ON THE COMPARISON OF OPTIMIZERS FOR IMBALANCED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Data imbalance is pervasive in machine learning, from rare words and anomalies to underrepresented patterns in heterogeneous or cross-tabulated data. We study idealized optimizers geometries in continuous time to model small-step training in deep learning. We assume that the source of imbalance is unobserved: the optimizer has only access to the aggregate training loss ignoring the exact contributions of the majority and minority groups. In this setting, we characterize a region where majority losses are optimized regardless of the admissible minority structure. We derive explicit equations of this zone and bounds on the time needed to leave it. These bounds exhibit a milder dependence on minority amplitude for sign, spectral, and Newton descent than for Euclidean gradient descent. Experiments with AdamW and Muon suggest similar advantages over SGD across language, tabular, and image tasks.

1 INTRODUCTION

Training data are rarely well balanced. Even when the observed class labels are carefully supervised, the data may contain latent minority groups associated with acquisition conditions, rare subclasses, demographic attributes, or rare linguistic patterns. Such groups arise for instance in hidden-subclass learning, long-tailed recognition Kearns et al. (2018); Sohoni et al. (2020); Liu et al. (2019), tabular data, or subgroup fairness. Related phenomena also occur in language modelling, where token and context frequencies are strongly heterogeneous by nature Kunstner et al. (2024); Kunstner & Bach (2025). These groups may be unknown before training or irreducibly imbalanced (rare words, many-features subgroups in tabular data), and in addition, in some cases methods that require their identities for resampling or reweighting are doomed to failure.

Building on the study of gradient descent bias in Bachoc et al. (2025), we focus on training with imbalanced data and study when the optimization trajectory leaves the region of parameter space in which the majority is systematically favored.

In our training model, the user need not know the precise majority–minority decomposition $L = L_1 + L_0$, where L_0 is a minority component and L_1 its associated majority. Only an amplitude $\psi > 0$, measuring the impact of L_0, is known. This assumption fits many real-life scenarios, from tabular data to LLMs as presented in Section 2. Within this idealized training setting, our protocol defines a minority awareness region associated with an optimizer (modelled by an ODE) and then studies the first time the resulting trajectory enters this region, leaving the place where majority is systematically favored. This entry time provides an indication of the optimizer's ability to deal with imbalanced data. We now describe this procedure more precisely.

Idealizing optimizer geometry. A direct theoretical comparison of modern optimizers is out of reach, and our modelling proceeds through two distinct idealizations. First, we retain their dominant update geometries while discarding other algorithmic features as weight decay, momentum, minibatching. Thus, classically, gradient descent represents SGD, sign descent matches Adam-type methods, and spectral descent corresponds to Muon and Scion-type methods (Pethick et al., 2025; Oikonomidis et al., 2026; Pethick et al., 2026). Although intractable in high dimension Newton's descent is studied to understand the impact of second-order information. Secondly, as typical learning rates are lower than 10^{-2}, we study optimizer vector fields in continuous time, that is solutions of $\dot{\theta}(t) \in -V_L(\theta(t))$ with $\theta(t) \in \mathbb{R}^d$ with $V_L : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ being multivalued.

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.

Comparing ODEs clocks Continuous time introduces a potential ambiguity in time comparisons: scaling the vector field V_L scales time by the same factor, so the minority awareness time described above is not intrinsically meaningful. To solve this apparent issue, we return to discrete methods, $\theta_{k+1} \in \theta_k - \eta V_L(\theta_k)$, for which the natural clock is $t \simeq \eta k$. Over a range of admissible step sizes,

$$\eta \in [\eta_{\min}, \eta_{\max}], \qquad \kappa := \frac{\eta_{\max}}{\eta_{\min}},$$

the resulting clock discrepancies are therefore controlled by a factor κ. Key to our approach is the fact that deep learning losses have rather narrow learning rates regimes across optimizers, e.g. for networks for vision tasks, learning rates typically range from $\eta_{\min} \simeq 10^{-3}$ to $\eta_{\max} \simeq 10^{-2}$, giving $\kappa \simeq 20$ (see Section D for discussions on tabular data and LLM). Our continuous-time model is therefore justified on two grounds: as explained earlier the small-step regime makes the continuous dynamics an acceptable model, while the moderate κ value for DL losses ensures that the corresponding time clocks remain comparable up to a controlled factor.

Minority awareness and entry time. As in Bachoc et al. (2025), we study an initial "minority-blind" regime in which majority learning can proceed regardless of the admissible latent minority structure. For each idealized optimizer (gradient descent, sign descent, spectral descent, or Newton's method) and each trajectory, we define t_{entry} as the first entry into the minority-aware zone. Within this zone, at least one admissible decomposition yields minority improvement at the majority's expense. This need not correspond to learning the true latent subgroup: awareness concerns at least one admissible minority at the prescribed level ψ. We therefore use t_{entry} as an indicator of exposure to imbalance, comparing how rapidly different optimizer geometries leave the initial majority-dominated regime.

Main results. Explicit equations of minority aware zones and entry-time estimates (both upper and lower) are derived for many optimizers: gradient descent, gradient clipping, sign descent, spectral descent, and Newton descent. Their dependence on the minority scale ψ differs sharply: normalized geometries can enjoy a substantially milder deterioration than Euclidean gradient descent as imbalance becomes severe, with additional gains governed by gradient support or matrix rank. Both are empirically related to the global dimension. Newton descent provides an idealized second-order reference confirming that clever geometric rescaling is key for balanced training.

Empirical evidences. We assess the relevance of entry times as indicators for imbalanced learning through two complementary approaches. First, in linear regression, where the latent decomposition and entry times can be computed, we empirically compare entry-time ratios across optimizers and find close agreement with our theoretical predictions. Second, in classification with deep networks and language modelling, where the minority level ψ and the entry time t_{entry} are not directly observable, we evaluate the practical counterparts of the optimizers studied in our analysis—SGD, AdamW, and Muon—by measuring their performance on minority and majority data. Across these experiments, rescaled methods show a systematic advantage over SGD in minority learning, while the relative performance of AdamW and Muon depends on the problem geometry. In language modelling, these results are consistent with the advantage of Adam-type normalization under heavy-tailed frequencies reported in (Kunstner et al., 2024; Kunstner & Bach, 2025) while corroborating the geometric benefits of Muon highlighted in recent literature (Vasudeva et al., 2025; Wang et al., 2025).

2 A MODEL OF IMBALANCE IN SUPERVISED LEARNING

Minorities in supervised learning We consider supervised learning samples $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, for $1 \leq i \leq n$, where $\mathcal{X}$ and $\mathcal{Y}$ are general sets. We are interested in the situation where a small number of pairs (x_i, y_i) (or equivalently of indices i) belong to an unknown minority group.

Typically, in binary classification where $\mathcal{Y} = \{0, 1\}$, there could be a small number of (x_i, y_i) with $y_i = 1$, that would form a minority group (that would be easily identified). Nevertheless, we are interested in the more challenging case where this minority group cannot be identified easily. In binary classification, we would have $\#\{i : y_i = 0\} \simeq \#\{i : y_i = 1\}$, but in the features $(x_i)_{i=1}^n$, we would have hidden minority structures.

