$

Hence, $\Delta_{ic}(\boldsymbol{W}) = 0$ for all $\boldsymbol{W}$. Choose $\boldsymbol{W}$ with $\boldsymbol{w}_c = \boldsymbol{x}_i$, $\boldsymbol{w}_{y_i} = 0$, and all other rows zero. Then $\Delta_{ic}(\boldsymbol{W}) = \|\boldsymbol{x}_i\|_2^2 > 0$, a contradiction. Therefore

$$\mathcal{M} = \mathcal{Z} \implies \mathcal{I}_\star = \varnothing.$$

Assume $\mathcal{I}_\star = \varnothing$. For each pair (i, c) choose $\boldsymbol{U}_{ic} \in \mathcal{C}$ such that $\Delta_{ic}(\boldsymbol{U}_{ic}) < 0$. Set

$$\widehat{\boldsymbol{U}} := \sum_{(i,c)} \boldsymbol{U}_{ic} \in \mathcal{C}.$$

It is immediate that $\Delta_{ic}(\widehat{\boldsymbol{U}}) < 0$ for all (i, c).

Conversely, suppose there exists $\widehat{\boldsymbol{U}} \in \mathcal{Z}$ with $\Delta_{ic}(\widehat{\boldsymbol{U}}) < 0$ for all (i, c). Then $\widehat{\boldsymbol{U}} \in \mathcal{C}$, and no pair can satisfy $\Delta_{ic}(\boldsymbol{U}) = 0$ for every $\boldsymbol{U} \in \mathcal{C}$. Thus $\mathcal{I}_\star = \varnothing$. Moreover, for any $\boldsymbol{H} \in \mathcal{Z}$, define

$$m(\boldsymbol{H}) := \max_{(i,c)} |\Delta_{ic}(\boldsymbol{H})|.$$

If $m(\boldsymbol{H}) = 0$, then $\boldsymbol{H} \in \mathcal{C} \subseteq \mathcal{M}$. If $m(\boldsymbol{H}) > 0$, choose

$$0 < \varepsilon < \frac{\min_{(i,c)}\{-\Delta_{ic}(\widehat{\boldsymbol{U}})\}}{m(\boldsymbol{H})}.$$

Then for every pair (i, c),

$$\Delta_{ic}(\widehat{\boldsymbol{U}} \pm \varepsilon \boldsymbol{H}) \leq -\min_{(i,c)}\{-\Delta_{ic}(\widehat{\boldsymbol{U}})\} + \varepsilon m(\boldsymbol{H}) < 0,$$

thus $\widehat{\boldsymbol{U}} \pm \varepsilon \boldsymbol{H} \in \mathcal{C}$. Therefore

$$\boldsymbol{H} = \frac{(\widehat{\boldsymbol{U}} + \varepsilon \boldsymbol{H}) - (\widehat{\boldsymbol{U}} - \varepsilon \boldsymbol{H})}{2\varepsilon} \in \operatorname{span}(\mathcal{C}) = \mathcal{M}.$$

Since $\boldsymbol{H} \in \mathcal{Z}$ was arbitrary, $\mathcal{Z} \subseteq \mathcal{M}$, and thus $\mathcal{M} = \mathcal{Z}$ and hence $\mathcal{E} = \{\mathbf{0}\}$. □

The assumption of $\boldsymbol{x}_i$ being all nonzero is unnecessary for the equivalence between (ii) and (iii).

C Failure of an example-wise split in multiclass classification

Consider $K = 3$-class classification over $D = 4$-dimensional space. The model is $f(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x}$ with $\boldsymbol{W} \in \mathbb{R}^{3\times 4}$. Take the dataset as $\mathcal{D} = \{(\boldsymbol{e}_1, 1), (2\boldsymbol{e}_1, 2), (\boldsymbol{e}_2, 3)\}$. The risk satisfies

$$3R(\boldsymbol{W}) = -\boldsymbol{w}_1^\top \boldsymbol{e}_1 + \log \sum_c e^{\boldsymbol{w}_c^\top \boldsymbol{e}_1} - 2\boldsymbol{w}_2^\top \boldsymbol{e}_1 + \log \sum_c e^{2\boldsymbol{w}_c^\top \boldsymbol{e}_1} - \boldsymbol{w}_3^\top \boldsymbol{e}_2 + \log \sum_c e^{\boldsymbol{w}_c^\top \boldsymbol{e}_2}$$

Suppose the examples are partitioned into a separable subset A_c and a subset A_S whose risk has a finite minimizer. If A_c contains both $(\boldsymbol{e}_1, 1)$ and $(2\boldsymbol{e}_1, 2)$, then by the separability, we must have

$$\begin{cases} (\boldsymbol{e}_1 - \boldsymbol{e}_2)^\top \boldsymbol{W}\boldsymbol{e}_1 > 0 \\ (\boldsymbol{e}_2 - \boldsymbol{e}_1)^\top \boldsymbol{W} 2\boldsymbol{e}_1 > 0, \end{cases}$$

a contradiction. Hence, A_S should contain at least one of the two points. If A_S contains both, decreasing only $\boldsymbol{w}_3^\top \boldsymbol{e}_1$ strictly decreases the loss from the first two examples and leaves the loss from the third example unchanged. Thus, the risk corresponding to A_S has no finite minimizer, regardless of which subset the third example belongs to.

If A_S contains exactly one of the first two samples, it is strictly separable even if the third sample belongs to A_S. When the first sample is present, take $\boldsymbol{w}_1 = \boldsymbol{e}_1$, $\boldsymbol{w}_3 = \boldsymbol{e}_2$, and $\boldsymbol{w}_2 = 0$; when the second is present, take $\boldsymbol{w}_2 = \boldsymbol{e}_1$, $\boldsymbol{w}_3 = \boldsymbol{e}_2$, and $\boldsymbol{w}_1 = 0$. Every pairwise score gap Δ_{ic} on the respective subset is strictly negative. Scaling this matrix makes its subset risk tend to zero, whereas the risk is positive at every finite parameter. Thus, no proposed A_S has a finite risk minimizer.

This example shows that the decomposition of Ji & Telgarsky (2019) does not extend to multiclass classification through an example-wise partition alone. The first two examples retain a positive limiting loss because of their conflicting labels, even though some of their class comparisons can be separated: class 1 and 3 for the first example, and class 2 and 3 for the second example.

D Reduced risk

Define the *recession function*

$$\rho(\boldsymbol{D}) := \frac{1}{N} \sum_{i=1}^{N} \max \left\{ 0, \max_{c \neq y_i} \Delta_{ic}(\boldsymbol{D}) \right\}.$$

Lemma D.1 (Recession function)**.** *For all $\boldsymbol{W}, \boldsymbol{D}$,*

$$\lim_{r\to\infty} \frac{R(\boldsymbol{W} + r\boldsymbol{D}) - R(\boldsymbol{W})}{r} = \rho(\boldsymbol{D}).$$

Consequently,

$$\mathcal{C} = \{\boldsymbol{D} \in \mathcal{Z} : \rho(\boldsymbol{D}) = 0\}.$$

For $\boldsymbol{D} \in \mathcal{Z}$, $\boldsymbol{D} \in \mathcal{C}$ if and only if $r \mapsto R(\boldsymbol{W} + r\boldsymbol{D})$ is nonincreasing for every $\boldsymbol{W}$ and $r \geq 0$; $\boldsymbol{D} \in \mathcal{L}$ if and only if $R(\boldsymbol{W} + r\boldsymbol{D}) = R(\boldsymbol{W})$ for all $\boldsymbol{W}, r$.

Proof. For each i,

$$\lim_{r\to\infty} \frac{1}{r} \log \left(1 + \sum_{c \neq y_i} \exp(\Delta_{ic}(\boldsymbol{W}) + r\Delta_{ic}(\boldsymbol{D})) \right) = \max \left\{ 0, \max_{c \neq y_i} \Delta_{ic}(\boldsymbol{D}) \right\}.$$

If $\boldsymbol{D} \in \mathcal{C}$, every exponential term is nonincreasing along the ray. Conversely, if $\boldsymbol{D} \notin \mathcal{C}$ then $\rho(\boldsymbol{D}) > 0$; thus the risk cannot be monotonically nonincreasing along $\boldsymbol{D}$. $R(\boldsymbol{W} + r\boldsymbol{D}) = R(\boldsymbol{W})$ for all r implies that both $\boldsymbol{D}$ and $-\boldsymbol{D}$ belong to $\mathcal{C}$, or $\boldsymbol{D} \in \mathcal{L}$; the reverse implication follows because every score gap then vanishes at $\boldsymbol{D}$. □

Lemma D.2 (A direction strictly decreasing every inactive pair)**.** *If $\mathcal{J}_\star \neq \varnothing$, there exists $\widehat{\boldsymbol{U}} \in \mathcal{C}$ such that*

$$\Delta_{ic}(\widehat{\boldsymbol{U}}) = 0 \quad ((i,c) \in \mathcal{I}_\star), \qquad \Delta_{ic}(\widehat{\boldsymbol{U}}) < 0 \quad ((i,c) \in \mathcal{J}_\star).$$

Proof. For every $(i,c) \in \mathcal{J}_\star$, choose $\boldsymbol{U}_{ic} \in \mathcal{C}$ with $\Delta_{ic}(\boldsymbol{U}_{ic}) < 0$. The sum of these matrices $\widehat{\boldsymbol{U}} = \sum_{(i,c)\in\mathcal{J}_\star} \boldsymbol{U}_{ic}$ belongs to $\mathcal{C}$. Then $\Delta_{ic}(\widehat{\boldsymbol{U}})$ for $(i,c) \in \mathcal{I}_\star$ vanishes on $\mathcal{C}$, whereas that for $(i,c) \in \mathcal{J}_\star$ is strictly negative. □

Recall

$$R_0(\boldsymbol{V}) = \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + \sum_{c:(i,c)\in\mathcal{I}_\star} e^{\Delta_{ic}(\boldsymbol{V})}\right), \qquad \boldsymbol{V} \in \mathcal{E}.$$

Theorem D.3 (Reduced risk)**.** *The reduced risk R_0 is coercive and strictly convex on $\mathcal{E}$, and strongly convex on every compact convex subset of $\mathcal{E}$. It has a unique minimizer $\boldsymbol{V}_\star$, with*

$$\inf_{\boldsymbol{W}} R(\boldsymbol{W}) = R_0(\boldsymbol{V}_\star).$$

The infimum of the risk R is attained at a finite parameter if and only if $\mathcal{J}_\star = \varnothing$.

Proof of Theorem 3.1. When $\mathcal{E} = \{\mathbf{0}\}$, $\boldsymbol{V}_\star = \mathbf{0}$ is the unique reduced-risk minimizer, and coercivity and the curvature claims are vacuous. Thus, assume $\mathcal{E} \neq \{\mathbf{0}\}$ from now on; the argument regarding infimum attainment applies to both cases.

For the reduced risk, its recession function on $\mathcal{E}$ is

$$\rho_0(\boldsymbol{V}) := \frac{1}{N}\sum_i \max\left\{0, \max_{c:(i,c)\in\mathcal{I}_\star} \Delta_{ic}(\boldsymbol{V})\right\}.$$

Suppose $\rho_0(\boldsymbol{V}) = 0$ for $\boldsymbol{V} \in \mathcal{E}$. If $\mathcal{J}_\star = \varnothing$, all score gaps are already nonpositive, thus $\boldsymbol{V} \in \mathcal{C} = \mathcal{L} = \mathcal{M}$ and $\boldsymbol{V} = \mathbf{0}$. If $\mathcal{J}_\star \neq \varnothing$, $\boldsymbol{V} + s\widehat{\boldsymbol{U}} \in \mathcal{C} \subseteq \mathcal{M}$ for sufficiently large s (Lemma D.2), implying $\boldsymbol{V} \in \mathcal{M} \cap \mathcal{E} = \{\mathbf{0}\}$. The function ρ_0 is continuous and strictly positive on the compact Frobenius unit sphere in $\mathcal{E}$. Hence its minimum on that sphere is positive. Positive homogeneity and $R_0 \geq \rho_0$ imply that $R_0(\boldsymbol{V}) \to \infty$ as $\|\boldsymbol{V}\|_{\mathrm{F}} \to \infty$, proving coercivity.

For each i, let

$$\mathcal{I}_i^\star := \{c \neq y_i : (i,c) \in \mathcal{I}_\star\}, \qquad q_{ic}(\boldsymbol{V}) := \frac{e^{\Delta_{ic}(\boldsymbol{V})}}{1 + \sum_{c\in\mathcal{I}_i^\star} e^{\Delta_{ic}(\boldsymbol{V})}} > 0.$$

Write $R_{0,i}(\boldsymbol{V}) := \log(1 + \sum_{c\in\mathcal{I}_i^\star} e^{\Delta_{ic}(\boldsymbol{V})})$. Then

$$\left\langle \boldsymbol{H}, \nabla^2 R_{0,i}(\boldsymbol{V})\boldsymbol{H}\right\rangle = \sum_{c\in\mathcal{I}_i^\star} q_{ic}(\boldsymbol{V})\Delta_{ic}(\boldsymbol{H})^2 - \left(\sum_{c\in\mathcal{I}_i^\star} q_{ic}(\boldsymbol{V})\Delta_{ic}(\boldsymbol{H})\right)^2 \geq 0,$$

and thus

$$\left\langle \boldsymbol{H}, \nabla^2 R_0(\boldsymbol{V})\boldsymbol{H}\right\rangle = \frac{1}{N}\sum_{i=1}^{N} \left\langle \boldsymbol{H}, \nabla^2 R_{0,i}(\boldsymbol{V})\boldsymbol{H}\right\rangle \geq 0.$$

Note that $\left\langle \boldsymbol{H}, \nabla^2 R_{0,i}(\boldsymbol{V})\boldsymbol{H}\right\rangle$ vanishes exactly when $\Delta_{ic}(\boldsymbol{H}) = 0$ for all $c \in \mathcal{I}_i^\star$. Hence,

$$\left\langle \boldsymbol{H}, \nabla^2 R_0(\boldsymbol{V})\boldsymbol{H}\right\rangle = 0 \implies \Delta_{ic}(\boldsymbol{H}) = 0 \quad \forall (i,c) \in \mathcal{I}_\star \implies \rho_0(\boldsymbol{H}) = 0.$$

Consequently,

$$\boldsymbol{H} \in \mathcal{E} \setminus \{\mathbf{0}\} \implies \left\langle \boldsymbol{H}, \nabla^2 R_0(\boldsymbol{V})\boldsymbol{H}\right\rangle > 0 \qquad \forall \boldsymbol{V} \in \mathcal{E}.$$

Thus R_0 is strictly convex on $\mathcal{E}$. More quantitatively, for a nonempty compact $K \subset \mathcal{E}$, define

$$\lambda_F(K) := \min_{\substack{\boldsymbol{V}\in\mathrm{conv}(K)\\ \boldsymbol{H}\in\mathcal{E},\ \|\boldsymbol{H}\|_{\mathrm{F}}=1}} \left\langle \boldsymbol{H}, \nabla^2 R_0(\boldsymbol{V})\boldsymbol{H}\right\rangle > 0.$$

For $\boldsymbol{V}_0, \boldsymbol{V}_1 \in K$, set $\boldsymbol{H} := \boldsymbol{V}_1 - \boldsymbol{V}_0$. Taylor's formula with integral remainder yields

$$R_0(\boldsymbol{V}_1) = R_0(\boldsymbol{V}_0) + \langle \nabla R_0(\boldsymbol{V}_0), \boldsymbol{H}\rangle + \int_0^1 (1-t)\left\langle \boldsymbol{H}, \nabla^2 R_0(\boldsymbol{V}_0 + t\boldsymbol{H})\boldsymbol{H}\right\rangle dt.$$

Since $\boldsymbol{V}_0 + t\boldsymbol{H} \in \operatorname{conv}(K)$,

$$R_0(\boldsymbol{V}_1) \geq R_0(\boldsymbol{V}_0) + \langle \nabla R_0(\boldsymbol{V}_0), \boldsymbol{H} \rangle + \frac{\lambda_F(K)}{2} \|\boldsymbol{H}\|_{\mathrm{F}}^2 .$$

Define

$$a_{\mathcal{E}} := \min_{\substack{\boldsymbol{H} \in \mathcal{E} \\ \|\boldsymbol{H}\|=1}} \|\boldsymbol{H}\|_{\mathrm{F}} > 0.$$

Then $\|\boldsymbol{H}\|_{\mathrm{F}} \geq a_{\mathcal{E}} \|\boldsymbol{H}\|$ for all $\boldsymbol{H} \in \mathcal{E}$. Hence

$$R_0(\boldsymbol{V}_1) \geq R_0(\boldsymbol{V}_0) + \langle \nabla R_0(\boldsymbol{V}_0), \boldsymbol{V}_1 - \boldsymbol{V}_0 \rangle + \frac{\lambda_F(K) a_{\mathcal{E}}^2}{2} \|\boldsymbol{V}_1 - \boldsymbol{V}_0\|^2 .$$

Coercivity and strict convexity imply the unique minimizer $\boldsymbol{V}_\star$ of R_0. Every $\boldsymbol{W} = \boldsymbol{W}^{\mathrm{n}} + \boldsymbol{U} + \boldsymbol{V}$ satisfies

$$R(\boldsymbol{W}) = \frac{1}{N} \sum_i \log \left(1 + \sum_{c:(i,c)\in\mathcal{I}_\star} e^{\Delta_{ic}(\boldsymbol{V})} + \sum_{c:(i,c)\in\mathcal{J}_\star} e^{\Delta_{ic}(\boldsymbol{U})+\Delta_{ic}(\boldsymbol{V})} \right) \geq R_0(\boldsymbol{V}) \geq R_0(\boldsymbol{V}_\star).$$

If $\mathcal{J}_\star \neq \varnothing$, take the strict recession direction from Lemma D.2. Then

$$R(\boldsymbol{V}_\star + s\widehat{\boldsymbol{U}}) \to R_0(\boldsymbol{V}_\star) \quad \text{as} \quad s \to \infty.$$

This proves $\inf R = R_0(\boldsymbol{V}_\star)$. At every finite parameter, at least one exponential term at an inactive pair is positive, thus for $\boldsymbol{W} = \boldsymbol{W}^{\mathrm{n}} + \boldsymbol{U} + \boldsymbol{V}$, $R(\boldsymbol{W}) > R_0(\boldsymbol{V}) \geq R_0(\boldsymbol{V}_\star)$, and the infimum is not attained. If instead $\mathcal{J}_\star = \varnothing$, every recession direction has all score gaps zero; hence $\mathcal{C} = \mathcal{M} = \mathcal{L}$ and $R(\boldsymbol{W}) = R_0(\boldsymbol{V})$. The infimum is attained precisely when $\boldsymbol{V} = \boldsymbol{V}_\star$. □

E RISK CONVERGENCE

Let $\mathcal{H}$ be a real vector subspace. For $\boldsymbol{G} \in \mathcal{H}$, the dual of the norm restricted to $\mathcal{H}$ is

$$\|\boldsymbol{G}\|_{\mathcal{H},*} := \max_{\substack{\boldsymbol{H} \in \mathcal{H} \\ \|\boldsymbol{H}\| \leq 1}} |\langle \boldsymbol{G}, \boldsymbol{H} \rangle| .$$

Thus, for every ambient matrix $\boldsymbol{G}$,

$$\|\Pi_{\mathcal{H}} \boldsymbol{G}\|_{\mathcal{H},*} = \max_{\substack{\boldsymbol{H} \in \mathcal{H} \\ \|\boldsymbol{H}\| \leq 1}} |\langle \boldsymbol{G}, \boldsymbol{H} \rangle| \leq \|\boldsymbol{G}\|_* . \tag{23}$$

For $\boldsymbol{U} \in \mathcal{M}$ and $\boldsymbol{V} \in \mathcal{E}$, define

$$n_i(\boldsymbol{U}, \boldsymbol{V}) := \sum_{c:(i,c)\in\mathcal{J}_\star} e^{\Delta_{ic}(\boldsymbol{U})+\Delta_{ic}(\boldsymbol{V})}, \qquad d_i(\boldsymbol{V}) := 1 + \sum_{c:(i,c)\in\mathcal{I}_\star} e^{\Delta_{ic}(\boldsymbol{V})}.$$

Lemma E.1. *For every $\boldsymbol{U} \in \mathcal{M}$ and $\boldsymbol{V} \in \mathcal{E}$,*

$$R(\boldsymbol{U} + \boldsymbol{V}) = R_0(\boldsymbol{V}) + \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + \frac{n_i(\boldsymbol{U}, \boldsymbol{V})}{d_i(\boldsymbol{V})} \right) = R_0(\boldsymbol{V}) + \widetilde{R}(\boldsymbol{U}; \boldsymbol{V}).$$

Proof. This is immediate from factoring $d_i(\boldsymbol{V})$ from each example loss:

$$R(\boldsymbol{U} + \boldsymbol{V}) - R_0(\boldsymbol{V}) = \frac{1}{N} \sum_i \log \left(1 + \frac{\sum_{c:(i,c)\in\mathcal{J}_\star} e^{\Delta_{ic}(\boldsymbol{U})+\Delta_{ic}(\boldsymbol{V})}}{1 + \sum_{c:(i,c)\in\mathcal{I}_\star} e^{\Delta_{ic}(\boldsymbol{V})}} \right). \qquad \square$$

We prove the two cases of Theorem 4.3 using Theorem 3.1 and Lemma A.2. Let $R_\star = R_0(\boldsymbol{V}_\star)$.