To model and motivate these minority structures, we introduce latent descriptors of the form

$$z_i = \phi(x_i, y_i) \in \mathcal{Z},$$

where $\phi : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ is an unknown map into some set $\mathcal{Z}$. Note that ϕ may only depend on the features, in which case it would hold $\phi(x_i, y_i) = \phi(x_i)$. Fixing a set $\mathcal{A} \subset \mathcal{Z}$, the index set

$$I_{\mathcal{A}} := \{i \in \{1, \dots, n\} : \phi(x_i, y_i) \in \mathcal{A}\} \tag{1}$$

may yield a latent minority group $(x_i, y_i)_{i \in I_{\mathcal{A}}}$. Let us give concrete cases.

a) Tabular Data. Suppose $z_\phi \in \mathcal{X} = \mathbb{R}^d$, define $\phi(x, y) = z_\phi^\top x \in \mathbb{R}$ and let $\phi(x, y)$ be a score such as age, weight, or wealth. Then for each threshold $t \in \mathbb{R}$ one may take $\mathcal{A}_t = (-\infty, t]$. In that case $I_{\mathcal{A}_t} = \{i : \phi(x_i) \leq t\}$ is a minority set of individuals with particularly low scores.

b) Hidden subclasses in classification. Consider multi-class classification where the labels $y_i \in \{1, \dots, C\}$ are balanced. Nevertheless, it is still possible that each sample also belongs to a subclass $h_i \in \{1, \dots, M\}$, corresponding to a viewpoint, shooting condition, scene configuration, scanner type, or image resolution. Fix a rare regime $h_\star \in \{1, \dots, M\}$, and define $\phi(x, y) \in \{1, \dots, M\}$ with $\mathcal{A} = \{h_\star\}$. Then $I_{\mathcal{A}} = \{i : \phi(x_i, y_i) = h_\star\}$. The minority group is then the rare visual regime $h_\star$. This corresponds to the hidden-subclass setting of Sohoni et al. (2020): the imbalance does not lie in the labels themselves, but within each class; see also Kearns et al. (2018).

c) Rare-word counts in autoregressive language models. Let $(s_t)_{1 \leq t \leq N}$ be a tokenized corpus with vocabulary $\mathcal{V}$. Each training sample is $x_t = (s_1, \dots, s_{t-1}), y_t = s_t$. For each token $v \in \mathcal{V}$, let $N_v := \#\{t : y_t = v\}$ and assume sorted counts $N_{(1)} \geq \cdots \geq N_{(|\mathcal{V}|)}$. We set $\phi(x_t, y_t) = N_{y_t}$, and for a given rank r, $\mathcal{A}_r = \{1, \dots, N_{(r)}\}$ isolates a minority

$$I_{\mathcal{A}_r} = \{t : N_{y_t} \leq N_{(r)}\}$$

whose target tokens fall into the long tail of the distribution (Piantadosi, 2014; Kunstner et al., 2024; Kunstner & Bach, 2025; Radford et al., 2019). If the empirical frequencies exhibit a power-law decay, $N_{(r)} \propto r^{-\alpha}$, then

$$\frac{|\mathcal{I}_{\mathcal{A}_r}|}{N} \approx \frac{\sum_{k=r}^{|\mathcal{V}|} k^{-\alpha}}{\sum_{k=1}^{|\mathcal{V}|} k^{-\alpha}}.$$

Empirical risk minimization Let $\ell(\theta; x, y) \in \mathbb{R}$ be a loss (typically associated to a neural network with parameters $\theta \in \mathbb{R}^d$). Define the empirical risk $L(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\theta; x_i, y_i)$. Given a minority group $I_{\mathcal{A}}$ as above, we can define the minority loss and the corresponding majority loss

$$L_0(\theta) = \frac{1}{n} \sum_{i \in I_{\mathcal{A}}} \ell(\theta; x_i, y_i), \quad L_1(\theta) = \frac{1}{n} \sum_{i \in \{1, \dots, n\} \setminus I_{\mathcal{A}}} \ell(\theta; x_i, y_i). \tag{2}$$

We have the loss decomposition $L = L_0 + L_1$, and, as discussed above, this decomposition is unknown in general, and even more, multiple notions of minority simultaneously give rise to multiple decompositions. This structural ambiguity is inherent to many learning problems; it appears in subgroup fairness (Kearns et al., 2018), hidden-subclass problems (Oakden-Rayner et al., 2020; Sohoni et al., 2020), and long-tailed recognition (Kang et al., 2020; Liu et al., 2019). In the sequel, we adopt a general yet theoretically tractable viewpoint, considering a general loss function L and a general set of admissible minority losses L_0.

The learning loss model With the above motivation in mind, for $d \in \mathbb{N}^*$, consider a loss $L : \mathbb{R}^d \to \mathbb{R}$, for which we make the standing assumption

Assumption 2.1 (Loss assumptions)**.** *L is C^2 with $L_\star := \inf_{\theta \in \mathbb{R}^d} L(\theta) > -\infty$.*

Motivated by the above, we assume that L is subject to an unknown decomposition

$$\forall \theta \in \mathbb{R}^d, \quad L(\theta) = L_1(\theta) + L_0(\theta),$$

where $L_1, L_0 : \mathbb{R}^d \to \mathbb{R}$ are twice continuously differentiable functions and, for some $\psi, D > 0$,

$$L_0 \in \mathcal{K}_{\min} := \left\{ M \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R}) \ : \ \forall \theta \in \mathbb{R}^d, \ \|\nabla M(\theta)\|_2 \leq \psi(\|\theta\|_2 + D) \right\}. \tag{3}$$

The abstract decomposition $L = L_1 + L_0$ should therefore be interpreted as a way to model uncertainty over minority structures while keeping the total objective fixed.

In view of the above, ψ may be understood as a bound on the effective mass of the latent minority group. Coming back to Equations 1 and 2, it is informally interpreted as

$$\psi \approx \frac{n_0}{n_0 + n_1} G,$$

where n_0 is the cardinality of $\mathcal{I}_\mathcal{A}$, $n_1 = n - n_0$ and G is the "average" Lipschitz constant of the gradients of $\ell(\cdot; x_i, y_i)$, for $i \in \mathcal{I}_\mathcal{A}$. When the loss function has bounded Hessian, we consider $\psi := \sigma_{\max}(\nabla^2 L_0)$ (largest singular value). Hence, as further discussed and quantified in Bachoc et al. (2025), L_0 can be small because the minority losses $\ell(\cdot; x_i, y_i)$ are scarce or because they have a small variability.

3 Learning process and minority aware zone

3.1 Learning with descent methods

The learning process is modeled by a continuous-time dynamical system driven by an optimizer $V_L : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$, which may be multivalued, that is $V_L(\theta) \subset \mathbb{R}^d$ for $\theta \in \mathbb{R}^d$. We refer to Aubin & Cellina (1984) for the theory. In the remainder we assume the following.

Assumption 3.1 (Dynamics assumptions)**.** *V_L is nonempty convex compact valued, has closed graph and is locally bounded.*

From Lemma A.1, for any $\theta_{\text{init}} \in \mathbb{R}^d$, there exists an absolutely continuous curve θ and $T > 0$ satisfying

$$\dot{\theta}(t) \in -V_L(\theta(t)) \quad \text{for a.e. } t \in [0, T), \qquad \theta(0) = \theta_{\text{init}}. \tag{4}$$

In the various settings we adopt, we will see that $T = +\infty$.

Denote by $dL(\theta)$ the differential mapping of L at θ. In the sequel, we shall consider various choices for V_L, for which the key mechanism of descent property holds:

$$dL(\theta) = 0 \Leftrightarrow V_L(\theta) \ni 0 \tag{5}$$

$$dL(\theta)(v) = \langle v, \nabla L(\theta) \rangle > 0, \ \forall v \in V_L(\theta), \ \forall \theta \in \mathbb{R}^d \setminus dL^{-1}(\{0\}), \tag{6}$$

which ensures that the loss decreases along trajectories of Equation 4.

Five idealized descent mechanisms are considered: gradient descent, gradient clipping, sign descent, spectral descent, and Newton descent. They share a common variational construction: gradient directions are selected by maximizing the differential over a norm ball, and the speed may then be adjusted through clipping or normalization. For simplicity, we first present this construction for a fixed norm $\|\cdot\|$ on $\mathbb{R}^d$. Parameter-dependent norms $\|\cdot\|_\theta$ would accommodate Newton descent.

Normal gradients For a linear form ℓ on $\mathbb{R}^d$, define its dual norm by $\|\ell\|_* := \max_{\|z\| \leq 1} \ell[z]$. For a differentiable function L, the corresponding norm-gradient writes

$$\operatorname{grad}_{\|\cdot\|} L(\theta) := \operatorname*{arg\,max}_{\|z\| \leq \|\mathrm{d}L(\theta)\|_*} \mathrm{d}L(\theta)[z], \tag{7}$$

while the unit norm-gradient directions are $\rho(\operatorname{grad}_{\|\cdot\|} L(\theta)) := \operatorname*{arg\,max}_{\|z\| \leq 1} \mathrm{d}L(\theta)[z]$. The gradient $\operatorname{grad}_{\|\cdot\|}$ could be multivalued depending on the curvature properties of the unit ball of the norm $\|\cdot\|$. From the definitions, we infer:

$$\forall v \in \operatorname{grad}_{\|\cdot\|} L(\theta), \quad \|v\| = \|\mathrm{d}L(\theta)\|_* \quad \text{with} \quad \mathrm{d}L(\theta)[v] = \|\mathrm{d}L(\theta)\|_*^2. \tag{8}$$

By standard arguments the multivalued mappings $\operatorname{grad}_{\|\cdot\|}$ and $\rho(\operatorname{grad}_{\|\cdot\|} L(\theta))$ are closed maps with nonempty convex compact values.

In the remainder the normal gradient descent and the normalized gradient descent will play a key role:

$$\dot{\theta}(t) \in -\operatorname{grad}_{\|\cdot\|} L(\theta(t)) \text{ and } \dot{\theta}(t) \in -\rho(\operatorname{grad}_{\|\cdot\|} L(\theta(t))).$$

We have the following well-known basic properties.

Gradient descent: For Euclidean gradient descent, choose $\|\cdot\| = \|\cdot\|_2$ so that Equation 7 gives

$$V_L(\theta) = \mathrm{grad}_{\|\cdot\|_2} L(\theta) = \nabla L(\theta), \qquad \dot{\theta}(t) = -\nabla L(\theta(t)).$$

Gradient clipping: Fix $\tau_{\mathrm{clip}} > 0$, and define $\mathrm{Clip}_{\tau_{\mathrm{clip}}}(g) := \begin{cases} g, & \|g\|_2 \leq \tau_{\mathrm{clip}}, \\ \tau_{\mathrm{clip}} \dfrac{g}{\|g\|_2}, & \|g\|_2 > \tau_{\mathrm{clip}}, \end{cases}$ whose dynamics is $\dot{\theta}(t) = -\mathrm{Clip}_{\tau_{\mathrm{clip}}}(\nabla L(\theta(t)))$.

Sign descent: Choose $\|\cdot\| = \|\cdot\|_\infty$, and consider $V_L = \rho(\mathrm{grad}_{\|\cdot\|_\infty} L(\theta))$. Define the scalar multivalued sign by $\mathrm{sign}(s) = 1$ for $s > 0$, $\mathrm{sign}(s) = -1$ for $s < 0$, and $\mathrm{sign}(0) = [-1, 1]$ and extend it component-wise. The dynamics given by V_L is called *sign descent*:

$$\dot{\theta}(t) \in -\mathrm{sign}(\nabla L(\theta(t))).$$

Spectral descent: Assume the parameter space is of the form $\mathbb{R}^{a \times b}$. The differential is represented by the matrix gradient through

$$\mathrm{d}L(\Theta)[H] = \mathrm{tr}(\nabla L(\Theta)^\top H) \qquad \forall \Theta \in \mathbb{R}^{a \times b} \qquad \forall H \in \mathbb{R}^{a \times b}.$$

The matrix Spectral norm (w.r.t. to Euclidean norm) is denoted by $\|\cdot\|_{\mathrm{sp}}$, its dual norm –the nuclear norm– is denoted by $\|\cdot\|_{\mathrm{nuc}}$. Spectral descent is the corresponding normalized spectral-norm descent:

$$\dot{\Theta}(t) \in -\rho(\mathrm{grad}_{\|\cdot\|_{\mathrm{sp}}} L(\Theta(t))). \tag{9}$$

More details about spectral descent are provided in Section A.5

Newton descent: Assume there exists $\mu > 0$ such that

$$\langle u, \nabla^2 L(\theta) u \rangle \geq \mu \|u\|_2^2 \qquad \forall \theta, u \in \mathbb{R}^d \tag{10}$$

and define the *Hessian norm* at θ in $\mathbb{R}^d$ by $\|u\|_\theta^2 := \langle u, \nabla^2 L(\theta) u \rangle$, for any u in $\mathbb{R}^d$. Equation 7 yields the Newton field $V_L(\theta) = \nabla^2 L(\theta)^{-1} \nabla L(\theta)$ with $\dot{\theta}(t) = -\nabla^2 L(\theta(t))^{-1} \nabla L(\theta(t))$.

3.2 Majority adverse zone and minority aware zone

For a fixed loss L, we use the admissible minority class $\mathcal{K}_{\min}$ defined in Equation (3). The corresponding admissible "majority losses"[1] are

$$\mathcal{K}_{\mathrm{maj}}(L) := \{L - M : \ M \in \mathcal{K}_{\min}\}. \tag{11}$$

For the sequel, fix a decomposition $L = L_1 + L_0$ with $L_0 \in \mathcal{K}_{\min}$ and $L_1 = L - L_0 \in \mathcal{K}_{\mathrm{maj}}(L)$, and consider a possibly multivalued optimizer $V_L : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$. Based on the majority adverse zone introduced in Bachoc et al. (2025), we introduce the following variation:

Definition 3.1 (Majority adverse zone)**.** *Consider a loss function $L \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$, an optimizer $V_L : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ satisfying Assumption 3.1, and $L_0 \in \mathcal{K}_{\min}$, yielding a fixed decomposition $L = L_1 + L_0$. The majority adverse zone associated with (L, V_L, L_0) is*

$$Z_{\mathrm{maj-adv}}(L_0) := \left\{ \theta \in \mathbb{R}^d : \sup_{v \in V_L(\theta)} \langle v, \nabla L_1(\theta) \rangle < 0 \right\}. \tag{12}$$

Equivalently, $\theta \in Z_{\mathrm{maj-adv}}(L_0)$ iff $\langle v, \nabla L_1(\theta) \rangle < 0$ for all $v \in V_L(\theta)$, which matches the definition in Bachoc et al. (2025) when $V_L = \nabla L$ and with $\leq$ in Equation 12, see Remark A.1.

As the majority-minority decomposition is unknown, we take the union of the majority adverse zones over all admissible minority components, to obtain what we called the minority aware zone.

[1]The term majority refers to the potential magnitude of the losses in $\mathcal{K}_{\mathrm{maj}}$ not necessarily to their actual magnitude as they may be for instance null functions (in the case where $L \in \mathcal{K}_{\min}$ and $L_0 = L$).

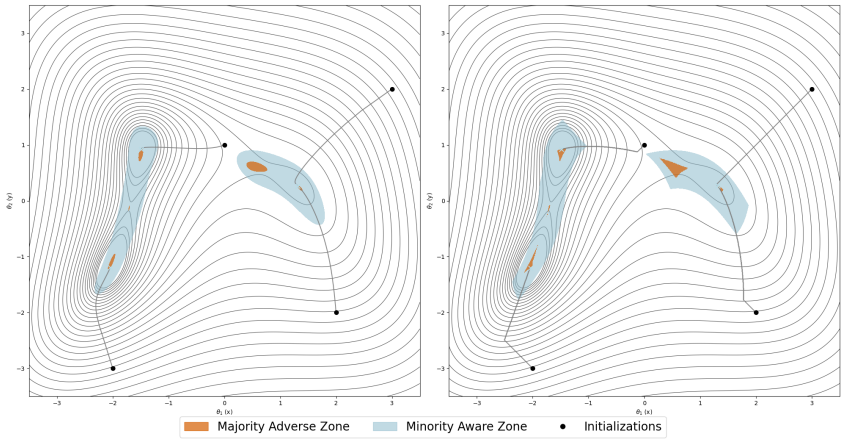


Figure 1: Landscape L, majority adverse, minority aware zones, learning curves for gradient descent (left) and sign descent (right).

Definition 3.2 (Minority aware zone)**.** *Consider a loss function $L \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$ and an optimizer $V_L : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ satisfying Assumption 3.1. The minority aware zone is*

$$Z_{\min} := \bigcup_{L_0 \in \mathcal{K}_{\min}} Z_{\mathrm{maj-adv}}(L_0).$$

As explained in the next lemma, the complement of $Z_{\min}$ is, roughly speaking, a minority-blind regime: the learning dynamic remains majority-favouring whatever the decomposition. Alternatively the minority aware zone may be seen as the locus of points where the optimizer favors at least one minority in some decomposition. In $Z_{\min}$ complete blindness ceases; see Lemma A.6.

Lemma 3.1 (Uniform majority dominance)**.** *Under Assumptions 2.1 and 3.1, suppose that V_L is single-valued. Then, for every $\theta \in \mathbb{R}^d$,*

$$\theta \notin Z_{\min} \iff dL_1(\theta)[-V_L(\theta)] \leq 0 \quad \textit{for every } L_1 \in \mathcal{K}_{\mathrm{maj}}(L).$$

Thus, every admissible majority loss is nonincreasing along a trajectory that remains outside $Z_{\min}$.

See also Theorem A.5 and the Appendix for other considerations on the minority awareness zone. An illustration is provided in Figure 1.