Equation 23 gives

$$\|\Pi_{\mathcal{E}} \nabla R(\boldsymbol{W})\|_{\mathcal{E},*} \leq \|\nabla R(\boldsymbol{W})\|_* .$$

For $\delta_t := R(\boldsymbol{W}_t) - R_\star$ and $\delta := R(\boldsymbol{W}) - R_\star$, smoothness gives

$$\begin{aligned} R(\boldsymbol{W}_{t+1}) &\le R(\boldsymbol{W}_t) - \eta_t \left\|\nabla R(\boldsymbol{W}_t)\right\|_* \langle \nabla R(\boldsymbol{W}_t), \boldsymbol{S}_t\rangle + \frac{\beta\eta_t^2 \left\|\nabla R(\boldsymbol{W}_t)\right\|_*^2}{2} \left\|\boldsymbol{S}_t\right\|^2 \\ &\le R(\boldsymbol{W}_t) - \eta_t \left(1 - \frac{\beta\eta_t}{2}\right) \left\|\nabla R(\boldsymbol{W}_t)\right\|_*^2, \end{aligned} \tag{24}$$

and thus $\delta_{t+1} \le \delta_t - \frac{\eta_t}{2} \left\|\nabla R(\boldsymbol{W}_t)\right\|_*^2 \le \delta_t$. If $\delta_0 = 0$, then $\nabla R(\boldsymbol{W}_0) = 0$ and all iterates remain equal to $\boldsymbol{W}_0$. For the sublevel estimates, fix any $b > R_\star$; the trajectory bounds below specialize to $b = R(\boldsymbol{W}_0)$ when $\delta_0 > 0$, and hence $R(\boldsymbol{W}_t) \le b$ for all t. Moreover, the set

$$\mathcal{K}_b := \{\boldsymbol{V} \in \mathcal{E} : R_0(\boldsymbol{V}) \le b\}$$

is compact and convex by Theorem 3.1. If $\mathcal{E} \neq \{\boldsymbol{0}\}$, define

$$\lambda_b := \min_{\substack{\boldsymbol{V}\in\mathcal{K}_b \\ \boldsymbol{H}\in\mathcal{E},\ \|\boldsymbol{H}\|=1}} \left\langle \boldsymbol{H}, \nabla^2 R_0(\boldsymbol{V})\boldsymbol{H}\right\rangle > 0.$$

This is a strong-convexity modulus on $\mathcal{K}_b$ for the restricted norm. If $\mathcal{E} = \{\boldsymbol{0}\}$, all statements involving $\boldsymbol{V}$ are trivial and we may set $\lambda_b = \beta$.

E.1 Nonempty $\mathcal{J}_\star$

Recall

$$\gamma_\star := \max_{\substack{\boldsymbol{U}\in\mathcal{M} \\ \|\boldsymbol{U}\|\le 1}} \min_{(i,c)\in\mathcal{J}_\star} \{-\Delta_{ic}(\boldsymbol{U})\} > 0.$$

Lemma D.2 and compactness of the unit ball give a max-margin matrix satisfying

$$\boldsymbol{U}_\star \in \arg\max_{\substack{\boldsymbol{U}\in\mathcal{M} \\ \|\boldsymbol{U}\|\le 1}} \min_{(i,c)\in\mathcal{J}_\star} \{-\Delta_{ic}(\boldsymbol{U})\}, \qquad \|\boldsymbol{U}_\star\| = 1, \qquad \Delta_{ic}(\boldsymbol{U}_\star) \le -\gamma_\star < 0 \quad ((i,c) \in \mathcal{J}_\star).$$

1. Excess-risk decomposition. Fix $\boldsymbol{W} = \boldsymbol{W}^{\mathrm{n}} + \boldsymbol{U} + \boldsymbol{V}$ with $R(\boldsymbol{W}) \le b$. Define

$$\zeta(\boldsymbol{W}) := \frac{1}{N} \sum_{(i,c)\in\mathcal{J}_\star} p_{ic}(\boldsymbol{W}) = \frac{1}{N} \sum_i \frac{n_i}{d_i + n_i}.$$

For $s > 0$,

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{1 - e^{-s}}{s}\right) = \frac{(1+s)e^{-s} - 1}{s^2} \le 0, \qquad \lim_{s\downarrow 0} \frac{1 - e^{-s}}{s} = 1.$$

Let $c_b := \frac{1 - e^{-Nb}}{Nb} > 0$. Then, for $s \in [0, Nb]$, $c_b s \le 1 - e^{-s} \le s$. Substitute $s = \log(1 + n_i/d_i)$:

$$c_b \widetilde{R}(\boldsymbol{U}; \boldsymbol{V}) \le \zeta(\boldsymbol{W}) \le \widetilde{R}(\boldsymbol{U}; \boldsymbol{V}). \tag{25}$$

Hence,

$$\begin{aligned} \left\|\nabla R(\boldsymbol{W})\right\|_* &\ge -\langle \nabla R(\boldsymbol{W}), \boldsymbol{U}_\star\rangle = -\frac{1}{N} \sum_{(i,c)\in\mathcal{I}} p_{ic}(\boldsymbol{W}) \Delta_{ic}(\boldsymbol{U}_\star) \\ &= -\frac{1}{N} \sum_{(i,c)\in\mathcal{J}_\star} p_{ic}(\boldsymbol{W}) \Delta_{ic}(\boldsymbol{U}_\star) \ge \gamma_\star \zeta(\boldsymbol{W}) \ge \gamma_\star c_b \widetilde{R}(\boldsymbol{U}; \boldsymbol{V}). \end{aligned} \tag{26}$$

2. Derivative of $\widetilde{R}$. Define

$$c_4 := 2 \max_{(i,c)\in\mathcal{I}} \left\|\Pi_{\mathcal{E}} \boldsymbol{M}_{(i,c)}\right\|_{\mathcal{E},*} = 2 \max_{(i,c)\in\mathcal{I}} \sup_{\substack{\boldsymbol{H}\in\mathcal{E} \\ \|\boldsymbol{H}\|\le 1}} |\Delta_{ic}(\boldsymbol{H})|.$$

In particular, $c_4 = 0$ when $\mathcal{E} = \{\mathbf{0}\}$. All gradients with respect to $\boldsymbol{V}$ are taken within $\mathcal{E}$. Since the gradient of Δ_{ic} within $\mathcal{E}$ is $\Pi_{\mathcal{E}}\boldsymbol{M}_{(i,c)}$,

$$\nabla_{\boldsymbol{V}} n_i(\boldsymbol{U},\boldsymbol{V}) = \sum_{c:(i,c)\in\mathcal{J}_\star} e^{\Delta_{ic}(\boldsymbol{U})+\Delta_{ic}(\boldsymbol{V})}\Pi_{\mathcal{E}}\boldsymbol{M}_{(i,c)},$$
$$\nabla d_i(\boldsymbol{V}) = \sum_{c:(i,c)\in\mathcal{I}_\star} e^{\Delta_{ic}(\boldsymbol{V})}\Pi_{\mathcal{E}}\boldsymbol{M}_{(i,c)}.$$

Thus,

$$\|\nabla_{\boldsymbol{V}} n_i\|_{\mathcal{E},*} \le \frac{c_4}{2} n_i, \qquad \|\nabla d_i\|_{\mathcal{E},*} \le \frac{c_4}{2}(d_i - 1).$$

Apply the chain rule:

$$\nabla_{\boldsymbol{V}} \log\left(1+\frac{n_i}{d_i}\right) = \frac{\nabla_{\boldsymbol{V}} n_i}{d_i+n_i} - \frac{n_i}{d_i+n_i}\frac{\nabla d_i}{d_i},$$
$$\left\|\nabla_{\boldsymbol{V}} \log\left(1+\frac{n_i}{d_i}\right)\right\|_{\mathcal{E},*} \le \frac{c_4}{2}\frac{n_i}{d_i+n_i}\left(1+\frac{d_i-1}{d_i}\right) \le c_4\frac{n_i}{d_i+n_i}.$$

Applying Equation 25 yields:

$$\left\|\nabla_{\boldsymbol{V}}\widetilde{R}(\boldsymbol{U};\boldsymbol{V})\right\|_{\mathcal{E},*} \le c_4\zeta(\boldsymbol{W}) \le c_4\widetilde{R}(\boldsymbol{U};\boldsymbol{V}). \tag{27}$$

3. Gradient of R_0. Because $\boldsymbol{V}, \boldsymbol{V}_\star \in \mathcal{K}_b$, strong convexity implies

$$R_0(\boldsymbol{V}_\star) \ge R_0(\boldsymbol{V}) + \langle \nabla R_0(\boldsymbol{V}), \boldsymbol{V}_\star - \boldsymbol{V}\rangle + \frac{\lambda_b}{2}\|\boldsymbol{V}-\boldsymbol{V}_\star\|^2.$$

Then,

$$\begin{aligned} R_0(\boldsymbol{V}) - R_\star &\le \langle \nabla R_0(\boldsymbol{V}), \boldsymbol{V}-\boldsymbol{V}_\star\rangle - \frac{\lambda_b}{2}\|\boldsymbol{V}-\boldsymbol{V}_\star\|^2 \\ &\le \|\nabla R_0(\boldsymbol{V})\|_{\mathcal{E},*}\|\boldsymbol{V}-\boldsymbol{V}_\star\| - \frac{\lambda_b}{2}\|\boldsymbol{V}-\boldsymbol{V}_\star\|^2 \\ &\le \frac{\|\nabla R_0(\boldsymbol{V})\|_{\mathcal{E},*}^2}{2\lambda_b}. \end{aligned} \tag{28}$$

Let $c_5 := \sqrt{\frac{2\lambda_b}{b-R_\star}}$. Since $0 \le (R_0(\boldsymbol{V}) - R_\star) \le b - R_\star$, it follows that

$$\|\nabla R_0(\boldsymbol{V})\|_{\mathcal{E},*} \ge \sqrt{2\lambda_b(R_0(\boldsymbol{V})-R_\star)} \ge c_5(R_0(\boldsymbol{V})-R_\star).$$

Differentiating Lemma E.1 within $\mathcal{E}$ gives

$$\Pi_{\mathcal{E}}\nabla R(\boldsymbol{W}) = \nabla R_0(\boldsymbol{V}) + \nabla_{\boldsymbol{V}}\widetilde{R}(\boldsymbol{U};\boldsymbol{V}).$$

Equations 27 and 28 give

$$\begin{aligned} \|\nabla R(\boldsymbol{W})\|_* &\ge \|\Pi_{\mathcal{E}}\nabla R(\boldsymbol{W})\|_{\mathcal{E},*} \\ &\ge \|\nabla R_0(\boldsymbol{V})\|_{\mathcal{E},*} - \left\|\nabla_{\boldsymbol{V}}\widetilde{R}(\boldsymbol{U};\boldsymbol{V})\right\|_{\mathcal{E},*} \\ &\ge \sqrt{2\lambda_b(R_0(\boldsymbol{V})-R_\star)} - c_4\widetilde{R}(\boldsymbol{U};\boldsymbol{V}) \ge c_5(R_0(\boldsymbol{V})-R_\star) - c_4\widetilde{R}(\boldsymbol{U};\boldsymbol{V}). \end{aligned} \tag{29}$$

4. Sublevel gradient bound. Multiply Equation 26 by $c_4 + c_5$ and Equation 29 by $\gamma_\star c_b$, and add:

$$(\gamma_\star c_b + c_4 + c_5)\|\nabla R(\boldsymbol{W})\|_* \ge \gamma_\star c_b c_5(R_0(\boldsymbol{V}) - R_\star + \widetilde{R}(\boldsymbol{U};\boldsymbol{V})) = \gamma_\star c_b c_5\delta.$$

Set

$$\kappa_b := \frac{\gamma_\star c_b c_5}{\gamma_\star c_b + c_5 + c_4} > 0. \tag{30}$$

This proves Lemma 4.2 for $\mathcal{J}_\star \neq \varnothing$:

$$\|\nabla R(\boldsymbol{W})\|_* \ge \kappa_b(R(\boldsymbol{W}) - R_\star) \qquad (R(\boldsymbol{W}) \le b).$$

We now prove Theorem 4.3 when $\mathcal{J}_\star \neq \varnothing$. Take $b = R(\boldsymbol{W}_0)$. Since $\mathcal{J}_\star \neq \varnothing$, Theorem 3.1 implies $\delta_t > 0$ at every finite iterate. Combining the gradient bound with Equation 24,

$$\delta_t - \delta_{t+1} \geq \frac{\eta_t \kappa_b^2}{2} \delta_t^2, \qquad \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} = \frac{\delta_t - \delta_{t+1}}{\delta_t \delta_{t+1}} \geq \frac{\eta_t \kappa_b^2}{2} \frac{\delta_t}{\delta_{t+1}} \geq \frac{\eta_t \kappa_b^2}{2}.$$

Summation from $j = 0$ to $t - 1$ yields

$$\delta_t^{-1} \geq \delta_0^{-1} + \frac{\kappa_b^2}{2} \sum_{j<t} \eta_j, \qquad R(\boldsymbol{W}_t) - R_\star \leq \left(\delta_0^{-1} + \frac{\kappa_b^2}{2} \sum_{j<t} \eta_j \right)^{-1}.$$

This proves Equation 8. If $\eta_t = \eta > 0$, then $\delta_t = O(t^{-1})$; if $\eta_t \asymp (t+1)^{-1/2}$, then $\delta_t = O(t^{-1/2})$.

E.2 Empty $\mathcal{J}_\star$

In this case, $\mathcal{M} = \mathcal{L}$ and

$$R(\boldsymbol{W}^{\mathrm{n}} + \boldsymbol{U} + \boldsymbol{V}) = R_0(\boldsymbol{V}).$$

Thus, the recession cone consists only of directions that do not change the loss. Fix any $b > R_\star$ and any $\boldsymbol{W}$ with $R(\boldsymbol{W}) \leq b$. From Equation 28,

$$\delta \leq \frac{\|\nabla R_0(\boldsymbol{V})\|_{\mathcal{E},*}^2}{2\lambda_b}.$$

In addition,

$$\Pi_{\mathcal{E}} \nabla R(\boldsymbol{W}) = \nabla R_0(\boldsymbol{V}), \qquad \|\nabla R(\boldsymbol{W})\|_*^2 \geq \|\nabla R_0(\boldsymbol{V})\|_{\mathcal{E},*}^2 \geq 2\lambda_b \delta.$$

In this case, $R|_{\mathcal{E}} = R_0$, and thus smoothness gives $\lambda_b \leq \beta$. This also proves Lemma 4.2 in the empty case, with $\kappa_b = \sqrt{2\lambda_b/(b - R_\star)}$, because $0 \leq \delta \leq b - R_\star$. Regarding the rate, assume $\delta_0 > 0$; otherwise the trajectory is stationary from the beginning. Take $b = R(\boldsymbol{W}_0) > R_\star$. Equation 24 gives

$$0 \leq \delta_{t+1} \leq \delta_t - \frac{\eta_t}{2} \|\nabla R(\boldsymbol{W}_t)\|_*^2 \leq (1 - \lambda_b \eta_t) \delta_t, \qquad 0 \leq 1 - \lambda_b \eta_t \leq 1.$$

Therefore

$$R(\boldsymbol{W}_t) - R_\star = \delta_t \leq \delta_0 \prod_{j=0}^{t-1} (1 - \lambda_b \eta_j) \leq \delta_0 e^{-\lambda_b \sum_{j<t} \eta_j}.$$

When $\mathcal{E} = \{\boldsymbol{0}\}$ as well, the risk is constant and $\delta_t = 0$ for every t.

Convergence of $\boldsymbol{V}_t$. We have

$$\frac{\lambda_b}{2} \|\boldsymbol{V}_t - \boldsymbol{V}_\star\|^2 \leq R_0(\boldsymbol{V}_t) - R_\star \leq \delta_t.$$

Consequently, Equation 9 holds and

$$\|\boldsymbol{V}_t - \boldsymbol{V}_\star\|^2 = \begin{cases} O((\sum_{j<t} \eta_j)^{-1}), & \mathcal{J}_\star \neq \varnothing, \\ O(e^{-\lambda_b \sum_{j<t} \eta_j}), & \mathcal{J}_\star = \varnothing. \end{cases}$$

F Margin and directional convergence

Assume $\mathcal{J}_\star \neq \varnothing$, $\boldsymbol{W}_0 = 0$, and Assumption 5.1. Set $b := R(\boldsymbol{W}_0) = \log K$. Theorem 4.3 implies

$$0 < \delta_t \leq \left(\delta_0^{-1} + \frac{\kappa_b^2}{2} \sum_{j=0}^{t-1} \eta_j \right)^{-1}. \tag{31}$$

When $\mathcal{E} \neq \{\boldsymbol{0}\}$, we restate the constants in Section E here:

$$\lambda_b := \min_{\substack{\boldsymbol{V} \in \mathcal{E},\ R_0(\boldsymbol{V}) \leq b \\ \boldsymbol{H} \in \mathcal{E},\ \|\boldsymbol{H}\| = 1}} \left\langle \boldsymbol{H}, \nabla^2 R_0(\boldsymbol{V}) \boldsymbol{H} \right\rangle > 0, \qquad c_4 := 2 \max_{(i,c) \in \mathcal{I}} \sup_{\substack{\boldsymbol{H} \in \mathcal{E} \\ \|\boldsymbol{H}\| \leq 1}} |\Delta_{ic}(\boldsymbol{H})|.$$

Again, if $\mathcal{E} = \{\boldsymbol{0}\}$, all statements involving $\boldsymbol{V}$ are trivial and we may set $\lambda_b = \beta$ and $c_4 = 0$.