3.3 Variational characterization of the minority aware zone

Variational representations Minority aware zones can be represented as level sets of functions that, in all cases considered below, admit explicit formulas. For $L_0 \in \mathcal{K}_{\min}$, set $L_1 = L - L_0$ and define

$$Q_{L_0}(\theta) := \sup_{v \in V_L(\theta)} \langle v, \nabla L_1(\theta) \rangle.$$

This function measures the worst, over all admissible velocities, instantaneous variation of the majority loss. It is made to detect zones where L_1 does not decrease, indeed: $Z_{\mathrm{maj-adv}}(L_0) = \{\theta \in \mathbb{R}^d : \ Q_{L_0}(\theta) < 0\}$. We now aggregate over all admissible minority losses by setting

$$\mathcal{Q}(\theta) := \inf_{L_0 \in \mathcal{K}_{\min}} Q_{L_0}(\theta) = \inf_{L_0 \in \mathcal{K}_{\min}} \sup_{v \in V_L(\theta)} \langle v, \nabla (L - L_0)(\theta) \rangle \,, \tag{13}$$

which is finite-valued due to the definition of $\mathcal{K}_{\min}$. We immediately prove the following lemma.

Lemma 3.2 (Minority zone and $\mathcal{Q}$)**.** *Under Assumption 3.1, and by definition of $\mathcal{K}_{\min}$, we have:*

$$\theta \in Z_{\min} \iff \mathcal{Q}(\theta) < 0, \quad \textit{for every } \theta \in \mathbb{R}^d.$$

Equations for the minority aware zone In order to obtain explicit equations defining $\mathcal{Q}$, we shall use the following key results:

Lemma 3.3 (Pointwise gradient envelope for $\mathcal{K}_{\min}$)**.** *Under Assumption 2.1, one has*

$$\{\nabla L_0(\theta) : L_0 \in \mathcal{K}_{\min}\} = \{\xi \in \mathbb{R}^d : \|\xi\|_2 \leq \psi(\|\theta\|_2 + D)\} \text{ for every } \theta \in \mathbb{R}^d. \tag{14}$$

Theorem 3.4 (Multivalued minority aware zone)**.** *Under Assumptions 2.1 and 3.1, one has, for every $\theta \in \mathbb{R}^d$,*

$$\mathcal{Q}(\theta) = \sup_{v \in V_L(\theta)} \Big(\langle v, \nabla L(\theta) \rangle - \psi(\|\theta\|_2 + D)\|v\|_2 \Big). \tag{15}$$

Consequently, $\theta \in Z_{\min} \iff \sup_{v \in V_L(\theta)} \Big(\langle v, \nabla L(\theta) \rangle - \psi(\|\theta\|_2 + D)\|v\|_2 \Big) < 0.$

Lemma 3.3 and Theorem 3.4 are proved in Section B.1. Next, we define two core quantities. For a norm $\|\cdot\|$ on $\mathbb{R}^d$, set

$$m_{\|\cdot\|}(\theta) := \inf_{v \in \rho\left(\operatorname{grad}_{\|\cdot\|} L(\theta)\right)} \|v\|_2, \quad c_{\|\cdot\|} := \inf_{v \in \mathbb{R}^d \setminus \{0\}} \frac{\|v\|_2}{\|v\|}. \tag{16}$$

Using Theorem 3.4 , we may now derive the explicit expressions of the minority aware zones. Results are gathered in Table 1 and details can be found in Corollary B.2 and Section B.1. Here crit denotes the set of critical points, $|\mathrm{supp}(\cdot)| = \|\cdot\|_0$ and $\mathrm{rank}(\cdot)$ is the rank of a matrix.

Dynamics	**Minority zones** $Z_{\min}$ $\theta \notin \mathrm{crit}\, L$
$V_L(\theta) = \rho(\operatorname{grad}_{\|\cdot\|} L(\theta))$	$\|\mathrm{d}L(\theta)\|_* < \psi m_{\|\cdot\|}(\theta)(\|\theta\|_2 + D)$.
$V_L(\theta) = \operatorname{grad}_{\|\cdot\|} L(\theta)$	$\|\mathrm{d}L(\theta)\|_*^2 < \psi m_{\|\cdot\|}(\theta)\|\mathrm{d}L(\theta)\|_*(\|\theta\|_2 + D)$
Gradient descent	$\|\nabla L(\theta)\|_2 < \psi(\|\theta\|_2 + D)$
Gradient clipping	$\|\nabla L(\theta)\|_2 < \psi(\|\theta\|_2 + D)$
Sign descent	$\|\nabla L(\theta)\|_1 < \psi\sqrt{\|\mathrm{supp}(\nabla L(\theta))\|}(\|\theta\|_2 + D)$
Spectral descent	$\|\nabla L(\Theta)\|_{\mathrm{nuc}} < \psi\sqrt{\mathrm{rank}(\nabla L(\Theta))}(\|\Theta\|_F + D)$
Newton descent	$\left\langle \nabla^2 L(\theta)^{-1}\nabla L(\theta), \nabla L(\theta) \right\rangle < \psi\|\nabla^2 L(\theta)^{-1}\nabla L(\theta)\|_2(\|\theta\|_2 + D)$ $(*)$

Table 1: Zoology of minority aware zones for different dynamics. $(*)$ Strongly convex loss.

4 Entry time and group minority awareness

Definition 4.1 (Entry time)**.** *Under Assumptions 2.1 and 3.1, let θ be an absolutely continuous solution to*

$$\theta(0) = \theta_{\mathrm{init}}, \qquad \dot{\theta}(t) \in -V_L(\theta(t)) \quad \text{for a.e. } t \geq 0.$$

The entry time along the trajectory θ is defined as $t_{\,\mathrm{entry}}(\theta) := \inf\{t \geq 0 : \theta(t) \in Z_{\min}\} \in [0, +\infty]$. When θ is unique, we may write $t_{\,\mathrm{entry}}(\theta) = t_{\,\mathrm{entry}}(\theta_{\mathrm{init}})$.

Entry time is a critical marker that structures the learning process in phases. Before $t_{\,\mathrm{entry}}$, there is a *majority dominance* phase: majority traits are learned independently of the minority structure, see Lemma 3.1 and Lemma A.4. Then when the trajectory enters $Z_{\min}$, some minority structures start being visible, see Lemma A.6. It is quite easy to see that the finiteness of $t_{\,\mathrm{entry}}(\theta)$ does not mean that the trajectory stays indefinitely in the minority aware zone, it may go out, but then we can provide similar estimates with new initializations. Entry times serve as theoretical indicators for the onset of majority-adverse behavior as described in Bachoc et al. (2025).

Theoretical bounds for entry time Recall $m_{\|\cdot\|}$ in Equation 16, and for a trajectory θ, consider the $m_{\|\cdot\|}$*-average along* θ:

$$\bar{m}(t) := \frac{1}{t}\int_0^t m_{\|\cdot\|}(\theta(s))\, ds \quad (t > 0), \tag{17}$$

$$\bar{M}_\theta := \frac{1}{t_{\,\mathrm{entry}}(\theta)}\int_0^{t_{\,\mathrm{entry}}(\theta)} m_{\|\cdot\|}(\theta(s))\, ds \text{ whenever } 0 < t_{\,\mathrm{entry}}(\theta) < \infty. \tag{18}$$

Assumption 4.1 (Quadratic growth at infinity)**.** *Let the constant D be the same as in the definition of $\mathcal{K}_{\min}$ (see Equation 3). Assume that there exists a positive constant $\lambda_{\max} > 0$ such that $\forall \theta \in \mathbb{R}^d, \quad L(\theta) - L_\star \leq \lambda_{\max}(\|\theta\|_2 + D)^2$.*

Theorem 4.1 (Entry time for normalized norm-gradient)**.** *Under Assumptions 2.1 and 3.1, let θ solve*

$$\dot{\theta}(t) \in -\rho(\operatorname{grad}_{\|\cdot\|} L(\theta(t))), \qquad \theta(0) = \theta_{\text{init}}.$$

If $\nabla L(\theta_{\text{init}}) \neq 0$, then $t_{\text{entry}}(\theta) \leq \frac{L(\theta_{\text{init}}) - L_\star}{\psi D \bar{M}_\theta}$. If Assumption 4.1 also holds and $t_{\text{entry}}(\theta) > 0$, then $t_{\text{entry}}(\theta) \leq \frac{2\sqrt{\lambda_{\max}}}{\psi \bar{M}_\theta} \sqrt{L(\theta_{\text{init}}) - L_\star}$.

Theorem 4.2 (Entry time for norm-gradient)**.** *Under Assumptions 2.1 and 3.1, let θ solve*

$$\dot{\theta}(t) \in -\operatorname{grad}_{\|\cdot\|} L(\theta(t)), \qquad \theta(0) = \theta_{\text{init}}.$$

If $\nabla L(\theta_{\text{init}}) \neq 0$, one has $t_{\text{entry}}(\theta) \leq \frac{L(\theta_{\text{init}}) - L_\star}{\psi^2 D^2 \bar{M}_\theta^2}$. If Assumption 4.1 also holds, and $t_{\text{entry}}(\theta) > 0$, then

$$t_{\text{entry}}(\theta) \leq \frac{\lambda_{\max}}{\psi^2 \bar{M}_\theta^2} \begin{cases} \frac{L(\theta_{\text{init}}) - L_\star}{\lambda_{\max} D^2} & \text{if } \frac{L(\theta_{\text{init}}) - L_\star}{\lambda_{\max} D^2} \leq e \\ e \log\left(\frac{L(\theta_{\text{init}}) - L_\star}{\lambda_{\max} D^2}\right) & \text{if } \frac{L(\theta_{\text{init}}) - L_\star}{\lambda_{\max} D^2} > e \end{cases}.$$

By applying Theorems 4.1 and 4.2 for the specific optimization dynamics introduced in Section 3.1, we derive various upper bounds of entry time presented in Table 2 (see details in Sections B.2.2, B.2.3, B.2.4 and B.2.5).

Dynamics	**Entry times in general**	**Entry times with** Assumption 4.1
Gradient descent	$t^{\text{GD}}_{\text{entry}} = \mathcal{O}\left(\frac{L(\theta_{\text{init}}) - L_\star}{\psi^2 D^2}\right)$ $(\star)$	$t^{\text{GD}}_{\text{entry}} = \mathcal{O}\left(\frac{\lambda_{\max}}{\psi^2} \log(\frac{L(\theta_{\text{init}}) - L_\star}{\lambda_{\max} D^2})\right)$
Gradient clipping	$t^{\text{GC}}_{\text{entry}} = \mathcal{O}\left(\frac{L(\theta_{\text{init}}) - L_\star}{\psi D \min\{\tau_{\text{clip}}, \psi D\}}\right)$	$t^{\text{GC}}_{\text{entry}} = \mathcal{O}\left(\frac{\sqrt{\lambda_{\max}(L(\theta_{\text{init}}) - L_\star)}}{\psi \min\{\tau_{\text{clip}}, \psi D\}}\right)$
Sign descent	$t^{\text{SD}}_{\text{entry}} = \mathcal{O}\left(\frac{L(\theta_{\text{init}}) - L_\star}{\psi D \bar{M}_{\text{SD}}}\right)$ $(\star)$	$t^{\text{SD}}_{\text{entry}} = \mathcal{O}\left(\frac{\sqrt{\lambda_{\max}(L(\theta_{\text{init}}) - L_\star)}}{\psi \bar{M}_{\text{SD}}}\right)$
Spectral descent	$t^{\text{Spec}}_{\text{entry}} = \mathcal{O}\left(\frac{L(\Theta_{\text{init}}) - L_\star}{\psi D \bar{R}}\right)$ $(\star)$	$t^{\text{Spec}}_{\text{entry}} = \mathcal{O}\left(\frac{\sqrt{\lambda_{\max}(L(\Theta_{\text{init}}) - L_\star)}}{\psi \bar{R}}\right)$
Newton descent	$t^{\text{Newton}}_{\text{entry}} = \mathcal{O}\left(\log(\frac{\|\nabla L(\theta_{\text{init}})\|_2}{\psi D})\right)$	$t^{\text{Newton}}_{\text{entry}} = \mathcal{O}\left(\log(\frac{\|\nabla L(\theta_{\text{init}})\|_2}{\psi D})\right)$

Table 2: Zoology of entry times for different dynamics (see also Table 4 in Section B.2). The term $\bar{M}_{\text{SD}}$ denotes the time-average of square root of the support size of ∇L prior to the entry time, and $\bar{R}$ denotes the time-average of square root of rank of ∇L over the same period. For the entry times in $(\star)$, we also provide lower bounds, that is, examples of L matching the upper bounds in Section B.3.

The relevance of entry times regarding the optimization of the actual geometries is also discussed in a particular case in Section E. Within our continuous-time idealization of the small-step regime, Table 2 highlights two ways in which optimizer geometry can facilitate departure from majority dominance:
— The entry-time bounds for sign and spectral descent scale as $\mathcal{O}(\psi^{-1})$, rather than $\mathcal{O}(\psi^{-2})$ for gradient descent. This suggests that normalization can mitigate the delay associated with increasingly weak minority components. Newton descent exhibits very weak dependence, of order $\log(1/\psi)$. In conclusion, the more severe the imbalance, the larger the advantage of normalized geometries of the optimizers.
— The factors $\bar{M}_{\text{SD}}$ and $\bar{R}$ may further tighten the sign and spectral descent bounds when trajectories maintain large gradient support or rank before entry. We note that comparing upper bounds does not establish that sign or spectral descent always enters earlier than GD on the same problem.

As a conclusion, when imbalance is suspected but group labels are unavailable —hidden subclasses, latent data, rare patterns— reweighting and resampling cannot be applied directly. Our results identify optimizer choice as a practical response: normalized and spectral geometries yield minority-awareness time bounds with a milder dependence on minority amplitude than Euclidean gradient descent. Together with the experiments in Section 5, these findings support favoring AdamW or Muon over vanilla SGD when training under imbalance, without requiring explicit identification of minority groups.

5 NUMERICAL EXPERIMENTS

We first provide a sanity check of our t_{entry} bounds on quadratic problems. We then test whether normalized gradient methods yield more balanced outcomes as our theory suggests. In practical deep learning settings, theoretical quantities like ψ and t_{entry} are intractable, we thus measure empirical accuracy and loss to assess performances with respect to imbalance.

Entry times through linear regression. In this experiment, we seek empirical scaling relations linking entry times to minority amplitude ψ and dimension d. Using the Superconductor and YearPredictionMSD datasets, we apply principal component analysis (PCA) to construct a coherent family of linear-regression problems across dimensions. Within each problem, we vary the minority-group size, yielding different amplitudes $\psi := \sigma_{\max}(\nabla^2 L_0)$. We then measure the entry times of gradient descent (GD) and sign descent (SD), and fit their ratio as a function of ψ and d using log–log regression. This yields the following empirical scaling relations for Superconductor (left) and YearPredictionMSD (right):

$$t_{\text{entry}}(\text{GD})/t_{\text{entry}}(\text{SD}) \propto d^{0.65}/\psi^{0.90}, \qquad t_{\text{entry}}(\text{GD})/t_{\text{entry}}(\text{SD}) \propto d^{0.51}/\psi^{0.37}.$$

These estimates show that the entry time of gradient descent effectively grows faster than that of Sign Descent in high-dimensional ($d >> 1$) and highly imbalanced ($\psi << 1$) regimes. Detailed experimental protocols are provided in Section C.1.

Hidden subclass classification - Forest Covertype. We predict forest cover type from 54 cartographic variables ($\sim 581,000$ samples). Following Collobert & Bengio (2001), we use balanced binary supervision (Lodgepole Pine vs. all other classes). In this formulation, the *Aspen* class forms the hidden minority, comprising just 1.6% of the dataset. Our results demonstrate that AdamW's early learning advantage over SGD is significantly magnified for this minority class (63.93% vs. 51.40%) compared to the majority class (90.99% vs 88.25%). This phenomenon also occurs below.

Catch-up time with CIFAR-10. To facilitate direct comparison with prior work, we replicate the experimental setup of Bachoc et al. (2025), wherein the *car* class is subsampled to form a minority group.

Metrics	AdamW	SGD	Muon
Average L_0 accuracy	91.82%	86.04%	96.48%
Average L_1 accuracy	95.45%	92.92%	98.39%

Table 3: Average accuracy over iterations $1-30$. This evaluation window was chosen because both L_0 and L_1 accuracies reach 99% after iteration 40 in our simulations.