F.1 Gradient and excess risk

Lemma F.1. *Define the constants*

$$c_6 := (\gamma_\star + c_4)^2/(2\lambda_b), \qquad c_7 := \gamma_\star(N + c_6) > 0.$$

Then, for every $t \geq 0$,

$$\|\nabla R(\boldsymbol{W}_t)\|_* \geq \gamma_\star \delta_t - c_7 \delta_t^2. \tag{32}$$

Proof. Decompose the risk

$$\delta_t = R_0(\boldsymbol{V}_t) - R_\star + \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + \frac{n_i(\boldsymbol{U}_t, \boldsymbol{V}_t)}{d_i(\boldsymbol{V}_t)}\right), \qquad 0 \leq \log\left(1 + \frac{n_i(\boldsymbol{U}_t, \boldsymbol{V}_t)}{d_i(\boldsymbol{V}_t)}\right) \leq N\delta_t.$$

For $0 \leq s \leq N\delta_t$, $1 - e^{-s} = e^{-s}(e^s - 1) \geq s e^{-N\delta_t}$. Hence

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} \frac{n_i(\boldsymbol{U}_t, \boldsymbol{V}_t)}{d_i(\boldsymbol{V}_t) + n_i(\boldsymbol{U}_t, \boldsymbol{V}_t)} &= \frac{1}{N}\sum_{i=1}^{N}\left[1 - \exp\left(-\log\left(1 + \frac{n_i(\boldsymbol{U}_t, \boldsymbol{V}_t)}{d_i(\boldsymbol{V}_t)}\right)\right)\right] \\ &\geq e^{-N\delta_t}(R(\boldsymbol{W}_t) - R_0(\boldsymbol{V}_t)).\end{aligned}$$

Take any primal max-margin matrix $\boldsymbol{U}_\star \in \mathcal{U}$. $-\Delta_{ic}(\boldsymbol{U}_\star) \geq \gamma_\star$ for $(i, c) \in \mathcal{J}_\star$. Therefore

$$\begin{aligned}\|\nabla R(\boldsymbol{W}_t)\|_* &\geq -\langle \nabla R(\boldsymbol{W}_t), \boldsymbol{U}_\star \rangle \\ &= -\frac{1}{N}\sum_{(i,c)\in\mathcal{J}_\star} \frac{e^{\Delta_{ic}(\boldsymbol{U}_t) + \Delta_{ic}(\boldsymbol{V}_t)}\Delta_{ic}(\boldsymbol{U}_\star)}{d_i(\boldsymbol{V}_t) + n_i(\boldsymbol{U}_t, \boldsymbol{V}_t)} \\ &\geq \gamma_\star e^{-N\delta_t}(R(\boldsymbol{W}_t) - R_0(\boldsymbol{V}_t)).\end{aligned} \tag{33}$$

When $\mathcal{E} \neq \{\boldsymbol{0}\}$, Equations 28 and 29 give

$$\|\nabla R(\boldsymbol{W}_t)\|_* \geq \sqrt{2\lambda_b(R_0(\boldsymbol{V}_t) - R_\star)} - c_4(R(\boldsymbol{W}_t) - R_0(\boldsymbol{V}_t)). \tag{34}$$

If $R_0(\boldsymbol{V}_t) - R_\star > c_6\delta_t^2$, then

$$\|\nabla R(\boldsymbol{W}_t)\|_* > (\gamma_\star + c_4)\delta_t - c_4(R(\boldsymbol{W}_t) - R_0(\boldsymbol{V}_t)) \geq \gamma_\star \delta_t.$$

Otherwise $R(\boldsymbol{W}_t) - R_0(\boldsymbol{V}_t) \geq \delta_t - c_6\delta_t^2$. If $\delta_t < (N + c_6)^{-1}$, Equation 33 gives

$$\begin{aligned}\|\nabla R(\boldsymbol{W}_t)\|_* &\geq \gamma_\star(1 - N\delta_t)(\delta_t - c_6\delta_t^2) \\ &= \gamma_\star\delta_t - \gamma_\star(N + c_6)\delta_t^2 + \gamma_\star N c_6 \delta_t^3 \geq \gamma_\star\delta_t - c_7\delta_t^2.\end{aligned}$$

If $\delta_t \geq (N + c_6)^{-1}$, the right side of Equation 32 is nonpositive, so the inequality is automatic. When $\mathcal{E} = \{\boldsymbol{0}\}$, use $R(\boldsymbol{W}_t) - R_0(\boldsymbol{V}_t) = \delta_t$ in Equation 33; the same conclusion holds. □

F.2 Summability of error terms

Lemma F.2. *Define the finite constants*

$$c_8 := 1 + \frac{\delta_0\kappa_b^2}{2\beta}, \qquad c_9 := \frac{4}{\kappa_b^2} c_8^{3/2}\sqrt{2\beta\delta_0},$$

$$c_{10} := \eta_0^2\sqrt{2\beta\delta_0} + \sqrt{\frac{4\beta}{\kappa_b^2}}\sum_{t=1}^{\infty} \frac{\eta_t^2}{\left(\sum_{j=0}^{t-1}\eta_j\right)^{1/2}}.$$

Then

$$\sum_{t=0}^{\infty} \eta_t\delta_t \|\nabla R(\boldsymbol{W}_t)\|_* \leq c_9, \qquad \sum_{t=0}^{\infty} \eta_t^2 \|\nabla R(\boldsymbol{W}_t)\|_* \leq c_{10}. \tag{35}$$

Proof. Note that

$$R_\star \le R\left(\boldsymbol{W}_t - \frac{\|\nabla R(\boldsymbol{W}_t)\|_*}{\beta}\boldsymbol{S}_t\right) \le R(\boldsymbol{W}_t) - \frac{\|\nabla R(\boldsymbol{W}_t)\|_*^2}{2\beta}, \qquad \|\nabla R(\boldsymbol{W}_t)\|_* \le \sqrt{2\beta\delta_t}.$$

For $t \ge 1$, Equation 31 implies

$$\|\nabla R(\boldsymbol{W}_t)\|_* \le \sqrt{\frac{4\beta}{\kappa_b^2}}\frac{1}{\left(\sum_{j=0}^{t-1}\eta_j\right)^{1/2}}.$$

This proves the second bound in Equation 35 from Assumption 5.1. For the first bound, note that

$$\frac{\delta_0^{-1} + \frac{\kappa_b^2}{2}\sum_{j=0}^{t}\eta_j}{\delta_0^{-1} + \frac{\kappa_b^2}{2}\sum_{j=0}^{t-1}\eta_j} \le 1 + \frac{\delta_0\kappa_b^2\eta_t}{2} \le c_8.$$

Thus

$$\begin{aligned}\int_{\delta_0^{-1}+\frac{1}{2}\kappa_b^2\sum_{j=0}^{t-1}\eta_j}^{\delta_0^{-1}+\frac{1}{2}\kappa_b^2\sum_{j=0}^{t}\eta_j} s^{-3/2}\,\mathrm{d}s &\ge \frac{\kappa_b^2\eta_t}{2}\left(\delta_0^{-1} + \frac{1}{2}\kappa_b^2\sum_{j=0}^{t}\eta_j\right)^{-3/2} \\ &\ge \frac{\kappa_b^2\eta_t}{2c_8^{3/2}}\left(\delta_0^{-1} + \frac{1}{2}\kappa_b^2\sum_{j=0}^{t-1}\eta_j\right)^{-3/2}.\end{aligned}$$

Together with Equation 31,

$$\begin{aligned}\sum_{t=0}^{\infty}\eta_t\delta_t\|\nabla R(\boldsymbol{W}_t)\|_* &\le \sqrt{2\beta}\sum_{t=0}^{\infty}\eta_t\left(\delta_0^{-1} + \frac{\kappa_b^2}{2}\sum_{j=0}^{t-1}\eta_j\right)^{-3/2} \\ &\le \frac{2}{\kappa_b^2}\sqrt{2\beta}c_8^{3/2}\int_{\delta_0^{-1}}^{\infty}s^{-3/2}\,\mathrm{d}s = \frac{4}{\kappa_b^2}c_8^{3/2}\sqrt{2\beta\delta_0} = c_9.\end{aligned}$$

□

Lemma F.3 (Logarithmic excess-risk)**.** *With*

$$c_{11} := |\log\delta_0| + \frac{\beta\gamma_\star}{2}c_{10} + c_7c_9,$$

one has, for every $t \ge 0$,

$$\log\delta_t \le c_{11} - \gamma_\star\sum_{j=0}^{t-1}\eta_j\|\nabla R(\boldsymbol{W}_j)\|_*. \tag{36}$$

Proof. Smoothness yields

$$0 < \delta_{t+1} \le \delta_t - \eta_t\left(1 - \frac{\beta\eta_t}{2}\right)\|\nabla R(\boldsymbol{W}_t)\|_*^2.$$

Apply $-\log(1-s) \ge s$ on $[0,1)$ and Equation 32:

$$\begin{aligned}\log\frac{\delta_t}{\delta_{t+1}} &\ge -\log\left(1 - \eta_t\left(1 - \frac{\beta\eta_t}{2}\right)\frac{\|\nabla R(\boldsymbol{W}_t)\|_*^2}{\delta_t}\right) \\ &\ge \eta_t\|\nabla R(\boldsymbol{W}_t)\|_*\left(1 - \frac{\beta\eta_t}{2}\right)\frac{\|\nabla R(\boldsymbol{W}_t)\|_*}{\delta_t} \\ &\ge \eta_t\|\nabla R(\boldsymbol{W}_t)\|_*\left(\gamma_\star - \frac{\beta\gamma_\star}{2}\eta_t - c_7\delta_t\right).\end{aligned}$$

Applying Equation 35 gives

$$\begin{aligned}\log\delta_t &\le \log\delta_0 - \gamma_\star\sum_{j=0}^{t-1}\eta_j\|\nabla R(\boldsymbol{W}_j)\|_* + \frac{\beta\gamma_\star}{2}c_{10} + c_7c_9 \\ &\le c_{11} - \gamma_\star\sum_{j=0}^{t-1}\eta_j\|\nabla R(\boldsymbol{W}_j)\|_*.\end{aligned}$$

□

F.3 INACTIVE SCORE GAPS AND EXCESS RISK

Define the following constants, which are finite:

$$c_{12} := Nb + \log((e-1)N) + \max_{\substack{\boldsymbol{V}\in\mathcal{E},\ R_0(\boldsymbol{V})\le b \\ (i,c)\in\mathcal{J}_\star}} |\Delta_{ic}(\boldsymbol{V})|, \qquad c_{13} := \max_{\substack{\boldsymbol{V}\in\mathcal{E} \\ R_0(\boldsymbol{V})\le b}} \|\boldsymbol{V}\|$$

$$c_{14} := \max_{(i,c)\in\mathcal{J}_\star} \sup_{\substack{\boldsymbol{U}\in\mathcal{M} \\ \|\boldsymbol{U}\|\le 1}} |\Delta_{ic}(\boldsymbol{U})| \ge \gamma_\star > 0, \qquad c_{15} := (c_{11}+c_{12}) + \gamma_\star c_{13} > 0.$$

Here $c_{12} > 0$ because $N \ge 1$, $b \ge 0$, and $e - 1 > 1$.

Lemma F.4. *For all sufficiently large t and every $(i,c) \in \mathcal{J}_\star$,*

$$-\Delta_{ic}(\boldsymbol{U}_t) \ge -\log\delta_t - c_{12} \ge \gamma_\star \sum_{j=0}^{t-1} \eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_* - (c_{11}+c_{12}).$$

Hence,

$$-\Delta_{ic}(\boldsymbol{U}_t) \ge \log\left(\delta_0^{-1} + \frac{\kappa_b^2}{2}\sum_{j=0}^{t-1}\eta_j\right) - c_{12}. \tag{37}$$

Proof. Note that

$$0 \le \log d_i(\boldsymbol{V}_t) \le \sum_{l=1}^{N} \log d_l(\boldsymbol{V}_t) = N R_0(\boldsymbol{V}_t) \le Nb.$$

Also,

$$\delta_t \ge \frac{1}{N}\log\left(1 + \frac{e^{\Delta_{ic}(\boldsymbol{U}_t)+\Delta_{ic}(\boldsymbol{V}_t)}}{d_i(\boldsymbol{V}_t)}\right) \quad \Rightarrow \quad e^{\Delta_{ic}(\boldsymbol{U}_t)+\Delta_{ic}(\boldsymbol{V}_t)} \le d_i(\boldsymbol{V}_t)(e^{N\delta_t}-1).$$

Since $\delta_t \to 0$, eventually $N\delta_t \le 1$ and $e^{N\delta_t} - 1 \le (e-1)N\delta_t$. Therefore

$$\begin{aligned} -\Delta_{ic}(\boldsymbol{U}_t) &\ge -\log\delta_t - \log((e-1)N) - \log d_i(\boldsymbol{V}_t) + \Delta_{ic}(\boldsymbol{V}_t) \\ &\ge -\log\delta_t - c_{12}. \end{aligned}$$

Now apply Lemma F.3 and Equation 31. □

Proof of Theorem 5.2. Every SD increment has norm $\eta_t \left\|\nabla R(\boldsymbol{W}_t)\right\|_*$. Since $\boldsymbol{W}_0 = 0$ and $\boldsymbol{W}_t = \boldsymbol{U}_t + \boldsymbol{V}_t$,

$$\|\boldsymbol{W}_t\| \le \sum_{j=0}^{t-1} \eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_*, \qquad \|\boldsymbol{U}_t\| \le \sum_{j=0}^{t-1} \eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_* + c_{13}. \tag{38}$$

Fix $(i,c) \in \mathcal{J}_\star$. As $\|\boldsymbol{V}_t\| \le c_{13}$,

$$-\Delta_{ic}(\boldsymbol{U}_t) \le |\Delta_{ic}(\boldsymbol{U}_t)| \le c_{14}\|\boldsymbol{U}_t\| \le c_{14}\left(\sum_{j=0}^{t-1}\eta_j\|\nabla R(\boldsymbol{W}_j)\|_* + c_{13}\right).$$

Equation 37 then gives

$$\sum_{j=0}^{t-1} \eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_* \ge \frac{1}{c_{14}}\log\left(\delta_0^{-1} + \frac{\kappa_b^2}{2}\sum_{j=0}^{t-1}\eta_j\right) - \frac{c_{12}}{c_{14}} - c_{13} \to \infty. \tag{39}$$

Applying $\frac{1}{2}\sum_{j=0}^{t-1}\eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_*^2 \le \delta_0$ and Cauchy–Schwarz inequality further gives

$$\left(\sum_{j=0}^{t-1}\eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_*\right)^2 \le \left(\sum_{j=0}^{t-1}\eta_j\right)\left(\sum_{j=0}^{t-1}\eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_*^2\right) \le 2\delta_0 \sum_{j=0}^{t-1}\eta_j. \tag{40}$$

By positive homogeneity of the minimum margin and its maximum over the unit ball,

$$\min_{(i,c)\in\mathcal{J}_\star}\{-\Delta_{ic}(\boldsymbol{U})\} \le \gamma_\star \|\boldsymbol{U}\| \qquad (\boldsymbol{U}\in\mathcal{M}).$$

Lemma F.4 and Equation 38 imply

$$\sum_{j=0}^{t-1}\eta_j \|\nabla R(\boldsymbol{W}_j)\|_* - \frac{c_{11}+c_{12}}{\gamma_\star} \le \|\boldsymbol{U}_t\| \le \sum_{j=0}^{t-1}\eta_j \|\nabla R(\boldsymbol{W}_j)\|_* + c_{13}. \tag{41}$$

In particular, $\|\boldsymbol{U}_t\|\to\infty$. For all sufficiently large t, the lower bound of $\min_{(i,c)\in\mathcal{J}_\star}\{-\Delta_{ic}(\boldsymbol{U}_t)\}$ is positive, hence:

$$\begin{aligned}\frac{\min_{(i,c)\in\mathcal{J}_\star}\{-\Delta_{ic}(\boldsymbol{U}_t)\}}{\|\boldsymbol{U}_t\|} &\ge \frac{\gamma_\star\sum_{j=0}^{t-1}\eta_j\|\nabla R(\boldsymbol{W}_j)\|_* - (c_{11}+c_{12})}{\sum_{j=0}^{t-1}\eta_j\|\nabla R(\boldsymbol{W}_j)\|_* + c_{13}}\\ &= \gamma_\star - \frac{c_{15}}{\sum_{j=0}^{t-1}\eta_j\|\nabla R(\boldsymbol{W}_j)\|_* + c_{13}}.\end{aligned}$$

The normalized margin is at most $\gamma_\star$ by definition. Equation 39 converts the denominator to the inverse-logarithmic rate in Equation 10. □

F.4 Directional convergence rate from modulus of convexity

Lemma F.5 (Generalized inverse of the modulus)**.** *Suppose the norm (restricted to $\mathcal{M}$) is strictly convex. Then $\omega_{\mathcal{M}}$ and $\omega_{\mathcal{M}}^\dagger$ are nondecreasing. Moreover,*

$$\omega_{\mathcal{M}}(0)=0, \qquad \omega_{\mathcal{M}}(r)>0\ (0<r\le 2), \qquad \lim_{s\downarrow 0}\omega_{\mathcal{M}}^\dagger(s)=0.$$

For $r\in[0,2]$ and $s\ge 0$,

$$\omega_{\mathcal{M}}(r)\le s \quad\Rightarrow\quad r\le\omega_{\mathcal{M}}^\dagger(s). \tag{42}$$

If $\omega_{\mathcal{M}}(r)\ge c_\omega r^q$ for some $c_\omega>0$ and $q\ge 2$, then

$$\omega_{\mathcal{M}}^\dagger(s)\le\min\{2,(s/c_\omega)^{1/q}\}.$$

Proof. For every $r\in[0,2]$, the set

$$\{\boldsymbol{H},\boldsymbol{H}'\in\mathcal{M}:\ \|\boldsymbol{H}\|=\|\boldsymbol{H}'\|=1,\ \|\boldsymbol{H}-\boldsymbol{H}'\|\ge r\}$$

is compact and nonempty. Thus $\omega_{\mathcal{M}}$ attains the minimum. For $r=0$, $\boldsymbol{H}'=\boldsymbol{H}$ gives $\omega_{\mathcal{M}}(0)=0$. In finite dimensional (Banach) spaces, strict convexity is equivalent to uniform convexity, which is equivalent to $\omega_{\mathcal{M}}(r)>0$ for all $r>0$ (Lindenstrauss & Tzafriri, 1979, Section 1.e). It is nondecreasing as the feasible set shrinks with r; monotonicity of $\omega_{\mathcal{M}}^\dagger$ immediately follows.

The set at each $s\ge 0$

$$\{r\in[0,2]:\omega_{\mathcal{M}}(r)\le s\}$$

contains zero and is contained in $[0,2]$, thus its supremum exists. Equation 42 follows directly from the definition. For every $\varepsilon\in(0,2]$, monotonicity gives $s<\omega_{\mathcal{M}}(\varepsilon)\Rightarrow\omega_{\mathcal{M}}^\dagger(s)\le\varepsilon$. Since $\omega_{\mathcal{M}}(\varepsilon)>0$, this proves $\lim_{s\downarrow0}\omega_{\mathcal{M}}^\dagger(s)=0$. Finally, $\omega_{\mathcal{M}}(r)\le s$ and $\omega_{\mathcal{M}}(r)\ge c_\omega r^q$ imply $r\le(s/c_\omega)^{1/q}$. □

Proof of Theorem 5.3. It is immediate that the max-margin direction is unique; otherwise, choose two distinct directions $\boldsymbol{U}_1,\boldsymbol{U}_2\in\mathcal{U}$ and take their average. As the norm is strictly convex, $\|(\boldsymbol{U}_1+\boldsymbol{U}_2)/2\|<1$ but the average yields the same margin $\gamma_\star$. Normalizing the average direction to have a norm of one contradicts the maximality of max-margin directions. Hence $\mathcal{U}=\{\boldsymbol{U}_\star\}$. Then,

$$\begin{aligned}\gamma_\star\left\|\frac{\boldsymbol{U}_t/\|\boldsymbol{U}_t\|+\boldsymbol{U}_\star}{2}\right\| &\ge \min_{(i,c)\in\mathcal{J}_\star}\left\{-\Delta_{ic}\left(\frac{\boldsymbol{U}_t/\|\boldsymbol{U}_t\|+\boldsymbol{U}_\star}{2}\right)\right\}\\ &\ge \frac{1}{2}\left(\frac{\min_{(i,c)\in\mathcal{J}_\star}\{-\Delta_{ic}(\boldsymbol{U}_t)\}}{\|\boldsymbol{U}_t\|}+\gamma_\star\right).\end{aligned}$$

Hence,

$$\begin{aligned}\omega_{\mathcal{M}}\left(\left\|\frac{\boldsymbol{U}_t}{\|\boldsymbol{U}_t\|}-\boldsymbol{U}_\star\right\|\right) &\le 1-\left\|\frac{\boldsymbol{U}_t/\|\boldsymbol{U}_t\|+\boldsymbol{U}_\star}{2}\right\| \\ &\le \frac{1}{2\gamma_\star}\left(\gamma_\star-\frac{\min_{(i,c)\in\mathcal{J}_\star}\{-\Delta_{ic}(\boldsymbol{U}_t)\}}{\|\boldsymbol{U}_t\|}\right) \\ &\le \frac{c_{15}}{2\gamma_\star\left(\sum_{j=0}^{t-1}\eta_j\left\|\nabla R(\boldsymbol{W}_j)\right\|_*+c_{13}\right)}\to 0.\end{aligned}$$

Applying Equation 42 to this bound and using $\omega_{\mathcal{M}}^{\dagger}(s)\to 0$ as $s\downarrow 0$ proves Equation 17.