Language modelling - SmallGPT2 on WikiText-103. Adopting the setup of Kunstner et al. (2024), we train a SmallGPT-2 model from scratch with AdamW, SGD, and Muon, tracking both majority and minority (the 5% rarest tokens) losses (Figure 2).

While the majority loss steadily decreases across all optimizers, the minority loss trajectories diverge significantly. The AdamW and Muon minority losses initially spike before decreasing, whereas the SGD minority loss strictly increases throughout the experiment. This behavior corroborates our theoretical finding that standard gradient dynamics inherently delay optimization of the minority signal L_0. The setting of the experiments and detailed results are given in Table 10.

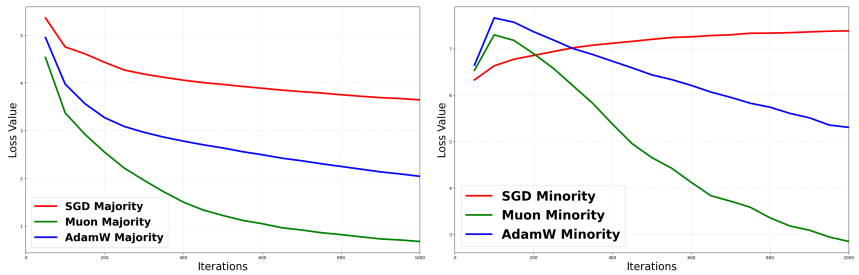


Figure 2: Evolution of training loss of SmallGPT2 from scratch on WikiText-103.

AI USE STATEMENT

In this work, we used generative AI tools to formulate mathematical claims, to assist in the writing of proofs, to provide critical ingredients for proving mathematical claims and to assist with translation. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods or interpret results. Generating synthetic data sets, cleaning and reformatting datasets to support qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for transcribe recording of research material, summarizing or analyzing existing literature and editing the paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We are committed to the reproducibility of our findings. Comprehensive details regarding the setup and execution of our numerical experiments are provided in the text. Theoretically, rigorous and complete proofs for all claims, theorems, and bounds can be found in the Appendix.

REFERENCES

Sercan O Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6679–6687, 2021.

Jean-Pierre Aubin and Arrigo Cellina. *Differential Inclusions: Set-Valued Maps and Viability Theory*, volume 264 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 1984.

François Bachoc, Jérôme Bolte, Ryan Boustany, and Jean-Michel Loubes. When majority rules, minority loses: Bias amplification of gradient descent. In *Advances in Neural Information Processing Systems 38*, 2025. URL https://openreview.net/forum?id=RDe4Ntw2oy. NeurIPS 2025 poster.

Ronan Collobert and Samy Bengio. Svmtorch: support vector machines for large-scale regression problems. *J. Mach. Learn. Res.*, 1:143–160, September 2001. ISSN 1532-4435. doi: 10.1162/15324430152733142. URL https://doi.org/10.1162/15324430152733142.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18932–18943, 2021.

Yinze Hu, Hongjun Xiang, Xingao Gong, and Hongyu Yu. A physical response-and-memory model for muon optimization, 2026. URL https://arxiv.org/abs/2608.22994.

Keller Jordan et al. Muon: An optimizer for hidden layers in neural networks, 2024.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1910.09217.

Andrej Karpathy. nanoGPT: The simplest, fastest repository for training/finetuning medium-sized GPTs. https://github.com/karpathy/nanoGPT, 2023.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2564–2572. PMLR, 2018. URL https://proceedings.mlr.press/v80/kearns18a.html.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Frederik Kunstner and Francis Bach. Scaling laws for gradient descent and sign descent for linear bigram models under Zipf's law. In *Advances in Neural Information Processing Systems 38*, 2025. URL https://openreview.net/forum?id=VUbwLjLkws. NeurIPS 2025 poster.

Frederik Kunstner, Alan Milligan, Robin Yadav, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why Adam outperforms gradient descent on language models. In *Advances in Neural Information Processing Systems 37*, volume 37, 2024. doi: 10.52202/079017-0948. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/350e718ff74062b4bac2c6ffd9e1ac66-Abstract-Conference.html.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2537–2546, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Liu_Large-Scale_Long-Tailed_Recognition_in_an_Open_World_CVPR_2019_paper.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL)*, pp. 151–159, 2020. URL https://doi.org/10.1145/3368555.3384468.

Konstantinos Oikonomidis, Jan Quan, Kimon Antonakopoulos, Antonio Silveti-Falls, Volkan Cevher, and Panagiotis Patrinos. Constrained stochastic spectral preconditioning converges for nonconvex objectives, 2026. URL https://arxiv.org/abs/2605.11850.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49069–49104. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/pethick25a.html.

Thomas Pethick, Wanyun Xie, Mete Erdogan, Kimon Antonakopoulos, Antonio Silveti-Falls, and Volkan Cevher. Generalized gradient norm clipping and non-euclidean (l_0, l_1)-smoothness, 2026. URL https://arxiv.org/abs/2506.01913.

Steven T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130, 2014. doi: 10.3758/s13423-014-0585-6. URL https://link.springer.com/article/10.3758/s13423-014-0585-6.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: a better, faster, scalable, and open tabular foundation model, 2026. URL https://arxiv.org/abs/2602.11139.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

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

Nimit Sohoni, Jared A. Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Advances in Neural Information Processing Systems 33*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e0688d13958a19e087e123148555e4b4-Abstract.html.

Ben S. Southworth, Shuai Jiang, Daniel McBride, Eric C. Cyr, and Stephen Thomas. Muon in vision transformers: Optimizer-recipe interactions and gradient spectra, 2026. URL https://arxiv.org/abs/2605.24770.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How muon's spectral design benefits generalization: A study on imbalanced data, 2025. URL https://arxiv.org/abs/2510.22980.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Y. F. Tan. Muon outperforms Adam in tail-end associative memory learning, 2025. URL https://arxiv.org/abs/2509.26030.

Yifeng Wang et al. Tabkanet: An efficient Kolmogorov-Arnold network for tabular data. *arXiv preprint arXiv:2405.10000*, 2024.

G.A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(92)90407-2. URL https://www.sciencedirect.com/science/article/pii/0024379592904072.

A Auxiliary results and comments

A.1 General comments

We summarize here several key aspects of our approach, with the aim of delineating what has been established and what remains to be investigated in a few points.

— We introduce a majority–minority model under limited prior knowledge. Within this framework, it is difficult to guarantee that all minority components will eventually be learned, since this also depends on the initialization. Nevertheless, by detecting the onset of minority learning, we obtain an informative indicator of when minority components begin to emerge. Our key contributions are an equation for a minority awareness zone for most optimizers and estimates for entry times.

This reconnects with the work of Bachoc et al. (2025) on the origin of bias of gradient methods and dominance of a majority learning region.

— We work with idealized dynamics rather than genuine optimizers. The purpose is to retain the core vector fields governing the dynamics while stripping away secondary algorithmic effects.

— We restrict attention to a class of functions arising from deep-learning losses, for which the relevant step sizes are small but vary over a moderate range, and for which the amplitude factor κ introduced in Section 1 remains modest. Thus, our analysis is intended to capture a regime that is simplified but still representative of the scales encountered in *deep learning* problems.

— All else being equal, entry times are comparable across the dynamics we consider. They therefore provide an indication, for a given initial condition, of when majority dominance begins to cease.

— Among the scaled optimizers considered, three clearly stand out: Adam, Muon and Newton's method. This suggests a promising direction, but the phenomenon remains to be understood more systematically.

Many aspects remain to be investigated. In particular:

— extending the analysis from idealized dynamics to genuine optimizers;

— developing more precise markers for minority detection,

— understanding more systematically why certain scaled optimizers exhibit markedly different minority-learning behavior.

A.2 Results on differential inclusions

Lemma A.1 (Aubin–Cellina local existence theorem)**.** *Let $\Gamma : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map with nonempty compact convex values. Assume that Γ has a closed graph and is locally bounded. Then, for every $\theta_{\mathrm{init}} \in \mathbb{R}^d$, there exist $T_{\max} > 0$ and an absolutely continuous map $\theta : [0, T_{\max}) \to \mathbb{R}^d$ such that*

$$\theta(0) = \theta_{\mathrm{init}}, \qquad \dot{\theta}(t) \in -\Gamma(\theta(t)) \quad \textit{for a.e. } t \in [0, T_{\max}).$$

This is a standard consequence of the existence theory for differential inclusions; see, for instance, (Aubin & Cellina, 1984, Chapter 2, Section 1).