For $\boldsymbol{W}_t$, we first derive an inequality between two normalized matrices. For nonzero matrices $\boldsymbol{H},\boldsymbol{H}'$,

$$\begin{aligned}\left\|\frac{\boldsymbol{H}}{\|\boldsymbol{H}\|}-\frac{\boldsymbol{H}'}{\|\boldsymbol{H}'\|}\right\| &\le \frac{\|\boldsymbol{H}-\boldsymbol{H}'\|}{\|\boldsymbol{H}\|}+\|\boldsymbol{H}'\|\left|\frac{1}{\|\boldsymbol{H}\|}-\frac{1}{\|\boldsymbol{H}'\|}\right| \\ &\le \frac{2\|\boldsymbol{H}-\boldsymbol{H}'\|}{\|\boldsymbol{H}\|}.\end{aligned} \tag{43}$$

Equation 43 applied with $\boldsymbol{H}=\boldsymbol{U}_t$ and $\boldsymbol{H}'=\boldsymbol{W}_t$ gives

$$\left\|\frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}-\frac{\boldsymbol{U}_t}{\|\boldsymbol{U}_t\|}\right\|\le\frac{2\|\boldsymbol{V}_t\|}{\|\boldsymbol{U}_t\|}\le\frac{2c_{13}}{\sum_{j=0}^{t-1}\eta_j\left\|\nabla R(\boldsymbol{W}_j)\right\|_*-(c_{11}+c_{12})/\gamma_\star}\to 0.$$

In particular,

$$\begin{aligned}\left\|\frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}-\boldsymbol{U}_\star\right\| &\le \omega_{\mathcal{M}}^{\dagger}\left(\frac{c_{15}}{2\gamma_\star\left(\sum_{j=0}^{t-1}\eta_j\left\|\nabla R(\boldsymbol{W}_j)\right\|_*+c_{13}\right)}\right) \\ &\quad+\frac{2c_{13}}{\sum_{j=0}^{t-1}\eta_j\left\|\nabla R(\boldsymbol{W}_j)\right\|_*-(c_{11}+c_{12})/\gamma_\star}\to 0.\end{aligned} \tag{44}$$

Under the power-type lower bound in Lemma F.5,

$$\left\|\frac{\boldsymbol{U}_t}{\|\boldsymbol{U}_t\|}-\boldsymbol{U}_\star\right\|\le\left(\frac{c_{15}}{2\gamma_\star c_\omega\left(\sum_{j=0}^{t-1}\eta_j\left\|\nabla R(\boldsymbol{W}_j)\right\|_*+c_{13}\right)}\right)^{1/q}.$$

Together with Equations 39 and 44, this gives the claimed $O((\log(\sum_{j<t}\eta_j))^{-1/q})$ rate for both errors, since $q\ge 2$. □

G COUNTEREXAMPLE WITH CONSTANT STEPS

G.1 SETUP

Write $\ell(s) = \log(1 + e^{-s})$ for the scalar logistic loss. Let

$$\mathcal{D}_{\text{const}} = \{(\boldsymbol{e}_1/20, +1),\ ((\boldsymbol{e}_1 + \boldsymbol{e}_2)/20, +1),\ (\boldsymbol{e}_3, +1),\ (\boldsymbol{e}_3, -1)\}. \tag{45}$$

The conflicting labels at $\boldsymbol{e}_3$ make these data nonseparable. For $\boldsymbol{w} = (x, y, z)^\top$, their risk is

$$R(\boldsymbol{w}) = \frac{1}{4}\left[\ell(x/20) + \ell((x + y)/20) + \ell(z) + \ell(-z)\right]. \tag{46}$$

Define the boundary vectors by

$$\boldsymbol{b}(a) = (1 - \psi(a), 1 + a + \psi(a), a)^\top,$$

where $\psi(a) = \frac{1}{2}(\max\{|a| - 1, 0\})^2$. Define the set $\mathbb{K} = \operatorname{conv}\{\pm\boldsymbol{b}(a) : -2 \le a \le 2\}$ and its Minkowski functional

$$\|\boldsymbol{w}\|_{\mathbb{K}} = \inf\{r > 0 : \boldsymbol{w} \in r\mathbb{K}\}. \tag{47}$$

The generating set $\{\pm\boldsymbol{b}(a) : -2 \le a \le 2\}$ is compact because $\boldsymbol{b}$ is continuous. Its convex hull $\mathbb{K}$ is compact in $\mathbb{R}^3$, convex, and balanced. $\mathbb{K}$ is absorbing as

$$\begin{pmatrix} | & | & | \\ \boldsymbol{b}(0) & \boldsymbol{b}(1) & \boldsymbol{b}(2) \\ | & | & | \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1/2 \\ 1 & 2 & 7/2 \\ 0 & 1 & 2 \end{pmatrix}$$

has a determinant of -1. Thus, for every $\boldsymbol{w} \neq \boldsymbol{0}$, there exist scalars $\alpha_0, \alpha_1, \alpha_2$ such that $\boldsymbol{w} = \alpha_0\boldsymbol{b}(0) + \alpha_1\boldsymbol{b}(1) + \alpha_2\boldsymbol{b}(2)$. Then, for every $r \ge \sum_{m=0}^{2} |\alpha_m|$,

$$\frac{\boldsymbol{w}}{r} = \sum_{m=0}^{2} \frac{|\alpha_m|}{r} \operatorname{sgn}(\alpha_m)\boldsymbol{b}(m) + \left(1 - \frac{\sum_{m=0}^{2} |\alpha_m|}{r}\right)\boldsymbol{0} \in \mathbb{K}.$$

Note that $\boldsymbol{0} \in \mathbb{K}$ by definition. Hence, $\|\boldsymbol{w}\|_{\mathbb{K}}$ is a seminorm (Rudin, 1991); compactness further implies that it is a norm. Its closed unit ball is $\mathbb{K}$.

The following identities describe the available movements for ($|a| \le 2$):

$$b_1(a)+b_2(a)-b_3(a) = 2, \qquad \frac{1}{2} \le b_1(a) \le 1, \qquad 0 \le b_1(a)+b_2(a) \le 4, \qquad |b_3(a)| \le 2. \tag{48}$$

Equation 48 implies that $b_2(a) \ge b_3(a)$. In particular, $(x + y - z)/2$ equals one on every $\boldsymbol{b}(a)$ and is at most one on $\mathbb{K}$. Scaling a vector to the unit ball gives

$$\frac{x + y - z}{2} \le \left\|(x, y, z)^\top\right\|_{\mathbb{K}}, \qquad \|\boldsymbol{b}(a)\|_{\mathbb{K}} = 1. \tag{49}$$

The points of $\mathbb{K}$ with first coordinate one are exactly

$$\mathbb{K} \cap \{x = 1\} = \{(1, 1 + a, a)^\top : |a| \le 1\}. \tag{50}$$

$\boldsymbol{x}_i$'s span $\mathbb{R}^3$, thus $\mathcal{S} = \{\boldsymbol{0}\}$ and Proposition B.1 gives $\mathcal{Z} = \mathbb{R}^3$. Using the recession-cone definition in Section 3 and the decomposition in Proposition B.3 yields

$$\mathcal{L} = \{\boldsymbol{0}\}, \quad \mathcal{C} = \{(x, y, 0)^\top : x \ge 0,\ x + y \ge 0\}, \quad \mathcal{M} = \mathbb{R}^2 \times \{\boldsymbol{0}\}, \quad \mathcal{E} = \operatorname{span}\{\boldsymbol{e}_3\}. \tag{51}$$

The third coordinate must be zero in every recession direction because the two conflicting examples require both $z \ge 0$ and $-z \ge 0$. Write $\boldsymbol{w}_t = \boldsymbol{u}_t + \boldsymbol{v}_t$ where $\boldsymbol{u}_t = (x_t, y_t, 0)^\top, \boldsymbol{v}_t = z_t\boldsymbol{e}_3$. For a unit vector in $\mathcal{M}$,

$$\frac{\min\{x, x + y\}}{20} \le \frac{x}{20} \le \frac{1}{20}.$$

Equality is attained at $(1, 1, 0)^\top$. By Equation 50, this is the only unit vector in $\mathcal{M}$ with $x = 1$. Hence

$$\gamma_\star = \frac{1}{20}, \qquad \boldsymbol{u}_\star = (1, 1, 0)^\top. \tag{52}$$

For $\|\boldsymbol{d}\|_{\mathbb{K}} \le 1$, Equation 48 gives $|d_1| \le 1$, $|d_1 + d_2| \le 4$, and $|d_3| \le 2$. Thus the four absolute sample scores at $\boldsymbol{d}$ are at most $1/20$, $4/20$, 2, and 2. The binary version of Lemma A.2, equivalently $\ell'' \le 1/4$ applied to Equation 46, gives

$$\left\langle \boldsymbol{d}, \nabla^2 R(\boldsymbol{w})\boldsymbol{d} \right\rangle \le \beta \|\boldsymbol{d}\|_{\mathbb{K}}^2, \qquad \beta := \frac{1}{4} \cdot \frac{1}{N} \left(\left(\frac{1}{20}\right)^2 + \left(\frac{4}{20}\right)^2 + 2 \cdot 2^2 \right) = \frac{3217}{6400} < 1. \quad (53)$$

The steepest descent is with respect to $\|\cdot\|_{\mathbb{K}}$ with the tie-breaking rule that chooses the direction of minimum Euclidean norm:

$$\begin{aligned} \boldsymbol{s}_t &= \operatorname*{arg\,min}_{\boldsymbol{s} \in \mathcal{F}_{\mathbb{K}}(\nabla R(\boldsymbol{w}_t))} \|\boldsymbol{s}\|_2, \\ \boldsymbol{w}_{t+1} &= \boldsymbol{w}_t - \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} \boldsymbol{s}_t, \qquad \boldsymbol{w}_0 = 0. \end{aligned} \quad (54)$$

Proposition G.1. *For every constant $0 < \eta \le 1$, the iterates of Equation 54 satisfy*

$$\gamma_\star - \limsup_{t \to \infty} \frac{\min_{i \in [2]} \boldsymbol{x}_i^\top \boldsymbol{u}_t}{\|\boldsymbol{u}_t\|_{\mathbb{K}}} > 0. \quad (55)$$

Moreover, $\|\boldsymbol{u}_t\|_{\mathbb{K}} \to \infty$ and $z_t \to 0$.

By Equation 51, the reduced risk is

$$R_0(z\boldsymbol{e}_3) = \frac{1}{4}[\ell(z) + \ell(-z)].$$

It is minimized uniquely at $z = 0$. Thus Theorem 3.1 identifies $R_\star = \frac{1}{2}\log 2$. Since $0 < \eta \le 1 < \beta^{-1}$, the constant schedule satisfies Assumption 4.1. The risk theorem, Theorem 4.3 yields

$$R(\boldsymbol{w}_t) \to R_\star, \qquad z_t \to 0. \quad (56)$$

The two remaining loss terms are nonnegative, and

$$0 \le \ell(x_t/20) + \ell((x_t + y_t)/20) = 4(R(\boldsymbol{w}_t) - R_0(z_t\boldsymbol{e}_3)) \le 4(R(\boldsymbol{w}_t) - R_\star) \to 0.$$

Because $\ell(s)$ tends to zero only when $s \to \infty$, both growing margins diverge:

$$x_t \to \infty, \qquad x_t + y_t \to \infty. \quad (57)$$

What remains to prove is that their growth is too slow relative to the parameter norm to attain Equation 52.

G.2 Steepest direction and update control

Define:

$$\begin{aligned} q_t &= \frac{1}{80(1 + e^{x_t/20})}, \quad p_t = \frac{1}{80(1 + e^{(x_t+y_t)/20})}, \\ h(z) &= \frac{e^z - 1}{4(e^z + 1)}, \qquad r_t = \frac{h(z_t) - p_t}{q_t}. \end{aligned} \quad (58)$$

Here q_t and p_t come from the first and second examples; $h(z_t)$ is the derivative contributed by the conflicting pair. The ratio r_t measures the imbalance that determines the choice of the third-coordinate movement. Its role follows from differentiation:

$$\begin{aligned} -\nabla R(\boldsymbol{w}_t) &= (q_t + p_t, p_t, -h(z_t))^\top, \\ \langle -\nabla R(\boldsymbol{w}_t), \boldsymbol{b}(a) \rangle &= q_t + 2p_t - q_t(\psi(a) + r_t a). \end{aligned} \quad (59)$$

Since $p_t, q_t > 0$, the gradient never vanishes. Write

$$L_t := \sum_{j<t} \eta \|\nabla R(\boldsymbol{w}_j)\|_{\mathbb{K},*}. \quad (60)$$

Lemma G.2. *If $|r_t| \le 1$, then*

$$\begin{aligned} \boldsymbol{s}_t = -\boldsymbol{b}(a_t), \qquad a_t = \begin{cases} 1 - r_t, & r_t < 0, \\ -1/2, & r_t = 0, \\ -1 - r_t, & r_t > 0, \end{cases} \\ \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} = q_t + 2p_t + q_t(|r_t| + r_t^2/2). \end{aligned} \tag{61}$$

In particular,

$$|a_t| \ge \frac{1}{2}, \qquad b_1(a_t) = 1 - \frac{1}{2} r_t^2, \qquad \boldsymbol{w}_{t+1} - \boldsymbol{w}_t = \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} \boldsymbol{b}(a_t). \tag{62}$$

Proof. Because $\mathbb{K}$ is symmetric, identifying $-\boldsymbol{s}_t$ amounts to maximizing against $-\nabla R(\boldsymbol{w}_t)$. By Equation 59, it amounts to minimizing $\psi(a) + r_t a$. Its derivative is

$$\frac{\mathrm{d}}{\mathrm{d}a}(\psi(a) + r_t a) = \begin{cases} a + 1 + r_t, & -2 < a < -1, \\ r_t, & -1 \le a \le 1, \\ a - 1 + r_t, & 1 < a < 2. \end{cases}$$

For $r_t > 0$ the minimum is at $a = -1 - r_t$; for $r_t < 0$ it is at $a = 1 - r_t$. The condition $|r_t| \le 1$ keeps these points in $[-2, 2]$. When $r_t = 0$, every $a \in [-1, 1]$ minimizes the expression.

We must also compare $-\boldsymbol{b}(a)$. Since $\psi(a) + r_t a$ is convex, its maximum on $[-2, 2]$ occurs at an endpoint. Using $\psi(2) = \psi(-2) = 1/2$ gives

$$\begin{aligned} \max_{|a| \le 2} \langle -\nabla R(\boldsymbol{w}_t), \boldsymbol{b}(a) \rangle &= q_t + 2p_t + q_t(|r_t| + r_t^2/2), \\ \max_{|a| \le 2} \langle -\nabla R(\boldsymbol{w}_t), -\boldsymbol{b}(a) \rangle &= -q_t/2 - 2p_t + 2q_t|r_t|. \end{aligned}$$

The first exceeds the second by $q_t(1 + (1 - |r_t|)^2/2) + 4p_t > 0$, thus only $\boldsymbol{b}(a)$ can maximize the functional.

For $r_t \ne 0$, the maximizing generator is unique. For $r_t = 0$, the face $\mathcal{F}_{\mathbb{K}}(-\nabla R(\boldsymbol{w}_t))$ is the segment in Equation 50. On that segment,

$$\left\| (1, 1 + a, a)^\top \right\|_2^2 = 1 + (1 + a)^2 + a^2 = \frac{3}{2} + 2(a + 1/2)^2.$$

The tie-breaking rule thus gives $a = -1/2$. □

The two consequences in Equation 62 explain the mechanism. A step of length $\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}$ increases x_t by only $\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} (1 - r_t^2/2)$, whereas the unit max-margin direction increases it by $\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}$. Also, the third coordinate moves by $\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} a_t$ with $|a_t| \ge 1/2$, even when $r_t = 0$. We will use that unavoidable movement to show that the imbalance cannot remain arbitrarily small at consecutive iterations.

Lemma G.3 (Control of the imbalance and step lengths)**.** *For every $t \ge 0$, $|r_t| \le 1$. Moreover,*

$$\begin{aligned} 0 < p_t, q_t \le \frac{1}{160}, \qquad |z_t| < \frac{3}{20}, \qquad 0 < \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} \le \frac{9}{320}, \\ \frac{\|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}}{q_t} \le \frac{13}{2}, \qquad q_t + 2p_t \le \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} \le \frac{5}{2}(q_t + 2p_t). \end{aligned} \tag{63}$$

Proof. Initially $p_0 = q_0 = 1/160$ and $h(z_0) = 0$, thus $r_0 = -1$. Suppose $|r_j| \le 1$ for every $0 \le j \le t$. We will prove the same bound at $t + 1$.

Summing Equation 62 and using Equation 48 gives

$$\frac{1}{2} L_t \le x_t \le L_t, \qquad x_t + y_t \ge 0, \qquad y_t \ge z_t, \qquad x_t + y_t - z_t = 2L_t. \tag{64}$$

The nonnegative scores imply $p_t, q_t \le 1/160$ by Equation 58. Since $|r_t| \le 1$, $|h(z_t)| = |p_t + q_t r_t| \le 1/80$. To convert this bound on $h(z_t)$ into a bound on z_t, differentiate the expression in Equation 58:

$$h'(z) = \frac{e^z}{2(1+e^z)^2} = \frac{1}{8} - 2h(z)^2.$$

Thus $0 < h'(z) \le 1/8$ everywhere. Since $h(0) = 0$, the mean value theorem gives $|h(z)| \le |z|/8$. In particular,

$$\frac{1}{9} < \frac{63}{512} = \frac{1}{8} - 2\left(\frac{1}{32}\right)^2 \le h'(z) \le \frac{1}{8} \qquad (|z| \le 1/4). \tag{65}$$

The mean value theorem now gives $h(3/20) \ge (3/20)/9 = 1/60 > 1/80$. Since h is odd and strictly increasing, $|h(z_t)| \le 1/80$ forces $|z_t| < 3/20$.

The inequality $y_t \ge z_t > -3/20$ from Equation 64 then gives

$$\frac{p_t}{q_t} = \frac{1 + e^{x_t/20}}{1 + e^{(x_t+y_t)/20}} \le \max\{1, e^{-y_t/20}\} \le e^{3/400} < 2.$$

Substituting this ratio and $|r_t| \le 1$ into Equation 61 gives $\|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}/q_t \le 13/2$. Using $p_t, q_t \le 1/160$ gives $\|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} \le 9/320$. The remaining comparison in Equation 63 follows from $0 \le |r_t| + r_t^2/2 \le 3/2$. Since $|a_t| \le 2$ and $\eta \le 1$, the whole third-coordinate step stays inside the interval where Equation 65 applies:

$$|z_t| + |z_{t+1} - z_t| \le \frac{3}{20} + 2\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} \le \frac{3}{20} + \frac{18}{320} = \frac{33}{160} < \frac{1}{4}. \tag{66}$$

The bounds in Equation 48 imply

$$0 \le \frac{x_{t+1} - x_t}{20} \le \frac{\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}}{20}, \qquad 0 \le \frac{(x_{t+1} + y_{t+1}) - (x_t + y_t)}{20} \le \frac{\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}}{5}.$$

For any $v \in \mathbb{R}$ and $a \ge 0$,

$$e^{-a} \le \frac{1 + e^v}{1 + e^{v+a}} \le 1, \qquad 1 - e^{-a} \le a.$$

Applying these inequalities to the two weights in Equation 58 yields

$$e^{-\eta\|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}/20} q_t \le q_{t+1} \le q_t, \qquad 0 \le q_t - q_{t+1} \le \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}\, q_t/20, \tag{67}$$

$$e^{-\eta\|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}/5} p_t \le p_{t+1} \le p_t, \qquad 0 \le p_t - p_{t+1} \le \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}\, p_t/5. \tag{68}$$

These quantify the small change of the two contributions during one update.