Corollary A.2 (Global existence for norm-gradient systems)**.** *Under Assumption 2.1, consider any of the following dynamical systems presented in Section 3.1: gradient descent, gradient clipping, sign descent, spectral descent and Newton descent (in this last case, assume in addition that Equation 10 holds). Then:*

(i) any of this dynamical system has at least one solution,

(ii) every maximal absolutely continuous solution of any of the above is global, that is,

$$T_{\max} = +\infty.$$

Proof. Existence. The multifunction

$$\ell \mapsto \arg\max_{\|z\| \leq 1} \ell[z]$$

has nonempty compact convex values, because it is the set of maximizers of a linear form over the compact convex unit ball of $\|\cdot\|$. Since $\theta \mapsto \mathrm{d}L(\theta)$ is continuous, both

$$\theta \mapsto \rho(\mathrm{grad}_{\|\cdot\|} L(\theta)) \qquad \text{and} \qquad \theta \mapsto \mathrm{grad}_{\|\cdot\|} L(\theta)$$

are upper semicontinuous and locally bounded. Hence Lemma A.1 provides a local absolutely continuous solution for both gradient and normalized gradient methods. Note that for Newton's method we may apply instead Arzela existence theorem (without uniqueness) and for gradient clipping it is more natural to apply Cauchy-Lipschitz.

Maximal solutions. We argue for gradient dynamics in an arbitrary norm $\|\cdot\|$, the other methods follow by similar arguments.

Let $\theta : [0, T_{\max}) \to \mathbb{R}^d$ be a maximal solution. For a.e. t, write

$$\dot{\theta}(t) \in -\mathrm{grad}_{\|\cdot\|} L(\theta(t)),$$

so

$$\frac{d}{dt} L(\theta(t)) = \mathrm{d}L(\theta(t))[\dot{\theta}(t)] = -\|\mathrm{d}L(\theta(t))\|_*^2. \tag{19}$$

As a consequence, for all $t \in [0, T_{\max})$

$$\int_0^t \|\mathrm{d}L(\theta(s))\|_*^2 ds \leq L(\theta_{\mathrm{init}}) - L_\star.$$

Thus if $T_{\max}$ was finite we would have a constant $C > 0$ such that

$$\int_0^{T_{\max}} \|\dot{\theta}(t)\|_2^2 dt \leq C \int_0^{T_{\max}} \|\mathrm{d}L(\theta(t))\|_*^2 dt \leq C \left(L(\theta_{\mathrm{init}}) - L_\star\right).$$

By Cauchy-Schwarz inequality, for $0 \leq s < t \leq T_{\max}$,

$$\|\theta(t) - \theta(s)\|_2 = \int_s^t \|\dot{\theta}(\tau)\|_2 d\tau \leq \sqrt{|t-s|} \sqrt{\int_0^t \|\dot{\theta}(\tau)\|_2^2 d\tau} \leq \sqrt{C|t-s|} \sqrt{L(\theta_{\mathrm{init}}) - L_\star}.$$

Hence we see that θ has a limit at $t = T_{\max}$ which contradicts the maximality of $T_{\max}$.

A similar reasoning provides that $T_{\max} = +\infty$ for Newton's method or gradient clipping. In the normalized case, the proof is even more direct because the velocity is globally bounded and the Cauchy property derived above boils down to

$$\|\theta(t) - \theta(s)\| \leq |t - s|$$

for all t, s in $[0, T_{\max})$.

In all cases uniqueness is not required. □

Absence of finite time convergence. While normalized gradient descent can lead in a finite time towards a critical point, this is not possible for the normal gradient descent despite multivaluedness.

Proposition A.1 (Absence of finite time convergence for the normal gradient descent)**.** *Under Assumption 2.1, let θ be an absolutely continuous solution to*

$$\dot{\theta}(t) \in -\operatorname{grad}_{\|\cdot\|} L(\theta(t))$$

with $dL(\theta_{\text{init}}) \neq 0$. Then for all time $t > 0$, $dL(\theta(t)) \neq 0$.

Proof. Suppose, toward a contradiction, that there exists $t_0 > 0$ such that

$$dL(\theta(t_0)) = 0, \text{i. e., } \nabla L(\theta(t_0)) = 0.$$

By the definition of the norm-gradient and norm equivalence in finite dimension, there exists a constant $C > 0$ such that for a.e. t,

$$\|\dot{\theta}(t)\|_2 \leq C\|\nabla L(\theta(t))\|_2.$$

Moreover, since L is C^2 and $\theta([0, t_0])$ is compact, ∇L is Lipschitz on a neighbourhood of $\theta([0, t_0])$. Let $\Lambda > 0$ be a corresponding Lipschitz constant.

For every $t \in [0, t_0]$, using $\nabla L(\theta(t_0)) = 0$, we get

$$\|\nabla L(\theta(t))\|_2 = \|\nabla L(\theta(t)) - \nabla L(\theta(t_0))\|_2 \leq \Lambda\|\theta(t) - \theta(t_0)\|_2 \leq \Lambda \int_t^{t_0} \|\dot{\theta}(s)\|_2 \, ds \leq \Lambda C \int_t^{t_0} \|\nabla L(\theta(s))\|_2 \, ds.$$

Define

$$g(t) := \|\nabla L(\theta(t))\|_2.$$

Then $g \geq 0$ is continuous and satisfies

$$g(t) \leq \Lambda C \int_t^{t_0} g(s) \, ds \qquad \text{for all } t \in [0, t_0].$$

Hence

$$g(t_0 - (t_0 - t)) \leq \Lambda C \int_0^{t_0 - t} g(t_0 - s) \, ds \qquad \text{for all } t \in [0, t_0].$$

Thus, by Gronwall's lemma, $g(t) = 0$ for all $t \in [0, t_0]$. In particular, $\nabla L(\theta(0)) = 0$, that is, $dL(\theta_{\text{init}}) = 0$ which is impossible. □

A.3 Bounds

We recall the definition of the dual norm

$$\|\ell\|_* := \max_{\|z\| \leq 1} \ell[z].$$

Lemma A.3 (Smoothness gap bounds for an arbitrary norm)**.** *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$, and let $\|\cdot\|_*$ be the corresponding operator norm on linear forms. Consider $h \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ which is L-smooth with respect to $\|\cdot\|$, i.e, with $L > 0$,*

$$\forall x, y \in \mathbb{R}^d, \qquad h(x) - h(y) - \mathrm{d}h(y)[x - y] \leq \frac{L}{2}\|x - y\|^2. \tag{20}$$

Assume that h reaches its minimum at $\theta_\star$. Then, for every $\theta \in \mathbb{R}^d$,

$$\frac{1}{2L}\|\mathrm{d}h(\theta)\|_*^2 \leq h(\theta) - h(\theta_\star) \leq \frac{L}{2}\|\theta - \theta_\star\|^2.$$

Proof. For the upper bound, apply the smoothness inequality, Equation 20, with $x = \theta$ and $y = \theta_\star$. Since $\theta_\star$ is a minimizer and $h \in C^1$, one has $\mathrm{d}h(\theta_\star) = 0$, so

$$h(\theta) - h(\theta_\star) \leq \frac{L}{2}\|\theta - \theta_\star\|^2.$$

For the lower bound, fix $\theta \in \mathbb{R}^d$. By compactness of the unit $\|\cdot\|$-ball and continuity of $\mathrm{d}h(\theta)$, there exists $v_\theta \in \mathbb{R}^d$ such that

$$\|v_\theta\| = 1, \qquad \mathrm{d}h(\theta)[v_\theta] = \|\mathrm{d}h(\theta)\|_*.$$

Apply the smoothness inequality with

$$y = \theta, \qquad x = \theta - \frac{\|\mathrm{d}h(\theta)\|_*}{L} v_\theta.$$

Then

$$\begin{aligned} h\left(\theta - \frac{\|\mathrm{d}h(\theta)\|_*}{L} v_\theta\right) &\leq h(\theta) - \frac{\|\mathrm{d}h(\theta)\|_*}{L} \mathrm{d}h(\theta)[v_\theta] + \frac{L}{2} \frac{\|\mathrm{d}h(\theta)\|_*^2}{L^2} \|v_\theta\|^2 \\ &= h(\theta) - \frac{1}{2L} \|\mathrm{d}h(\theta)\|_*^2. \end{aligned}$$