By Equation 66 and the mean value theorem, there is a slope $\theta_t \in [1/9, 1/8]$ such that

$$\begin{aligned} h(z_{t+1}) - h(z_t) &= \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}\, a_t \theta_t, \\ q_{t+1} r_{t+1} &= q_t r_t + \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}\, a_t \theta_t + (p_t - p_{t+1}). \end{aligned} \tag{69}$$

For $r_t \ne 0$, Equation 61 gives

$$q_t r_t + \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}\, a_t \theta_t = \operatorname{sgn}(r_t) \left[(q_t - \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}\, \theta_t)|r_t| - \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}\, \theta_t \right].$$

Here $q_t - \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}\, \theta_t \ge q_t - 13q_t/16 = 3q_t/16 > 0$ by Equation 63. The bracketed expression is therefore an increasing linear function of $|r_t| \in [0, 1]$. Its values at the endpoints are $-\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}\, \theta_t$ and $q_t - 2\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}\, \theta_t$. If the second endpoint is negative, its absolute value is at most $\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}\, \theta_t$, because $q_t \ge \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}\, \theta_t$. Thus the absolute value at either endpoint is bounded as follows:

$$\begin{aligned} |q_t r_t + \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}\, a_t \theta_t| &\le \max\{\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}/8,\ q_t - 2\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}/9\} \\ &\le q_t - \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}/40. \end{aligned} \tag{70}$$

For the last inequality, $q_t \geq 2\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}/13 \geq (1/8 + 1/40)\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}$, and $2/9 \geq 1/40$. When $r_t = 0$, the choice $a_t = -1/2$ gives an absolute value at most $\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}/16$, thus Equation 70 also holds.

By Equation 63, $p_t/5 + q_t/20 < 1/40$. Combining Equations 67, 68, 69, and 70, we obtain

$$\begin{aligned}|q_{t+1}r_{t+1}| + (q_t - q_{t+1}) &\leq |q_t r_t + \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} a_t \theta_t| + (p_t - p_{t+1}) + (q_t - q_{t+1}) \\ &\leq q_t - \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} \left(\frac{1}{40} - \frac{p_t}{5} - \frac{q_t}{20}\right) < q_t.\end{aligned}$$

Subtracting $q_t - q_{t+1}$ gives $|q_{t+1}r_{t+1}| < q_{t+1}$, hence $|r_{t+1}| < 1$. This closes the induction and proves all the bounds. □

Proof of Proposition G.1. The risk and third-coordinate limits were established in Equation 56. By Lemma G.3, the step formula in Lemma G.2 is valid for every t.

By Equations 49 and 64,

$$L_t = \frac{x_t + y_t - z_t}{2} \leq \|\boldsymbol{w}_t\|_{\mathbb{K}} \leq \sum_{j<t} \|\boldsymbol{w}_{j+1} - \boldsymbol{w}_j\|_{\mathbb{K}} = L_t. \tag{71}$$

Thus the norm equals the accumulated distance exactly. Since $x_t \leq L_t$ by Equation 64, the limit $x_t \to \infty$ in Equation 57 gives $L_t \to \infty$. Thus,

$$\big|\|\boldsymbol{u}_t\|_{\mathbb{K}} - L_t\big| \leq |z_t| \, \|\boldsymbol{e}_3\|_{\mathbb{K}} \to 0, \qquad \frac{\|\boldsymbol{u}_t\|_{\mathbb{K}}}{L_t} \to 1. \tag{72}$$

Summing the first-coordinate update in Equation 62 gives

$$\frac{x_t}{L_t} = 1 - \frac{1}{2}\frac{\sum_{j<t} \eta \|\nabla R(\boldsymbol{w}_j)\|_{\mathbb{K},*} \, r_j^2}{L_t} \qquad (t \geq 1). \tag{73}$$

The fraction on the right is a weighted average of r_j^2. To obtain a margin deficit, it suffices to show that this average stays bounded away from zero.

The third coordinate always moves by at least $\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}/2$, by Equation 62. Its effect on h is at least one-ninth of that movement, by Equations 65 and 66. Consequently, Equations 68 and 69 give

$$\begin{aligned}|q_{t+1}r_{t+1} - q_t r_t| &\geq |h(z_{t+1}) - h(z_t)| - |p_{t+1} - p_t| \\ &\geq \eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} \left(\frac{1}{18} - \frac{p_t}{5}\right) \geq \frac{\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}}{20} \geq \frac{\eta q_t}{20}.\end{aligned}$$

Here $p_t \leq 1/160$ and $\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} \geq \eta q_t$ come from Equation 63. On the other hand, $q_{t+1} \leq q_t$ from Equation 67 gives

$$|q_{t+1}r_{t+1} - q_t r_t| \leq q_t(|r_t| + |r_{t+1}|).$$

Dividing by $q_t > 0$ yields

$$|r_t| + |r_{t+1}| \geq \frac{\eta}{20}, \qquad r_t^2 + r_{t+1}^2 \geq \frac{\eta^2}{800}. \tag{74}$$

Thus at least one of each two consecutive steps has a nonnegligible loss of first-coordinate progress.

The last bound in Equation 63, together with Equations 67 and 68, gives

$$\begin{aligned}\|\nabla R(\boldsymbol{w}_{t+1})\|_{\mathbb{K},*} &\geq (q_{t+1} + 2p_{t+1}) \geq e^{-\eta\|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}/5}(q_t + 2p_t) \\ &\geq \frac{2}{5}e^{-9/1600} \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} \geq \frac{1}{3} \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}, \\ \|\nabla R(\boldsymbol{w}_{t+1})\|_{\mathbb{K},*} &\leq \frac{5}{2}(q_{t+1} + 2p_{t+1}) \leq \frac{5}{2}(q_t + 2p_t) \leq \frac{5}{2} \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}.\end{aligned} \tag{75}$$

The lower bound uses $\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} \le 9/320$ and $e^{-9/1600} > 5/6$. $\|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}$ is bounded by three times $\|\nabla R(\boldsymbol{w}_{t+1})\|_{\mathbb{K},*}$ and vice versa. Hence $\min\{\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}, \eta \|\nabla R(\boldsymbol{w}_{t+1})\|_{\mathbb{K},*}\} \ge (\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} + \eta \|\nabla R(\boldsymbol{w}_{t+1})\|_{\mathbb{K},*})/4$. Combining this with Equation 74 gives

$$\begin{aligned} &\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} r_t^2 + \eta \|\nabla R(\boldsymbol{w}_{t+1})\|_{\mathbb{K},*} r_{t+1}^2 \\ &\ge \min\{\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*}, \eta \|\nabla R(\boldsymbol{w}_{t+1})\|_{\mathbb{K},*}\}(r_t^2 + r_{t+1}^2) \\ &\ge \frac{\eta^2}{3200}(\eta \|\nabla R(\boldsymbol{w}_t)\|_{\mathbb{K},*} + \eta \|\nabla R(\boldsymbol{w}_{t+1})\|_{\mathbb{K},*}). \end{aligned} \tag{76}$$

Apply Equation 76 to the disjoint pairs $(0,1),(2,3),\dots$ below time t. There is at most one unpaired step; its length is at most $9\eta/320$ by Equation 63. All terms $\eta \|\nabla R(\boldsymbol{w}_j)\|_{\mathbb{K},*} r_j^2$ are nonnegative, thus

$$\sum_{j<t} \eta \|\nabla R(\boldsymbol{w}_j)\|_{\mathbb{K},*} r_j^2 \ge \frac{\eta^2}{3200}\left(L_t - \frac{9\eta}{320}\right).$$

Since $L_t \to \infty$, this proves

$$\liminf_{t\to\infty} \frac{\sum_{j<t} \eta \|\nabla R(\boldsymbol{w}_j)\|_{\mathbb{K},*} r_j^2}{L_t} \ge \frac{\eta^2}{3200}. \tag{77}$$

The minimum margin over the two separable examples is

$$\min_{i\in[2]} \boldsymbol{x}_i^\top \boldsymbol{u}_t = \frac{\min\{x_t, x_t + y_t\}}{20} \le \frac{x_t}{20}.$$

Also, $x_t/L_t \in [1/2, 1]$ by Equation 64; since $\|\boldsymbol{u}_t\|_{\mathbb{K}}/L_t \to 1$, replacing the denominator L_t by $\|\boldsymbol{u}_t\|_{\mathbb{K}}$ changes this bounded ratio by a quantity tending to zero.

Finally, combining Equations 64,72,73, and 77, we obtain

$$\begin{aligned} \limsup_{t\to\infty} \frac{\min\{x_t, x_t + y_t\}}{20 \|\boldsymbol{u}_t\|_{\mathbb{K}}} &\le \frac{1}{20} \limsup_{t\to\infty} \frac{x_t}{L_t} \\ &\le \frac{1}{20}\left(1 - \frac{\eta^2}{6400}\right) < \gamma_\star. \end{aligned} \tag{78}$$

The deficit is positive for every fixed $\eta > 0$. □

H COUNTEREXAMPLE INSIDE LINEALITY

Lemma H.1 (Connected accumulation sets)**.** *Let $\{\boldsymbol{U}_t\}_t$ be a sequence in a compact metric space $(\mathcal{X}, d)$ such that $d(\boldsymbol{U}_t, \boldsymbol{U}_{t+1}) \to 0$. Then its accumulation set is nonempty, compact, and connected.*

Proof. Let Ω denote the accumulation set of $\{\boldsymbol{U}_t\}_t$. Compactness of $\mathcal{X}$ implies that Ω is nonempty, and Ω is closed in $\mathcal{X}$, hence compact.

Suppose, for contradiction, that Ω is disconnected. Then $\Omega = A \cup B$, where A and B are disjoint nonempty compact sets. Since A and B are compact and disjoint,

$$d(A, B) := \inf_{\boldsymbol{A}\in A, \boldsymbol{B}\in B} d(\boldsymbol{A}, \boldsymbol{B}) > 0.$$

Choose disjoint open neighborhoods $\mathcal{O}_A$ and $\mathcal{O}_B$ of A and B, respectively, such that

$$d(\mathcal{O}_A, \mathcal{O}_B) > 0.$$

There exists T such that

$$\boldsymbol{U}_t \in \mathcal{O}_A \cup \mathcal{O}_B \qquad \text{for every } t \ge T.$$

Since A and B are both nonempty subsets of the accumulation set, the sequence visits both $\mathcal{O}_A$ and $\mathcal{O}_B$ infinitely often. Hence there are infinitely many indices $t \ge T$ for which $\boldsymbol{U}_t$ and $\boldsymbol{U}_{t+1}$ belong to different neighborhoods. For each such t,

$$d(\boldsymbol{U}_t, \boldsymbol{U}_{t+1}) \ge d(\mathcal{O}_A, \mathcal{O}_B) > 0,$$

contradicting $d(\boldsymbol{U}_t, \boldsymbol{U}_{t+1}) \to 0$. Therefore Ω is connected. □

Proof of Proposition 6.1. Compactness of the unit sphere in $\mathcal{M}$ and Equation 10 place every accumulation point in $\mathcal{U}$. Since $\boldsymbol{U}_t = \boldsymbol{W}_t - \boldsymbol{V}_t$,

$$\|\boldsymbol{U}_{t+1} - \boldsymbol{U}_t\| \le \eta_t \|\nabla R(\boldsymbol{W}_t)\|_* + \|\boldsymbol{V}_{t+1} - \boldsymbol{V}_t\| \to 0.$$

Here smoothness gives $\|\nabla R(\boldsymbol{W}_t)\|_* \le \sqrt{2\beta\delta_t} \to 0$; $\eta_t \le \beta^{-1}$ and $\boldsymbol{V}_t$ converges. Apply Equation 43:

$$\left\| \frac{\boldsymbol{U}_{t+1}}{\|\boldsymbol{U}_{t+1}\|} - \frac{\boldsymbol{U}_t}{\|\boldsymbol{U}_t\|} \right\| \le \frac{2\,\|\boldsymbol{U}_{t+1} - \boldsymbol{U}_t\|}{\|\boldsymbol{U}_t\|} \to 0.$$

Recall that $\|\boldsymbol{U}_t\| \to \infty$; Lemma H.1 finishes the proof. □

We prove Proposition 6.2. The norming face permits both data-flat, zero-sum directions and common-class shifts. The selected trajectory uses only a fixed two-dimensional subspace of the former for rotation. We identify the full norming face before specifying that selector.

H.1 Setup

Take $K = 3$, $D = 4$, and

$$\mathcal{D} = \{(\boldsymbol{e}_1, 1), (2\boldsymbol{e}_1, 2), (\boldsymbol{e}_2, 3)\}.$$

Use the following Frobenius-orthogonal subspaces:

$$\begin{aligned}
\mathcal{N}_0 &:= \{\mathbf{1}\boldsymbol{a}^\top : \boldsymbol{a} \in \mathbb{R}^4\},\\
\mathcal{L}_0 &:= \{[\boldsymbol{w}_1^\top, \boldsymbol{w}_2^\top, \boldsymbol{w}_3^\top]^\top : \boldsymbol{w}_c \in \operatorname{span}\{\boldsymbol{e}_3, \boldsymbol{e}_4\},\ \textstyle\sum_{c=1}^3 \boldsymbol{w}_c = 0\},\\
\mathcal{H} &:= \{[\boldsymbol{w}_1^\top, \boldsymbol{w}_2^\top, \boldsymbol{w}_3^\top]^\top : \boldsymbol{w}_c \in \operatorname{span}\{\boldsymbol{e}_1, \boldsymbol{e}_2\},\ \textstyle\sum_{c=1}^3 \boldsymbol{w}_c = 0\}.
\end{aligned}$$

Equality of all score gaps to zero forces equality of the three scores on $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$. Consequently,

$$\mathcal{S} = \mathcal{N}_0 \oplus_F \mathcal{L}_0, \qquad \mathcal{H} = \mathcal{S}^\perp, \qquad \mathbb{R}^{3\times 4} = \mathcal{H} \oplus_F \mathcal{L}_0 \oplus_F \mathcal{N}_0,$$

with $\dim \mathcal{H} = 4$ and $\dim \mathcal{S} = 8$. Define

$$\|\boldsymbol{W}\|_\diamond := \max\left\{\|\Pi_{\mathcal{H}}\boldsymbol{W}\|_{\mathrm{F}},\ \|\Pi_{\mathcal{L}_0}\boldsymbol{W}\|_{\mathrm{F}} + \|\Pi_{\mathcal{N}_0}\boldsymbol{W}\|_{\mathrm{F}}\right\}. \tag{79}$$

The two entries of the maximum are seminorms on $\mathbb{R}^{3\times 4}$, hence their maximum is homogeneous and satisfies the triangle inequality. Moreover,

$$\|\boldsymbol{W}\|_\diamond = 0 \iff \Pi_{\mathcal{H}}\boldsymbol{W} = \Pi_{\mathcal{L}_0}\boldsymbol{W} = \Pi_{\mathcal{N}_0}\boldsymbol{W} = 0 \iff \boldsymbol{W} = 0.$$

Thus $\|\cdot\|_\diamond$ is a norm. For any nonzero $\boldsymbol{A} \in \mathcal{H}$, Cauchy–Schwarz yields

$$\langle \boldsymbol{A}, \boldsymbol{D}\rangle = \langle \boldsymbol{A}, \Pi_{\mathcal{H}}\boldsymbol{D}\rangle \le \|\boldsymbol{A}\|_{\mathrm{F}}\,\|\Pi_{\mathcal{H}}\boldsymbol{D}\|_{\mathrm{F}} \le \|\boldsymbol{A}\|_{\mathrm{F}} \qquad (\|\boldsymbol{D}\|_\diamond \le 1).$$

Equality holds if and only if $\Pi_{\mathcal{H}}\boldsymbol{D} = \boldsymbol{A}/\|\boldsymbol{A}\|_{\mathrm{F}}$. Hence,

$$\mathcal{F}_\diamond(\boldsymbol{A}) = \left\{ \frac{\boldsymbol{A}}{\|\boldsymbol{A}\|_{\mathrm{F}}} + \boldsymbol{B} + \boldsymbol{Q} : \boldsymbol{B} \in \mathcal{L}_0,\ \boldsymbol{Q} \in \mathcal{N}_0,\ \|\boldsymbol{B}\|_{\mathrm{F}} + \|\boldsymbol{Q}\|_{\mathrm{F}} \le 1 \right\}, \qquad \|\boldsymbol{A}\|_{\diamond,*} = \|\boldsymbol{A}\|_{\mathrm{F}}. \tag{80}$$

Differences of members of any one such face span $\mathcal{L}_0 \oplus \mathcal{N}_0$. As $\boldsymbol{A}$ ranges over $\mathcal{H}\setminus\{\mathbf{0}\}$, the members $\boldsymbol{A}/\|\boldsymbol{A}\|_{\mathrm{F}}$ span $\mathcal{H}$. Since $\mathcal{S}^\perp = \mathcal{H}$,

$$\mathcal{Z} = \mathbb{R}^{3\times 4}, \qquad \mathcal{L} = \mathcal{S}, \qquad \mathcal{N} = \{\mathbf{0}\}.$$

For every direction $\boldsymbol{D}$,

$$\left\langle \boldsymbol{D}, \nabla^2 R(\boldsymbol{W})\boldsymbol{D}\right\rangle \le \frac{1}{N}\sum_{i=1}^N \frac{1}{4}\max_{c,c'} \|\boldsymbol{e}_c - \boldsymbol{e}_{c'}\|_2^2\,\|\boldsymbol{x}_i\|_2^2\,\|\Pi_{\mathcal{H}}\boldsymbol{D}\|_{\mathrm{F}}^2 = \|\Pi_{\mathcal{H}}\boldsymbol{D}\|_{\mathrm{F}}^2 \le \|\boldsymbol{D}\|_\diamond^2.$$

Here $N = 3$ and $\sum_i \|\boldsymbol{x}_i\|_2^2 = 6$. Thus $\beta = 1$ is a valid smoothness constant. Fix $0 < \eta_0 \le 1$.

Nonvanishing gradient. The matrix

$$\widehat{\boldsymbol{U}} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ -2 & 2 & 0 & 0 \end{bmatrix} \in \mathcal{H}$$

has pairwise values

$$\Delta_{12}(\widehat{\boldsymbol{U}}) = \Delta_{21}(\widehat{\boldsymbol{U}}) = 0,$$
$$\Delta_{13}(\widehat{\boldsymbol{U}}) = -3, \qquad \Delta_{23}(\widehat{\boldsymbol{U}}) = -6, \qquad \Delta_{31}(\widehat{\boldsymbol{U}}) = \Delta_{32}(\widehat{\boldsymbol{U}}) = -3.$$

For every $\boldsymbol{U} \in \mathcal{C}$, the first two samples also imply

$$\Delta_{12}(\boldsymbol{U}) \le 0, \qquad \Delta_{21}(\boldsymbol{U}) = -2\Delta_{12}(\boldsymbol{U}) \le 0 \quad \Longrightarrow \quad \Delta_{12}(\boldsymbol{U}) = \Delta_{21}(\boldsymbol{U}) = 0.$$

Thus, $\mathcal{I}_\star = \{(1,2),(2,1)\}$ and $\mathcal{J}_\star = \{(1,3),(2,3),(3,1),(3,2)\} \neq \varnothing$. Theorem 3.1 implies that R has no finite minimizer. Since R is convex and differentiable, $\nabla R(\boldsymbol{W}) \neq 0$ for every finite $\boldsymbol{W}$.

Rotating tie-breaking rule. Choose the Frobenius-orthonormal matrices

$$\boldsymbol{L}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \boldsymbol{L}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

In addition, let $\mathcal{P} := \operatorname{span}\{\boldsymbol{L}_1, \boldsymbol{L}_2\} \subset \mathcal{L}_0$ and define the linear map $\mathcal{Q} : \mathcal{P} \to \mathcal{P}$ by

$$\mathcal{Q}\boldsymbol{L}_1 = \boldsymbol{L}_2, \qquad \mathcal{Q}\boldsymbol{L}_2 = -\boldsymbol{L}_1.$$

For every $\boldsymbol{L} \in \mathcal{P}$,

$$\langle \boldsymbol{L}, \mathcal{Q}\boldsymbol{L} \rangle = 0, \qquad \|\mathcal{Q}\boldsymbol{L}\|_{\mathrm{F}} = \|\boldsymbol{L}\|_{\mathrm{F}}.$$

We now describe the tie-breaking rule. For simplicity, write $\boldsymbol{W}_t^\ell := \Pi_{\mathcal{L}} \boldsymbol{W}_t$. At the origin, set

$$-\boldsymbol{S}_0 = -\frac{\nabla R(\boldsymbol{0})}{\|\nabla R(\boldsymbol{0})\|_{\mathrm{F}}} + \frac{\boldsymbol{L}_1 + \boldsymbol{L}_2}{\sqrt{2}}.$$

For $t \ge 1$, set

$$-\boldsymbol{S}_t = -\frac{\nabla R(\boldsymbol{W}_t)}{\|\nabla R(\boldsymbol{W}_t)\|_{\mathrm{F}}} + \frac{\boldsymbol{W}_t^\ell + \mathcal{Q}\boldsymbol{W}_t^\ell}{\sqrt{2}\,\|\boldsymbol{W}_t^\ell\|_{\mathrm{F}}}, \qquad \eta_t = \frac{\eta_0}{t+1}. \tag{81}$$

Then

$$\boldsymbol{W}_1^\ell = \eta_0 \|\nabla R(\boldsymbol{0})\|_{\mathrm{F}} \frac{\boldsymbol{L}_1 + \boldsymbol{L}_2}{\sqrt{2}} \in \mathcal{P}, \qquad \|\boldsymbol{W}_1^\ell\|_{\mathrm{F}} = \eta_0 \|\nabla R(\boldsymbol{0})\|_{\mathrm{F}} > 0, \qquad \Pi_{\mathcal{N}_0} \boldsymbol{W}_1 = \boldsymbol{0}.$$

Suppose $\boldsymbol{W}_t^\ell \in \mathcal{P} \setminus \{\boldsymbol{0}\}$ and $\Pi_{\mathcal{N}_0} \boldsymbol{W}_t = \boldsymbol{0}$. Then $\mathcal{Q}\boldsymbol{W}_t^\ell$ is well-defined and

$$\frac{\boldsymbol{W}_t^\ell + \mathcal{Q}\boldsymbol{W}_t^\ell}{\sqrt{2}\,\|\boldsymbol{W}_t^\ell\|_{\mathrm{F}}} \in \mathcal{P}, \qquad \left\| \frac{\boldsymbol{W}_t^\ell + \mathcal{Q}\boldsymbol{W}_t^\ell}{\sqrt{2}\,\|\boldsymbol{W}_t^\ell\|_{\mathrm{F}}} \right\|_{\mathrm{F}}^2 = \frac{\|\boldsymbol{W}_t^\ell\|_{\mathrm{F}}^2 + \|\mathcal{Q}\boldsymbol{W}_t^\ell\|_{\mathrm{F}}^2}{2\,\|\boldsymbol{W}_t^\ell\|_{\mathrm{F}}^2} = 1.$$

The first term of $\boldsymbol{S}_t$ belongs to $\mathcal{H}$, and the second term belongs to $\mathcal{L}_0$ with Frobenius norm one. By Equation 80,

$$\|\boldsymbol{S}_t\|_\diamond = 1, \qquad \langle \nabla R(\boldsymbol{W}_t), \boldsymbol{S}_t \rangle = \|\nabla R(\boldsymbol{W}_t)\|_{\mathrm{F}} = \|\nabla R(\boldsymbol{W}_t)\|_{\diamond,*}.$$

The update has zero $\mathcal{N}_0$-component. Its lineality projection is

$$\boldsymbol{W}_{t+1}^\ell = \boldsymbol{W}_t^\ell + \frac{\eta_t \|\nabla R(\boldsymbol{W}_t)\|_{\mathrm{F}}}{\sqrt{2}} \left(\frac{\boldsymbol{W}_t^\ell}{\|\boldsymbol{W}_t^\ell\|_{\mathrm{F}}} + \mathcal{Q}\frac{\boldsymbol{W}_t^\ell}{\|\boldsymbol{W}_t^\ell\|_{\mathrm{F}}} \right) \in \mathcal{P},$$

and its squared norm satisfies

$$\|\boldsymbol{W}_{t+1}^\ell\|_{\mathrm{F}}^2 = \left(\|\boldsymbol{W}_t^\ell\|_{\mathrm{F}} + \frac{\eta_t \|\nabla R(\boldsymbol{W}_t)\|_{\mathrm{F}}}{\sqrt{2}} \right)^2 + \frac{\eta_t^2 \|\nabla R(\boldsymbol{W}_t)\|_{\mathrm{F}}^2}{2} > 0. \tag{82}$$

Thus induction gives

$$\boldsymbol{W}_t^\ell \in \mathcal{P} \setminus \{\boldsymbol{0}\} \quad (t \ge 1), \qquad \Pi_{\mathcal{N}_0} \boldsymbol{W}_t = \boldsymbol{0} \quad (t \ge 0).$$

Therefore, the steepest descent update in Equation 81 is well-defined.

H.2 RADIAL GROWTH AND ROTATION

Since $R(\boldsymbol{W}) = R(\Pi_{\mathcal{H}}\boldsymbol{W})$,

$$\Pi_{\mathcal{H}}\boldsymbol{W}_{t+1} = \Pi_{\mathcal{H}}\boldsymbol{W}_t - \eta_t \nabla R(\Pi_{\mathcal{H}}\boldsymbol{W}_t), \qquad \Delta_{ic}(\boldsymbol{W}_t) = \Delta_{ic}(\Pi_{\mathcal{H}}\boldsymbol{W}_t). \tag{83}$$

These are exactly Frobenius GD and its score gaps.

If $\sum_{j\geq 0} \eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_{\mathrm{F}}$ were finite, Equation 83 would make $\{\Pi_{\mathcal{H}}\boldsymbol{W}_t\}_t$ Cauchy, with a finite limit $\boldsymbol{Y} \in \mathcal{H}$. Then

$$\left\|\nabla R(\boldsymbol{W}_t)\right\|_{\mathrm{F}} = \left\|\nabla R(\Pi_{\mathcal{H}}\boldsymbol{W}_t)\right\|_{\mathrm{F}} \to \left\|\nabla R(\boldsymbol{Y})\right\|_{\mathrm{F}} > 0.$$

But this contradicts $\sum_t \eta_t = \infty$. Hence

$$\sum_{j=0}^{t-1} \eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_{\mathrm{F}} \to \infty. \tag{84}$$

Radial growth. Equation 82 gives, for $t \geq 1$,

$$\frac{\eta_t \left\|\nabla R(\boldsymbol{W}_t)\right\|_{\mathrm{F}}}{\sqrt{2}} \leq \left\|\boldsymbol{W}_{t+1}^{\ell}\right\|_{\mathrm{F}} - \left\|\boldsymbol{W}_t^{\ell}\right\|_{\mathrm{F}} \leq \eta_t \left\|\nabla R(\boldsymbol{W}_t)\right\|_{\mathrm{F}}.$$

Since $\left\|\boldsymbol{W}_1^{\ell}\right\|_{\mathrm{F}} = \eta_0 \left\|\nabla R(\boldsymbol{0})\right\|_{\mathrm{F}}$, induction yields

$$\frac{1}{\sqrt{2}} \sum_{j=0}^{t-1} \eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_{\mathrm{F}} \leq \left\|\boldsymbol{W}_t^{\ell}\right\|_{\mathrm{F}} \leq \sum_{j=0}^{t-1} \eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_{\mathrm{F}}. \tag{85}$$

In particular, $\left\|\boldsymbol{W}_t^{\ell}\right\|_{\mathrm{F}} \to \infty$.

Angular growth. The gradient is globally bounded from Lemma A.1:

$$\left\|\nabla R(\boldsymbol{W})\right\|_{\mathrm{F}} \leq \frac{\sqrt{2}}{3} \sum_{i=1}^{3} \left\|\boldsymbol{x}_i\right\|_2 = \frac{4\sqrt{2}}{3}, \qquad \eta_t \left\|\nabla R(\boldsymbol{W}_t)\right\|_{\mathrm{F}} \to 0.$$

We write

$$\frac{\boldsymbol{W}_t^{\ell}}{\left\|\boldsymbol{W}_t^{\ell}\right\|_{\mathrm{F}}} = \cos\theta_t \, \boldsymbol{L}_1 + \sin\theta_t \, \boldsymbol{L}_2$$

with $\theta_t \in \mathbb{R}$ with $\theta_1 = \pi/4$; we allow the angle θ_t to be outside $[0, 2\pi]$. Then

$$\theta_{t+1} - \theta_t = \arctan\left(\frac{\eta_t \left\|\nabla R(\boldsymbol{W}_t)\right\|_{\mathrm{F}}}{\sqrt{2}\left\|\boldsymbol{W}_t^{\ell}\right\|_{\mathrm{F}} + \eta_t \left\|\nabla R(\boldsymbol{W}_t)\right\|_{\mathrm{F}}}\right) > 0. \tag{86}$$

Recall that $\frac{x}{2} \leq \arctan x \leq x$ for $x \in [0, 1]$. From Equation 85, for sufficiently large t,

$$\theta_{t+1} - \theta_t \geq \frac{1}{2\sqrt{2}} \frac{\eta_t \left\|\nabla R(\boldsymbol{W}_t)\right\|_{\mathrm{F}}}{\sum_{j=0}^{t} \eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_{\mathrm{F}}} \geq \frac{1}{4\sqrt{2}} \log \frac{\sum_{j=0}^{t} \eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_{\mathrm{F}}}{\sum_{j=0}^{t-1} \eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_{\mathrm{F}}}.$$

The second inequality holds as $\frac{\eta_t \left\|\nabla R(\boldsymbol{W}_t)\right\|_{\mathrm{F}}}{\sum_{j=0}^{t-1} \eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_{\mathrm{F}}} \to 0$ and $\frac{x}{1+x} \geq \frac{1}{2}\log(1+x)$ for $x \in [0, 1]$. Hence, Equation 84 implies that $\theta_t \to \infty$. On the other hand,

$$0 \leq \theta_{t+1} - \theta_t \leq \frac{\eta_t \left\|\nabla R(\boldsymbol{W}_t)\right\|_{\mathrm{F}}}{\sum_{j=0}^{t-1} \eta_j \left\|\nabla R(\boldsymbol{W}_j)\right\|_{\mathrm{F}} + \eta_t \left\|\nabla R(\boldsymbol{W}_t)\right\|_{\mathrm{F}}} \to 0.$$

Fix any $\varphi \in [0, 2\pi)$. For each sufficiently large integer j, define the first crossing time

$$t_j := \min\{t : \theta_t \geq 2\pi j + \varphi\}$$

Then $0 \leq \theta_{t_j} - (2\pi j + \varphi) < \theta_{t_j} - \theta_{t_j - 1}$ and thus $\theta_{t_j} - (2\pi j + \varphi) \to 0$. This implies that

$$\frac{\boldsymbol{W}_{t_j}^{\ell}}{\left\|\boldsymbol{W}_{t_j}^{\ell}\right\|_{\mathrm{F}}} \to \cos\varphi \, \boldsymbol{L}_1 + \sin\varphi \, \boldsymbol{L}_2.$$

Thus every point of the circle is an accumulation point. Conversely, every normalized lineality iterate belongs to the closed circle. Therefore,

$$\left\{\text{accumulation points of } \frac{\boldsymbol{W}_t^{\ell}}{\left\|\boldsymbol{W}_t^{\ell}\right\|_{\mathrm{F}}}\right\} = \{\cos\varphi \, \boldsymbol{L}_1 + \sin\varphi \, \boldsymbol{L}_2 : \varphi \in [0, 2\pi)\}.$$

In particular, the lineality component does not converge in direction.

Directional nonconvergence of $\boldsymbol{W}_t$ and $\boldsymbol{U}_t$. Since $\boldsymbol{S}_t$ has norm one,

$$\|\boldsymbol{W}_t\|_\diamond \le \sum_{j=0}^{t-1} \eta_j \|\nabla R(\boldsymbol{W}_j)\|_{\mathrm{F}}, \qquad \frac{\|\Pi_{\mathcal{L}} \boldsymbol{W}_t\|_{\mathrm{F}}}{\|\boldsymbol{W}_t\|_\diamond} \ge \frac{1}{\sqrt{2}}.$$

If $\boldsymbol{W}_t / \|\boldsymbol{W}_t\|_\diamond$ converged, the projection of the limit onto $\mathcal{L}$ would be nonzero. Normalizing that projection would force $\boldsymbol{W}_t^\ell / \|\boldsymbol{W}_t^\ell\|_{\mathrm{F}}$ to converge, a contradiction.

Theorem 4.3 gives $\boldsymbol{V}_t \to \boldsymbol{V}_\star$. Since $\mathcal{N} = \{\boldsymbol{0}\}$, we have $\boldsymbol{W}_t = \boldsymbol{U}_t + \boldsymbol{V}_t$. Moreover,

$$\|\boldsymbol{W}_t\|_\diamond \ge \|\boldsymbol{W}_t^\ell\|_{\mathrm{F}} \longrightarrow \infty, \qquad \|\boldsymbol{U}_t\|_\diamond \ge \|\boldsymbol{W}_t\|_\diamond - \|\boldsymbol{V}_t\|_\diamond \longrightarrow \infty.$$

For all sufficiently large t,

$$\begin{aligned}\left\| \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|_\diamond} - \frac{\boldsymbol{U}_t}{\|\boldsymbol{U}_t\|_\diamond} \right\|_\diamond &\le \frac{\|\boldsymbol{V}_t\|_\diamond}{\|\boldsymbol{W}_t\|_\diamond} + \frac{|\|\boldsymbol{U}_t\|_\diamond - \|\boldsymbol{W}_t\|_\diamond|}{\|\boldsymbol{W}_t\|_\diamond} \\ &\le \frac{2\|\boldsymbol{V}_t\|_\diamond}{\|\boldsymbol{W}_t\|_\diamond} \longrightarrow 0.\end{aligned}$$

Convergence of $\boldsymbol{U}_t / \|\boldsymbol{U}_t\|_\diamond$ would therefore imply convergence of $\boldsymbol{W}_t / \|\boldsymbol{W}_t\|_\diamond$, a contradiction.

I Counterexample outside lineality

We prove Theorem 6.3. The parameter is $\boldsymbol{w} \in \mathbb{R}^3$, and its signed score on sample i is $y_i \boldsymbol{w}^\top \boldsymbol{x}_i$; the score is $-y_i \boldsymbol{w}^\top \boldsymbol{x}_i$.

I.1 Setup

Fix $m \ge 7$ and take

$$\mathcal{D}_{\mathrm{sep}} = \{(\boldsymbol{e}_1, +1), (2\boldsymbol{e}_1 + \boldsymbol{e}_2, +1), (m\boldsymbol{e}_1 + \boldsymbol{e}_3, +1), (-m\boldsymbol{e}_1 + \boldsymbol{e}_3, -1)\} \subset \mathbb{R}^3 \times \{-1, +1\}. \tag{87}$$

Both labels occur, and $\boldsymbol{e}_1$ strictly separates the samples with signed scores $1, 2, m, m$. The features span $\mathbb{R}^3$, thus $\mathcal{S} = \{\boldsymbol{0}\}$. For $\boldsymbol{w} = (x, y, z)^\top$, the empirical risk is

$$R_{\mathrm{sep}}(\boldsymbol{w}) = \frac{1}{4}\left[\log(1 + e^{-x}) + \log(1 + e^{-(2x+y)}) + \log(1 + e^{-(mx+z)}) + \log(1 + e^{-(mx-z)})\right]. \tag{88}$$

Define the thresholds and boundary vectors

$$\begin{aligned} &b_0 = 0, \quad b_1 = 10, \quad b_j = (1 + j^2) b_{j-1} \quad (j \ge 2), \\ &\boldsymbol{q}_j = \left(1 - \sum_{l=j}^{\infty} 2^{-l-1} e^{-b_l},\ 1 + 2^{-j},\ \frac{(-1)^j}{4}\right)^\top \quad (j \ge 1), \qquad \boldsymbol{q}_\pm = (1, 1, \pm 1/4)^\top. \end{aligned} \tag{89}$$

Then $\boldsymbol{q}_{2j} \to \boldsymbol{q}_+$ and $\boldsymbol{q}_{2j+1} \to \boldsymbol{q}_-$. Let $\mathbb{B}_2 := \{\boldsymbol{d} \in \mathbb{R}^3 : \|\boldsymbol{d}\|_2 \le 1\}$, and set

$$\mathbb{B} := \mathrm{cl}\,\mathrm{conv}\left(\frac{1}{10}\mathbb{B}_2 \cup \{\pm \boldsymbol{q}_j : j \ge 1\}\right). \tag{90}$$

Then $\pm\boldsymbol{q}_+, \pm\boldsymbol{q}_- \in \mathbb{B}$. Define its Minkowski functional by

$$\|\boldsymbol{w}\|_{\mathbb{B}} := \inf\{r > 0 : \boldsymbol{w} \in r\mathbb{B}\}, \qquad \boldsymbol{w} \in \mathbb{R}^3. \tag{91}$$

Every vector in $\mathbb{B}$ has bounded Euclidean norm, thus $\mathbb{B}$ is compact, convex, absorbing (as it contains a neighborhood of the origin), and balanced. Hence, Equation 91 is a seminorm (Rudin, 1991); compactness further implies that it is a norm. Its closed unit ball is $\mathbb{B}$. Its dual norm and norming face are

$$\|\boldsymbol{g}\|_{\mathbb{B},*} = \max_{\boldsymbol{d} \in \mathbb{B}} \langle \boldsymbol{g}, \boldsymbol{d} \rangle, \qquad \mathcal{F}_{\mathbb{B}}(\boldsymbol{g}) := \arg\max_{\boldsymbol{d} \in \mathbb{B}} \langle \boldsymbol{g}, \boldsymbol{d} \rangle \quad (\boldsymbol{g} \ne 0). \tag{92}$$

For $\boldsymbol{w} = (x, y, z)^\top$, the four scores are $-x, -(2x + y), -(mx + z), -mx + z$; if $\boldsymbol{w} \in \mathcal{C}$ then $x \geq 0, y \geq -2x, |z| \leq mx$ hold, and they strictly hold at $\boldsymbol{e}_1$. Thus,

$$\mathcal{Z} = \mathcal{M} = \mathbb{R}^3, \qquad \mathcal{N} = \mathcal{L} = \mathcal{E} = \{\mathbf{0}\}, \qquad \mathcal{I}_\star = \varnothing, \qquad \mathcal{J}_\star = \{(i, -y_i) : i \in [4]\}.$$

As the absolute value of the first coordinate of every element of $\mathbb{B}$ is at most 1, $\gamma_\star \leq 1$. However,

$$\mathbb{B} \cap \{\boldsymbol{d} : d_1 = 1\} = \operatorname{conv}\{\boldsymbol{q}_+, \boldsymbol{q}_-\} = \{(1, 1, s)^\top : |s| \leq 1/4\}. \tag{93}$$

Adding the limit points $\pm\boldsymbol{q}_+$ and $\pm\boldsymbol{q}_-$ of $\boldsymbol{q}_j$ to the generating set in Equation 90 makes that set compact; its convex hull is exactly $\mathbb{B}$. Every point in this enlarged generating set has first coordinate at most 1, with equality only at $\boldsymbol{q}_+$ and $\boldsymbol{q}_-$. A convex combination has first coordinate 1 only if every point with a positive coefficient has first coordinate 1. This proves Equation 93.

The vectors in the set of Equation 93 have scores $-y_i \boldsymbol{w}^\top \boldsymbol{x}_i$ of $-1, -3, -m - s, -m + s$. Hence

$$\gamma_{\mathrm{sep},\star} = 1, \qquad \mathcal{U}_{\mathrm{sep},\star} = \{(1, 1, s)^\top : |s| \leq 1/4\}. \tag{94}$$

The absolute values of four scores for $\|\boldsymbol{d}\|_{\mathbb{B}} \leq 1$ are bounded by $1, 7/2, m + 1/4, m + 1/4$, respectively. Since the second derivative of scalar logistic loss is at most $1/4$, we may take the global smoothness constant of the risk as

$$\beta_{\mathrm{sep}} := \frac{1}{4} \cdot \frac{1}{N} \left[1^2 + \left(\frac{7}{2}\right)^2 + 2\left(m + \frac{1}{4}\right)^2 \right]. \tag{95}$$

Fix $0 < \eta_0 < \beta_{\mathrm{sep}}^{-1}$ and $\eta_t = \eta_0/(t+1)$. These steps satisfy Assumption 5.1, thus Theorems 4.3 and 5.2 apply and give

$$\|\boldsymbol{w}_t\|_{\mathbb{B}} \to \infty, \qquad \sum_{j=0}^{t-1} \eta_j \|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_j)\|_{\mathbb{B},*} \to \infty. \tag{96}$$

By Proposition 6.1, the accumulation set is a nonempty compact connected subset of Equation 94. It remains to prove that both endpoints are in the accumulation set.

Consider maximizing $\langle (1, \varepsilon, \omega)^\top, \boldsymbol{d} \rangle$ over $\boldsymbol{d} \in \mathbb{B}$. For $l \neq j$, Equation 89 gives $\langle (1, \varepsilon, 0)^\top, \boldsymbol{q}_j - \boldsymbol{q}_l \rangle = \sum_{m=j}^{l-1} 2^{-m-1}(\varepsilon - e^{-b_m})$ for $l > j$ and $\sum_{m=l}^{j-1} 2^{-m-1}(e^{-b_m} - \varepsilon)$ for $l < j$. Both expressions are positive for $e^{-b_j} < \varepsilon < e^{-b_{j-1}}$. For $j \geq 2$, on the region

$$2e^{-b_j} \leq \varepsilon \leq \frac{1}{2} e^{-b_{j-1}}, \tag{97}$$

the gaps satisfy

$$\begin{aligned} \langle (1, \varepsilon, 0)^\top, \boldsymbol{q}_j - \boldsymbol{q}_l \rangle &\geq 2^{-j-1}(\varepsilon - e^{-b_j}) \geq 2^{-j-2}\varepsilon \qquad (l > j), \\ \langle (1, \varepsilon, 0)^\top, \boldsymbol{q}_j - \boldsymbol{q}_l \rangle &\geq 2^{-j}(e^{-b_{j-1}} - \varepsilon) \geq 2^{-j-1} e^{-b_{j-1}} \quad (l < j). \end{aligned} \tag{98}$$

Moreover, $\langle (1, \varepsilon, 0)^\top, \boldsymbol{q}_j - \boldsymbol{q}_\pm \rangle \geq 2^{-j}(\varepsilon - e^{-b_j}) > 0$. Set

$$c_{\mathrm{sep}} := 1 - \frac{1}{2} e^{-10} > 0.$$

Then $(\boldsymbol{q}_j)_1 \geq c_{\mathrm{sep}}$ for all j.

Lemma I.1. *If $\varepsilon \in (0, 1/2]$ and $|\omega| \leq 1/m$, then every maximizer of $\langle (1, \varepsilon, \omega)^\top, \boldsymbol{d} \rangle$ over $\boldsymbol{d} \in \mathbb{B}$ belongs to* $\operatorname{cl}\operatorname{conv}\{\boldsymbol{q}_j : j \geq 1\}$.

Proof. If $\varepsilon \in (0, 1/2]$ and $|\omega| \leq 1/m$, $\langle (1, \varepsilon, \omega)^\top, \boldsymbol{q}_j \rangle$ is at least $c_{\mathrm{sep}} - 1/(4m)$, while every $-\boldsymbol{q}_j$ has score at most $-c_{\mathrm{sep}} + 1/(4m)$. The small Euclidean ball has

$$\max_{\boldsymbol{d} \in \frac{1}{10}\mathbb{B}_2} \langle (1, \varepsilon, \omega)^\top, \boldsymbol{d} \rangle = \frac{1}{10}\sqrt{1 + \varepsilon^2 + \omega^2} \leq \frac{1}{10}\sqrt{1 + \frac{1}{4} + \frac{1}{m^2}} < c_{\mathrm{sep}} - \frac{1}{4m}.$$

Thus the maximizers belong to $\operatorname{cl}\operatorname{conv}\{\boldsymbol{q}_j : j \geq 1\}$. □

For each $j \geq 2$, taking $\omega = 0$ and $e^{-b_j} < \varepsilon < e^{-b_{j-1}}$ shows that $\boldsymbol{q}_j$ uniquely maximizes the functional. Thus it is exposed in $\mathbb{B}$ and $\|\boldsymbol{q}_j\|_{\mathbb{B}} = 1$; when $j = 1$, take $e^{-b_1} < \varepsilon < 1/2$. Hence, $\|\boldsymbol{q}_\pm\|_{\mathbb{B}} = 1$.

Changing ω from zero changes the difference of $\langle (1, \varepsilon, \omega)^\top, \boldsymbol{d} \rangle$ at any two vectors in $\{\boldsymbol{q}_j : j \geq 1\} \cup \{\boldsymbol{q}_+, \boldsymbol{q}_-\}$ by at most $|\omega|/2$. If also $|\omega| \leq 64\varepsilon^2$ and Equation 97 holds for some j, then

$$\frac{|\omega|}{2} \leq 32\varepsilon^2 \leq 16e^{-b_{j-1}}\varepsilon, \qquad \frac{|\omega|}{2} \leq 8e^{-2b_{j-1}}.$$

The first is strictly below $2^{-j-2}\varepsilon$ when $2^{j+6}e^{-b_{j-1}} < 1$; the second is strictly below $2^{-j-1}e^{-b_{j-1}}$ when $2^{j+4}e^{-b_{j-1}} < 1$. Thus, these are smaller than both gaps in Equation 98 whenever $2^{j+6}e^{-b_{j-1}} < 1$; this holds for all sufficiently large j, since $b_{j-1} \geq 10 \cdot 5^{j-2}$. Hence, for every sufficiently large j, if

$$2e^{-b_j} \leq \varepsilon \leq \tfrac{1}{2}e^{-b_{j-1}}, \qquad |\omega| \leq \min\{1/m, 64\varepsilon^2\} \tag{99}$$

then $\mathcal{F}_{\mathbb{B}}((1, \varepsilon, \omega)^\top) = \{\boldsymbol{q}_j\}$.

I.2 Growth of parameter

Using the same sign convention as Equation 2, write the steepest-descent update as

$$\boldsymbol{s}_t \in \mathcal{F}_{\mathbb{B}}(\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)), \qquad \boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\|_{\mathbb{B},*}\, \boldsymbol{s}_t, \qquad \boldsymbol{w}_0 = 0. \tag{100}$$

Write $\boldsymbol{w}_t = (x_t, y_t, z_t)^\top$. The gradient components satisfy

$$\begin{aligned}
-[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_1 &= \frac{1}{4}\left(\frac{1}{1+e^{x_t}} + \frac{2}{1+e^{2x_t+y_t}} + \frac{m}{1+e^{mx_t+z_t}} + \frac{m}{1+e^{mx_t-z_t}}\right), \\
-[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_2 &= \frac{1}{4(1+e^{2x_t+y_t})}, \\
-[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_3 &= \frac{1}{4}\left(\frac{1}{1+e^{mx_t+z_t}} - \frac{1}{1+e^{mx_t-z_t}}\right).
\end{aligned} \tag{101}$$

In particular, $-[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_1 > 0$, thus the gradient is never zero, and

$$0 < \frac{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_2}{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_1} \leq \frac{1}{2}, \qquad \left|\frac{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_3}{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_1}\right| \leq \frac{1}{m}. \tag{102}$$

By Lemma I.1,

$$-\boldsymbol{s}_t \in \mathrm{cl\,conv}\{\boldsymbol{q}_j : j \geq 1\}, \qquad c_{\mathrm{sep}} \leq (-\boldsymbol{s}_t)_1 \leq 1, \qquad 1 \leq (-\boldsymbol{s}_t)_2 \leq \frac{3}{2}, \qquad |(-\boldsymbol{s}_t)_3| \leq \frac{1}{4}.$$

Summing the update gives

$$\begin{aligned}
c_{\mathrm{sep}} \sum_{l=0}^{t-1} \eta_l \|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_l)\|_{\mathbb{B},*} \leq x_t \leq \sum_{l=0}^{t-1} \eta_l \|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_l)\|_{\mathbb{B},*}, \\
\sum_{l=0}^{t-1} \eta_l \|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_l)\|_{\mathbb{B},*} \leq y_t \leq \frac{3}{2} \sum_{l=0}^{t-1} \eta_l \|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_l)\|_{\mathbb{B},*}, \\
|z_t| \leq \frac{1}{4} \sum_{l=0}^{t-1} \eta_l \|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_l)\|_{\mathbb{B},*}.
\end{aligned} \tag{103}$$

The dual gradient norm is bounded:

$$\begin{aligned}
\|\nabla R_{\mathrm{sep}}(\boldsymbol{w})\|_{\mathbb{B},*} &\leq \sup_{\|\boldsymbol{d}\|_{\mathbb{B}} \leq 1} |\langle \nabla R_{\mathrm{sep}}(\boldsymbol{w}), \boldsymbol{d} \rangle| \leq \frac{1}{N} \sum_i \sup_{\|\boldsymbol{d}\|_{\mathbb{B}} \leq 1} |\boldsymbol{d}^\top \boldsymbol{x}_i| \\
&\leq \frac{1}{4}\left(1 + \frac{7}{2} + 2\left(m + \frac{1}{4}\right)\right) = \frac{2m+5}{4}.
\end{aligned}$$

Therefore

$$0 < (x_{t+1} + y_{t+1}) - (x_t + y_t) \le \frac{5}{2}\eta_t \left\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\right\|_{\mathbb{B},*} \to 0. \tag{104}$$

This shows that $x_t + y_t$ increases to infinity with vanishing increments. In addition,

$$(m-1)x_t \pm z_t \ge (m-1)x_t - |z_t| \ge \left((m-1)c_{\mathrm{sep}} - \frac{1}{4}\right) \sum_{l=0}^{t-1} \eta_l \left\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_l)\right\|_{\mathbb{B},*}.$$

Since $(m-1)c_{\mathrm{sep}} - 1/4 > 0$, Equation 96 shows that $(m-1)x_t \pm z_t$ tend to infinity.

I.3 Gradient ratios

Since $x_t, y_t \ge 0$,

$$\frac{2(1+e^{x_t})}{1+e^{2x_t+y_t}} \le 4e^{-x_t-y_t}, \qquad \frac{m(1+e^{x_t})}{1+e^{mx_t \pm z_t}} \le 2me^{-((m-1)x_t \pm z_t)}.$$

The right-hand sides tend to zero. Hence, for all sufficiently large t,

$$\frac{2}{1+e^{2x_t+y_t}} + \frac{m}{1+e^{mx_t+z_t}} + \frac{m}{1+e^{mx_t-z_t}} \le \frac{1}{1+e^{x_t}},$$

and

$$\frac{1}{4(1+e^{x_t})} \le -[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_1 \le \frac{2}{4(1+e^{x_t})}.$$

Using the second identity in Equation 101 and

$$\frac{1}{2}e^{-(x_t+y_t)} \le \frac{1+e^{x_t}}{1+e^{2x_t+y_t}} \le 2e^{-(x_t+y_t)},$$

we obtain

$$\left| -\log \frac{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_2}{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_1} - (x_t + y_t) \right| \le \log 4. \tag{105}$$

In addition,

$$\begin{aligned}\left|\frac{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_3}{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_1}\right| &\le (1+e^{x_t})\left(\frac{1}{1+e^{mx_t+z_t}} + \frac{1}{1+e^{mx_t-z_t}}\right) \\ &\le 2e^{x_t}e^{-mx_t}(e^{-z_t}+e^{z_t}) \le 4e^{-((m-1)x_t-|z_t|)}.\end{aligned}$$

The assumption $m \ge 7$ gives $(m-1)c_{\mathrm{sep}} - \frac{1}{4} > 5$. Consequently,

$$(m-1)x_t - |z_t| \ge \left((m-1)c_{\mathrm{sep}} - \frac{1}{4}\right) \sum_{l=0}^{t-1} \eta_l \left\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_l)\right\|_{\mathbb{B},*} \ge 2(x_t + y_t).$$

Combining these estimates with Equation 105 gives

$$\left|\frac{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_3}{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_1}\right| \le 64\left(\frac{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_2}{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_1}\right)^2 \tag{106}$$

for all sufficiently large t. Thus, for every sufficiently large j,

$$b_{j-1} + \log 8 \le x_t + y_t < b_j - \log 8 \tag{107}$$

satisfies Equation 99 with $\varepsilon = \frac{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_2}{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_1}$ and $\omega = \frac{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_3}{[\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)]_1}$. In particular,

$$\boldsymbol{s}_t = -\boldsymbol{q}_j$$

whenever Equation 107 holds and j is sufficiently large; the steepest direction is therefore unique.

I.4 Dominant updates and directional nonconvergence

Define the first-crossing times

$$t_j^- := \min\{t : x_t + y_t \geq b_{j-1} + \log 8\}, \qquad t_j^+ := \min\{t : x_t + y_t \geq b_j - \log 8\}.$$

Both are finite and diverge to infinity with j. Moreover, from Equation 104, for every sufficiently large j,

$$b_{j-1} + \log 8 \leq x_{t_j^-} + y_{t_j^-} < b_{j-1} + \log 8 + 1,$$
$$b_j - \log 8 \leq x_{t_j^+} + y_{t_j^+} < b_j - \log 8 + 1.$$

Hence $t_j^- < t_j^+$, and every direction of update with $t_j^- \leq t < t_j^+$ is $\boldsymbol{q}_j$. Since $(\boldsymbol{q}_j)_1 + (\boldsymbol{q}_j)_2 = 2 + 2^{-j} - \sum_{l=j}^{\infty} 2^{-l-1} e^{-b_l}$,

$$\frac{b_j - b_{j-1} - 2\log 8 - 1}{2 + 2^{-j} - \sum_{l=j}^{\infty} 2^{-l-1} e^{-b_l}} \leq \sum_{t=t_j^-}^{t_j^+ - 1} \eta_t \left\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\right\|_{\mathbb{B},*} \leq \frac{b_j - b_{j-1} - 2\log 8 + 1}{2 + 2^{-j} - \sum_{l=j}^{\infty} 2^{-l-1} e^{-b_l}}. \quad (108)$$

Meanwhile, Equation 103 implies

$$\sum_{t=0}^{t_j^- - 1} \eta_t \left\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\right\|_{\mathbb{B},*} \leq \frac{b_{j-1} + \log 8 + 1}{1 + c_{\mathrm{sep}}}.$$

Using $b_j - b_{j-1} = j^2 b_{j-1}$ and $2 + 2^{-j} - \sum_{l=j}^{\infty} 2^{-l-1} e^{-b_l} \leq 5/2$ gives

$$\frac{\sum_{t=0}^{t_j^- - 1} \eta_t \left\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\right\|_{\mathbb{B},*}}{\sum_{t=t_j^-}^{t_j^+ - 1} \eta_t \left\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\right\|_{\mathbb{B},*}} \leq \frac{5(b_{j-1} + \log 8 + 1)}{2(1 + c_{\mathrm{sep}})(j^2 b_{j-1} - 2\log 8 - 1)} \to 0. \quad (109)$$

We have that

$$\boldsymbol{w}_{t_j^+} = \boldsymbol{w}_{t_j^-} + \left(\sum_{t=t_j^-}^{t_j^+ - 1} \eta_t \left\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\right\|_{\mathbb{B},*}\right) \boldsymbol{q}_j, \qquad \left\|\boldsymbol{w}_{t_j^-}\right\|_{\mathbb{B}} \leq \sum_{t=0}^{t_j^- - 1} \eta_t \left\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\right\|_{\mathbb{B},*}.$$

Also, $\|\boldsymbol{q}_j\|_{\mathbb{B}} = 1$, thus we have

$$\left\| \frac{\boldsymbol{w}_{t_j^+}}{\sum_{t=0}^{t_j^+ - 1} \eta_t \left\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\right\|_{\mathbb{B},*}} - \boldsymbol{q}_j \right\|_{\mathbb{B}} \leq 2 \frac{\sum_{t=0}^{t_j^- - 1} \eta_t \left\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\right\|_{\mathbb{B},*}}{\sum_{t=0}^{t_j^+ - 1} \eta_t \left\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\right\|_{\mathbb{B},*}} \to 0. \quad (110)$$

Equation 110 also gives

$$\frac{\left\|\boldsymbol{w}_{t_j^+}\right\|_{\mathbb{B}}}{\sum_{t=0}^{t_j^+ - 1} \eta_t \left\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\right\|_{\mathbb{B},*}} \to 1.$$

Therefore,

$$\frac{\boldsymbol{w}_{t_{2j}^+}}{\left\|\boldsymbol{w}_{t_{2j}^+}\right\|_{\mathbb{B}}} \to \boldsymbol{q}_+, \qquad \frac{\boldsymbol{w}_{t_{2j+1}^+}}{\left\|\boldsymbol{w}_{t_{2j+1}^+}\right\|_{\mathbb{B}}} \to \boldsymbol{q}_-. \quad (111)$$

This proves directional nonconvergence. By Proposition 6.1, the accumulation set is connected and contained in the max-margin segment Equation 94. Since it contains both endpoints, it equals the entire segment.

J Nonseparable counterexample outside lineality

We prove Corollary 6.4 by appending one coordinate and a conflicting pair.

J.1 SETUP

Embed the four inputs of Equation 87 in $\mathbb{R}^4$ by appending zero, and add $(\boldsymbol{e}_4, +1)$ and $(\boldsymbol{e}_4, -1)$. Thus

$$\mathcal{D}_{\mathrm{ns}} = \{(\boldsymbol{e}_1, +1), (2\boldsymbol{e}_1 + \boldsymbol{e}_2, +1), (m\boldsymbol{e}_1 + \boldsymbol{e}_3, +1), (-m\boldsymbol{e}_1 + \boldsymbol{e}_3, -1), (\boldsymbol{e}_4, +1), (\boldsymbol{e}_4, -1)\},$$

a subset of $\mathbb{R}^4 \times \{-1, +1\}$. The last two examples prevent strict separation. Define

$$\iota : \mathbb{R}^3 \to \mathbb{R}^4, \qquad \iota(\boldsymbol{w}) := (\boldsymbol{w}^\top, 0)^\top.$$

With $\|\cdot\|_{\mathbb{B}}$ defined in Equation 91, use the norm

$$\|\boldsymbol{w}\|_{\mathrm{ns}} := \max\{\|\boldsymbol{w}_{1:3}\|_{\mathbb{B}}, |w_4|\}, \qquad \boldsymbol{w} = (w_1, w_2, w_3, w_4)^\top \in \mathbb{R}^4. \tag{112}$$

Its unit ball is $\mathbb{B} \times [-1, 1]$. The embedding is isometric for both the optimization norms and the Euclidean norms:

$$\|\iota(\boldsymbol{w})\|_{\mathrm{ns}} = \|\boldsymbol{w}\|_{\mathbb{B}}, \qquad \|\iota(\boldsymbol{w})\|_2 = \|\boldsymbol{w}\|_2. \tag{113}$$

The features span $\mathbb{R}^4$, hence $\mathcal{S}_{\mathrm{ns}} = \{\boldsymbol{0}\}$ and $\mathcal{Z}_{\mathrm{ns}} = \mathbb{R}^4$. In particular, $\mathcal{N}_{\mathrm{ns}} = \mathcal{L}_{\mathrm{ns}} = \{\boldsymbol{0}\}$. For $\boldsymbol{w} = (x, y, z, v)^\top$, v and $-v$ are the two newly introduced scores, and thus every recession direction has $v = 0$. The score inequalities of the first four examples yield $x \geq 0, y \geq -2x, |z| \leq mx$ and strictly hold at $\boldsymbol{e}_1$. Hence

$$\mathcal{M}_{\mathrm{ns}} = \mathbb{R}^3 \times \{\boldsymbol{0}\}, \qquad \mathcal{E}_{\mathrm{ns}} = \mathrm{span}\{\boldsymbol{e}_4\}, \qquad \mathcal{I}_{\mathrm{ns},\star} = \{(5, -1), (6, +1)\}.$$

The first four comparisons constitute $\mathcal{J}_{\mathrm{ns},\star}$. The max-margin value is one, and the max-margin set is $\iota(\mathcal{U}_{\mathrm{sep},\star})$. The reduced risk is

$$R_0(v\boldsymbol{e}_4) = \frac{1}{6}(\log(1 + e^{-v}) + \log(1 + e^{v})), \qquad \boldsymbol{v}_\star = 0.$$

These identify the spaces and objectives needed to apply the preceding theorems.

J.2 TRAJECTORY EMBEDDING

For every $\boldsymbol{w} \in \mathbb{R}^3$ and $v \in \mathbb{R}$,

$$R_{\mathrm{ns}}((\boldsymbol{w}^\top, v)^\top) = \frac{2}{3} R_{\mathrm{sep}}(\boldsymbol{w}) + \frac{1}{6}(\log(1 + e^{-v}) + \log(1 + e^{v})).$$

At $v = 0$,

$$R_{\mathrm{ns}}(\iota(\boldsymbol{w})) = \frac{2}{3} R_{\mathrm{sep}}(\boldsymbol{w}) + \frac{1}{3}\log 2, \qquad \nabla R_{\mathrm{ns}}(\iota(\boldsymbol{w})) = \frac{2}{3}\iota(\nabla R_{\mathrm{sep}}(\boldsymbol{w})). \tag{114}$$

The unit ball yields

$$\left\|(\boldsymbol{g}^\top, h)^\top\right\|_{\mathrm{ns},*} = \|\boldsymbol{g}\|_{\mathbb{B},*} + |h|.$$

For $\boldsymbol{g} \neq 0$, its norming face at $\iota(\boldsymbol{g})$ is

$$\mathcal{F}_{\mathrm{ns}}(\iota(\boldsymbol{g})) = \{(\boldsymbol{d}^\top, v)^\top : \boldsymbol{d} \in \mathcal{F}_{\mathbb{B}}(\boldsymbol{g}),\ |v| \leq 1\}, \qquad \|\iota(\boldsymbol{g})\|_{\mathrm{ns},*} = \|\boldsymbol{g}\|_{\mathbb{B},*}. \tag{115}$$

Since

$$\left\|(\boldsymbol{d}^\top, v)^\top\right\|_2^2 = \|\boldsymbol{d}\|_2^2 + v^2,$$

the unique minimum-Euclidean member has $v = 0$ and is the embedding of the minimum-Euclidean member of $\mathcal{F}_{\mathbb{B}}(\boldsymbol{g})$. Existence follows from compactness of the faces and uniqueness from strict convexity of the squared Euclidean norm.

Let $\boldsymbol{w}_t^{\mathrm{ns}}$ be the steepest descent trajectory under the tie-breaking rule of minimum Euclidean norm, initialized at zero, and let $\boldsymbol{s}_t^{\mathrm{ns}}$ be the corresponding norming direction at $\boldsymbol{w}_t^{\mathrm{ns}}$, that is,

$$\boldsymbol{s}_t^{\mathrm{ns}} := \operatorname*{arg\,min}_{\boldsymbol{d} \in \mathcal{F}_{\mathrm{ns}}(\nabla R_{\mathrm{ns}}(\boldsymbol{w}_t^{\mathrm{ns}}))} \|\boldsymbol{d}\|_2.$$

Let $\boldsymbol{w}_t$, η_t, and $\boldsymbol{s}_t$ be the trajectory, step size, and norming direction from Theorem 6.3 with the same minimum-Euclidean-norm tie-breaking:

$$\boldsymbol{s}_t := \operatorname*{arg\,min}_{\boldsymbol{d} \in \mathcal{F}_{\mathbb{B}}(\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t))} \|\boldsymbol{d}\|_2.$$

We apply the tie-breaking rule to $\boldsymbol{s}_t$ so that we have $\boldsymbol{s}_t^{\mathrm{ns}} = \iota(\boldsymbol{s}_t)$. Indeed, if $\boldsymbol{w}_t^{\mathrm{ns}} = \iota(\boldsymbol{w}_t)$, then

$$\boldsymbol{s}_t^{\mathrm{ns}} = \iota(\boldsymbol{s}_t), \qquad \|\nabla R_{\mathrm{ns}}(\boldsymbol{w}_t^{\mathrm{ns}})\|_{\mathrm{ns},*} = \frac{2}{3}\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\|_{\mathbb{B},*}.$$

Choose

$$\eta_t^{\mathrm{ns}} = \frac{3}{2}\eta_t. \tag{116}$$

Then

$$\boldsymbol{w}_{t+1}^{\mathrm{ns}} = \iota(\boldsymbol{w}_t) - \frac{3}{2}\eta_t \frac{2}{3}\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\|_{\mathbb{B},*}\,\iota(\boldsymbol{s}_t) = \iota\left(\boldsymbol{w}_t - \eta_t\|\nabla R_{\mathrm{sep}}(\boldsymbol{w}_t)\|_{\mathbb{B},*}\,\boldsymbol{s}_t\right) = \iota(\boldsymbol{w}_{t+1}).$$

Write $\boldsymbol{w}_t^{\mathrm{ns}} = \boldsymbol{u}_t^{\mathrm{ns}} + \boldsymbol{v}_t^{\mathrm{ns}}$, where $\boldsymbol{u}_t^{\mathrm{ns}} \in \mathcal{M}_{\mathrm{ns}}$ and $\boldsymbol{v}_t^{\mathrm{ns}} \in \mathcal{E}_{\mathrm{ns}}$. Induction therefore proves

$$\boldsymbol{w}_t^{\mathrm{ns}} = \boldsymbol{u}_t^{\mathrm{ns}} = \iota(\boldsymbol{w}_t), \qquad \boldsymbol{v}_t^{\mathrm{ns}} = 0 \qquad (t \geq 0). \tag{117}$$

A valid smoothness constant for the extended risk is

$$\beta_{\mathrm{ns}} := \frac{1}{4}\cdot\frac{1}{N}\left[1^2 + \left(\frac{7}{2}\right)^2 + 2\left(m + \frac{1}{4}\right)^2 + 2\cdot 1^2\right] = \frac{2}{3}\beta_{\mathrm{sep}} + \frac{1}{12}.$$

Choose

$$\eta_t = \frac{\eta_0}{t+1}, \qquad 0 < \eta_0 < \min\left\{\beta_{\mathrm{sep}}^{-1}, \frac{2}{3\beta_{\mathrm{ns}}}\right\}.$$

Both sequences satisfy Assumption 5.1; Theorems 4.3 and 5.2 supply risk and margin convergence. Equations 113 and 117 imply

$$\frac{\boldsymbol{u}_t^{\mathrm{ns}}}{\|\boldsymbol{u}_t^{\mathrm{ns}}\|_{\mathrm{ns}}} = \iota\left(\frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|_{\mathbb{B}}}\right).$$

Thus the accumulation set for both normalized $\boldsymbol{u}_t^{\mathrm{ns}}$ and normalized $\boldsymbol{w}_t^{\mathrm{ns}}$ is $\{(1,1,s,0)^\top : |s| \leq 1/4\}$; neither converges in direction.

K Directional convergence without convergence of steepest directions

Asymptotic alignment of the negative gradient and the parameter holds for Euclidean gradient flow under the hypotheses of Ji & Telgarsky (2020). The following example separates parameter directional convergence from convergence of steepest directions.

For $\boldsymbol{W} \in \mathbb{R}^{K\times D}$, define the entrywise norm and its dual by

$$\|\boldsymbol{W}\|_{1,1} := \sum_{a=1}^{K}\sum_{b=1}^{D}|W_{ab}|, \qquad \|\boldsymbol{G}\|_{\max} := \max_{a,b}|G_{ab}|.$$

Let $\boldsymbol{E}_{ab} \in \mathbb{R}^{3\times 4}$ be the standard basis matrix and let $\boldsymbol{e}_b \in \mathbb{R}^4$ be the b-th coordinate vector. Consider the dataset of $N = 6$ samples with $K = 3$ classes and $D = 4$,

$$\mathcal{D}_{\mathrm{align}} = \big((\boldsymbol{e}_1,1),(\boldsymbol{e}_1,1),(\boldsymbol{e}_2,1),(\boldsymbol{e}_3,1),(\boldsymbol{e}_3,2),(\boldsymbol{e}_3,3)\big).$$

The repetition of the first sample is intentional. Use Equation 2 with norm $\|\cdot\|_{1,1}$ and minimum-Frobenius-norm tie-breaking rule among the norming directions.

Proposition K.1 (Directional convergence with oscillating steepest directions)**.** *If $\eta_t > 0$, $\eta_t \to 0$, and $\sum_{t=0}^{\infty}\eta_t = \infty$, then the minimum-Frobenius steepest direction $\boldsymbol{S}_t$ does not converge, whereas*

$$\frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|_{1,1}} \to \frac{1}{2}(\boldsymbol{E}_{11} + \boldsymbol{E}_{12}).$$

Proof. We first show that $\boldsymbol{W}_t = a_t\boldsymbol{E}_{11} + b_t\boldsymbol{E}_{12}$ for all $t \geq 0$. For $\boldsymbol{W} = a\boldsymbol{E}_{11} + b\boldsymbol{E}_{12}$, the three examples at $\boldsymbol{e}_3$ have score gap zero. Hence,

$$R(\boldsymbol{W}) = \frac{1}{6}\left[2\log(1+2e^{-a}) + \log(1+2e^{-b}) + 3\log 3\right].$$

Set

$$g_1(a) := \frac{2}{3(2+e^a)} > 0, \qquad g_2(b) := \frac{1}{3(2+e^b)} > 0.$$

The full gradient is

$$\nabla R(\boldsymbol{W}) = \begin{pmatrix} -g_1(a) & -g_2(b) & 0 & 0 \\ g_1(a)/2 & g_2(b)/2 & 0 & 0 \\ g_1(a)/2 & g_2(b)/2 & 0 & 0 \end{pmatrix},$$

thus $\|\nabla R(\boldsymbol{W})\|_{\max} = \max\{g_1(a), g_2(b)\}$. The minimum-Frobenius member therefore satisfies

$$-\boldsymbol{S}_t = \begin{cases} \boldsymbol{E}_{11}, & g_1(a_t) > g_2(b_t), \\ \boldsymbol{E}_{12}, & g_1(a_t) < g_2(b_t), \\ (\boldsymbol{E}_{11} + \boldsymbol{E}_{12})/2, & g_1(a_t) = g_2(b_t). \end{cases}$$

Induction proves that $\boldsymbol{W}_t = a_t \boldsymbol{E}_{11} + b_t \boldsymbol{E}_{12}$ with nondecreasing $a_t, b_t \geq 0$. Let

$$\alpha_t := \eta_t \max\{g_1(a_t), g_2(b_t)\} > 0.$$

Since g_1 and g_2 are bounded, $\alpha_t \to 0$. The update of a_t and b_t is then

$$(a_{t+1}, b_{t+1}) = \begin{cases} (a_t + \alpha_t, b_t), & g_1(a_t) > g_2(b_t), \\ (a_t, b_t + \alpha_t), & g_1(a_t) < g_2(b_t), \\ (a_t + \frac{\alpha_t}{2}, b_t + \frac{\alpha_t}{2}), & g_1(a_t) = g_2(b_t). \end{cases}$$

We show that neither strict inequality holds eventually. Suppose not; if $g_1(a_t) > g_2(b_t)$ for all $t \geq T$, then $b_t = b_T$ and

$$a_{t+1} - a_t = \eta_t g_1(a_t) \geq \eta_t g_2(b_T).$$

Then $a_t \to \infty$, contradicting $g_1(a_t) > g_2(b_T) > 0$. If $g_1(a_t) < g_2(b_t)$ for all $t \geq T$, then $a_t = a_T$ and

$$b_{t+1} - b_t = \eta_t g_2(b_t) \geq \eta_t g_1(a_T).$$

This forces $b_t \to \infty$, contradicting $g_2(b_t) > g_1(a_T) > 0$. Moreover, the equality $g_1(a_t) = g_2(b_t)$ also cannot eventually hold. If $g_1(a_t) = g_2(b_t)$, then

$$\frac{g_1(a_{t+1})}{g_2(b_{t+1})} = 2 \cdot \frac{2 + \exp(b_t + \frac{\alpha_t}{2})}{2 + \exp(a_t + \frac{\alpha_t}{2})} < 1.$$

If there are infinitely many ties, $(\boldsymbol{E}_{11} + \boldsymbol{E}_{12})/2$ and $\boldsymbol{E}_{12}$ both occur infinitely often. If there are only finitely many ties, $\boldsymbol{E}_{11}$ and $\boldsymbol{E}_{12}$ both occur infinitely often. In either case, $\boldsymbol{S}_t$ does not converge.

We now show the directional convergence of $\boldsymbol{W}_t$. Define

$$h_t := a_t - b_t - \log(2 + 2e^{-b_t}).$$

Notice that $g_1(a_t) > g_2(b_t) \Leftrightarrow h_t < 0$ and $g_1(a_t) < g_2(b_t) \Leftrightarrow h_t > 0$. Then

$$h_{t+1} = \begin{cases} h_t + \alpha_t, & h_t < 0, \\ h_t - \alpha_t + \log(2 + 2e^{-b_t}) - \log(2 + 2e^{-b_t - \alpha_t}), & h_t > 0, \\ \log(2 + 2e^{-b_t}) - \log(2 + 2e^{-b_t - \frac{\alpha_t}{2}}), & h_t = 0. \end{cases}$$

For $b \geq 0$, the derivative of $\log(2+2e^{-b})$ is $-1/(1+e^b) \in [-1/2, 0)$. When $h_t = 0$, $0 \leq h_{t+1} \leq \frac{\alpha_t}{4}$ by the mean value theorem. When $h_t > 0$, the decrement $h_t - h_{t+1}$ belongs to $[\frac{\alpha_t}{2}, \alpha_t]$. Thus, a step that preserves a nonzero sign decreases $|h_t|$, whereas a sign change or departure from a tie forces $|h_t|$ to be at most α_t. Hence,

$$|h_{t+1}| \leq \max\{|h_t|, \alpha_t\}.$$

There should be arbitrarily late sign changes in h_t or $h_t = 0$. Given $\varepsilon > 0$, choose T so that $\alpha_t < \varepsilon$ for all $t \geq T$ and then let $t_T \geq T$ with $h_{t_T} h_{t_T+1} \leq 0$. Then $|h_{t_T+1}| \leq \alpha_{t_T} < \varepsilon$, and $|h_t| < \varepsilon$ for all $t > t_T$. Thus $h_t \to 0$.

Both sequences $\{a_t\}_t$ and $\{b_t\}_t$ are nondecreasing. Since $a_t - b_t$ is bounded, boundedness of either sequence would imply boundedness of the other. Then $\max\{g_1(a_t), g_2(b_t)\} \geq c > 0$. But

$$(a_{t+1} + b_{t+1}) - (a_t + b_t) = \alpha_t \geq c\eta_t,$$

which contradicts boundedness because $\sum_t \eta_t = \infty$. Therefore $a_t, b_t \to \infty$ and $a_t - b_t \to \log 2$.

For $t \geq 1$, $\|\boldsymbol{W}_t\|_{1,1} = a_t + b_t > 0$, and thus

$$\left\| \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|_{1,1}} - \frac{1}{2}(\boldsymbol{E}_{11} + \boldsymbol{E}_{12}) \right\|_{1,1} = \frac{|a_t - b_t|}{a_t + b_t} \to 0. \qquad \square$$

L NORM-DEPENDENT MAX-MARGIN DUALITY

Let $\Delta(\mathcal{J}_\star)$ be the probability simplex over $\mathcal{J}_\star$. Define the score-gap map $\mathcal{T}_\mathcal{J} : \mathcal{M} \to \mathbb{R}^{\mathcal{J}_\star}$ by $\mathcal{T}_\mathcal{J} \boldsymbol{U} = (\Delta_{ic}(\boldsymbol{U}))_{(i,c)\in\mathcal{J}_\star}$. The (Frobenius) adjoint $\mathcal{T}_\mathcal{J}^* : \mathbb{R}^{\mathcal{J}_\star} \to \mathcal{M}$ of $\mathcal{T}_\mathcal{J}$ is determined by

$$\left\langle \mathcal{T}_\mathcal{J}^* \boldsymbol{g}, \boldsymbol{U} \right\rangle = \sum_{(i,c)\in\mathcal{J}_\star} g_{ic} \Delta_{ic}(\boldsymbol{U}).$$

By the definition of $\boldsymbol{M}_{(i,c)}$ in Section A, this adjoint is explicitly

$$\mathcal{T}_\mathcal{J}^* \boldsymbol{g} = \Pi_\mathcal{M} \left(\sum_{(i,c)\in\mathcal{J}_\star} g_{ic} \boldsymbol{M}_{(i,c)} \right) \in \mathcal{M}.$$

Theorem L.1 (Margin duality)**.** *If $\mathcal{J}_\star \neq \varnothing$, then*

$$\gamma_\star := -\min_{\boldsymbol{U}\in\mathcal{M}} \left\{ \max_{(i,c)\in\mathcal{J}_\star} \Delta_{ic}(\boldsymbol{U}) : \|\boldsymbol{U}\| \leq 1 \right\} = \min_{\boldsymbol{g}\in\Delta(\mathcal{J}_\star)} \left\| \mathcal{T}_\mathcal{J}^* \boldsymbol{g} \right\|_{\mathcal{M},*} > 0. \tag{118}$$

For every primal optimum $\boldsymbol{U}_\star$ and every dual optimum $\boldsymbol{g}_\star$,

$$\left\langle \mathcal{T}_\mathcal{J}^* \boldsymbol{g}_\star, \boldsymbol{U}_\star \right\rangle = -\gamma_\star, \qquad \boldsymbol{U}_\star \in \operatorname*{arg\,min}_{\substack{\boldsymbol{U}\in\mathcal{M} \\ \|\boldsymbol{U}\|\leq 1}} \left\langle \mathcal{T}_\mathcal{J}^* \boldsymbol{g}_\star, \boldsymbol{U} \right\rangle. \tag{119}$$

If the norm (restricted to $\mathcal{M}$) is strictly convex, then $\boldsymbol{U}_\star$ is unique.

Proof. Choose $\widehat{\boldsymbol{U}}$ from Lemma D.2: $\Delta_{ic}(\widehat{\boldsymbol{U}}) < 0$ for every $(i,c) \in \mathcal{J}_\star$. After normalizing $\widehat{\boldsymbol{U}}$, this gives $\gamma_\star > 0$.

Sion's minimax duality (Sion, 1958) gives

$$\begin{aligned}
\min_{\boldsymbol{g}\in\Delta(\mathcal{J}_\star)} \left\| \mathcal{T}_\mathcal{J}^* \boldsymbol{g} \right\|_{\mathcal{M},*} &= -\max_{\boldsymbol{g}\in\Delta(\mathcal{J}_\star)} \min_{\substack{\boldsymbol{U}\in\mathcal{M} \\ \|\boldsymbol{U}\|\leq 1}} \left\langle \mathcal{T}_\mathcal{J}^* \boldsymbol{g}, \boldsymbol{U} \right\rangle \\
&= -\min_{\substack{\boldsymbol{U}\in\mathcal{M} \\ \|\boldsymbol{U}\|\leq 1}} \max_{\boldsymbol{g}\in\Delta(\mathcal{J}_\star)} \left\langle \mathcal{T}_\mathcal{J}^* \boldsymbol{g}, \boldsymbol{U} \right\rangle \\
&= -\min_{\substack{\boldsymbol{U}\in\mathcal{M} \\ \|\boldsymbol{U}\|\leq 1}} \max_{(i,c)\in\mathcal{J}_\star} \Delta_{ic}(\boldsymbol{U}) = \gamma_\star.
\end{aligned}$$

The second equality holds as the domain of $\min$ and $\max$ are both compact and convex, and the objective $\left\langle \mathcal{T}_\mathcal{J}^* \boldsymbol{g}, \boldsymbol{U} \right\rangle$ is bilinear. Every primal optimum has norm one: otherwise rescaling would increase its positive minimum margin. Primal optimality gives $\Delta_{ic}(\boldsymbol{U}_\star) \leq -\gamma_\star$ for $(i,c) \in \mathcal{J}_\star$, and hence

$$\left\langle \mathcal{T}_\mathcal{J}^* \boldsymbol{g}_\star, \boldsymbol{U}_\star \right\rangle \leq -\gamma_\star.$$

Hölder's inequality gives the reverse bound:

$$\left\langle \mathcal{T}_\mathcal{J}^* \boldsymbol{g}_\star, \boldsymbol{U}_\star \right\rangle \geq -\left\| \mathcal{T}_\mathcal{J}^* \boldsymbol{g}_\star \right\|_{\mathcal{M},*} \|\boldsymbol{U}_\star\| = -\gamma_\star.$$

Thus equality holds.

If two distinct unit primal optima existed under a strictly convex norm, their midpoint would have a norm strictly below one and all inactive margins at least $\gamma_\star$. Rescaling the midpoint to unit norm would produce a larger margin than $\gamma_\star$, a contradiction. □