Since $h(\theta_\star) \leq h(x)$ for every x, this yields

$$\frac{1}{2L} \|\mathrm{d}h(\theta)\|_*^2 \leq h(\theta) - h(\theta_\star).$$

□

A.4 Extra-material

Proposition A.2 (Subdifferential of the dual norm)**.** *Let E be a finite-dimensional real vector space, and let $\|\cdot\|$ be a norm on E. Its dual norm is defined by*

$$\|V\|_* = \max_{\|Z\| \leq 1} \langle V, Z \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the duality bracket.[2] *Then, $\partial \|V\|_* = \operatorname*{argmax}_{\|Z\| \leq 1} \langle V, Z \rangle$, for every V in E^*. Equivalently,*

$$Z \in \partial \|V\|_* \quad \Longleftrightarrow \quad \|Z\| \leq 1 \ \text{ and } \ \langle V, Z \rangle = \|V\|_*.$$

Proof. Recall that the subdifferential of the convex function $V \mapsto \|V\|_*$ is defined by

$$Z \in \partial \|V\|_*$$

if and only if, for every $Y \in E^*$,

$$\|Y\|_* \geq \|V\|_* + \langle Y - V, Z \rangle.$$

We prove both inclusions. First, assume

$$Z \in \operatorname*{argmax}_{\|X\| \leq 1} \langle V, X \rangle,$$

so that

$$\|Z\| \leq 1 \text{ and } \langle V, Z \rangle = \|V\|_*.$$

For any $Y \in E^*$, by definition of the dual norm,

$$\langle Y, Z \rangle \leq \|Y\|_* \|Z\| \leq \|Y\|_*. \tag{21}$$

Using $\langle V, Z \rangle = \|V\|_*$, we get

$$\langle Y, Z \rangle = \langle V, Z \rangle + \langle Y - V, Z \rangle = \|V\|_* + \langle Y - V, Z \rangle.$$

Hence by Equation 21

$$\|Y\|_* \geq \|V\|_* + \langle Y - V, Z \rangle, \text{ i.e., } Z \in \partial \|V\|_*.$$

Conversely, assume

$$Z \in \partial \|V\|_*.$$

[2] In our setting E is identified with its dual with the Euclidean scalar product (Frobenius-like for matrices). So, after identification the bracket is actually the standard scalar product.

Then, for every $Y \in E^*$,

$$\|Y\|_* \geq \|V\|_* + \langle Y - V, Z \rangle.$$

Taking $Y = 0$, we obtain $\langle V, Z \rangle \geq \|V\|_*$. On the other hand, the dual norm always satisfies $\langle V, Z \rangle \leq \|V\|_* \|Z\|$.

To conclude we show that $\|Z\| \leq 1$. Let $Y = tH$, where $t > 0$ and $H \in E^*$. The subgradient inequality gives

$$t\|H\|_* \geq \|V\|_* + \langle tH - V, Z \rangle.$$

That is,

$$t\|H\|_* \geq \|V\|_* - \langle V, Z \rangle + t\langle H, Z \rangle.$$

Divide by t:

$$\|H\|_* \geq \frac{\|V\|_* - \langle V, Z \rangle}{t} + \langle H, Z \rangle.$$

Letting $t \to +\infty$, we get

$$\langle H, Z \rangle \leq \|H\|_* \qquad \text{for every } H \in E^*.$$

Taking the supremum over all H with $\|H\|_* \leq 1$, we obtain

$$\|Z\| \leq 1,$$

because $\|\cdot\|$ is the dual norm of $\|\cdot\|_*$ in finite dimension.

Therefore

$$\|Z\| \leq 1.$$

By definition of $\|V\|_*$, this implies

$$\langle V, Z \rangle \leq \|V\|_*.$$

Together with the earlier inequality

$$\langle V, Z \rangle \geq \|V\|_*,$$

we conclude that

$$\langle V, Z \rangle = \|V\|_*.$$

Hence

$$Z \in \operatorname*{argmax}_{\|X\| \leq 1} \langle V, X \rangle.$$

This proves the equality $\partial \|V\|_* = \operatorname*{argmax}_{\|Z\| \leq 1} \langle V, Z \rangle$. □

Remark A.1. *In order to have the convenient variational representation in Equation 13 of $Z_{\min}$, our Definition 3.1 differs slightly from the corresponding definition in Bachoc et al. (2025) (where only univalued V_L are considered), by using a strict inequality. Observe that $Z_{\mathrm{maj-adv}}(L_0)$ could be empty in some cases —for instance when $L_0 = 0$ or $L_0 = -\epsilon L$ with $\epsilon > 0$ small.*

A.5 Spectral Descent

Definition A.1. *Let $a, b \in \mathbb{N}$ and M be an $a \times b$ non-zero matrix. A compact eigenvalue decomposition of M is a decomposition $M = U\Sigma V^\top$ where U is $a \times r$, with $r \in \mathbb{N}$, $r \leq \min(a, b)$, with $U^\top U = I_r$, where Σ is $r \times r$ and diagonal with non-zero diagonal values, and where V is $b \times r$ with $V^\top V = I_r$.*

Assume the parameter space is of the form $\mathbb{R}^{a \times b}$. The differential is represented by the matrix gradient through

$$\mathrm{d}L(\Theta)[H] = \mathrm{tr}(\nabla L(\Theta)^\top H) \qquad \forall \Theta \in \mathbb{R}^{a \times b} \qquad \forall H \in \mathbb{R}^{a \times b}.$$

The matrix Spectral norm (w.r.t. to Euclidean norm) is denoted by $\|\cdot\|_{\mathrm{sp}}$, its dual norm –the nuclear norm– is denoted by $\|\cdot\|_{\mathrm{nuc}}$. Let us denote by $\mathcal{S} = \partial \|\cdot\|_{\mathrm{nuc}}$ the subdifferential mapping of the nuclear norm. From classical results, see Proposition A.2, one has,

$$\mathcal{S}(M) = \operatorname*{arg\,max}_{\|Z\|_{\mathrm{sp}} \leq 1} \mathrm{tr}(M^\top Z) \text{ for all } M \text{ in } \mathbb{R}^{a \times b}, \tag{22}$$

and, using (Watson, 1992, Example 2, Section 2), if $M = U\Sigma V^\top$ is a compact singular value decomposition (see Definition A.1), then

$$\mathcal{S}(M) = \left\{UV^\top + W : U^\top W = 0,\ WV = 0,\ \|W\|_{\mathrm{sp}} \leq 1\right\}.$$

Therefore, by definition of normalized gradients,

$$\rho(\mathrm{grad}_{\|\cdot\|_{\mathrm{sp}}} L(\Theta)) = \mathcal{S}(\nabla L(\Theta)). \tag{23}$$

Spectral descent is the corresponding normalized spectral-norm descent:

$$\dot{\Theta}(t) \in -\rho(\mathrm{grad}_{\|\cdot\|_{\mathrm{sp}}} L(\Theta(t))). \tag{24}$$

Since $\Theta \mapsto \mathcal{S}(\nabla L(\Theta))$ has nonempty compact convex and bounded values and closed graph, standard existence results for differential inclusions again apply to Equation 9 (Lemma A.1 and corollary A.2).

Note that the (unnormalized) spectral-norm descent would be

$$\dot{\Theta}(t) \in -\,\mathrm{grad}_{\|\cdot\|_{\mathrm{sp}}} L(\Theta(t)) \Leftrightarrow \dot{\Theta}(t) \in -\|\nabla L(\Theta(t))\|_{\mathrm{nuc}}\, \rho(\mathrm{grad}_{\|\cdot\|_{\mathrm{sp}}} L(\Theta(t)))$$

with

$$v \in -\,\mathrm{grad}_{\|\cdot\|_{\mathrm{sp}}} L(\Theta) \implies \|v\|_{\mathrm{sp}} = \|\nabla L(\Theta)\|_{\mathrm{nuc}}.$$

A.6 Majority adverse and Minority awareness zones

Our definition of a majority adverse zone differs slightly from the one in Bachoc et al. (2025) because of the strict inequality (Remark A.1). As a result, in our setting the minority aware zone $Z_{\min}$ typically has nonempty interior, see Corollary B.1. This is a benefit of its definition, compared to $Z_{\mathrm{maj-adv}}(L_0)$ which can be empty for some choices of L_0 and non-empty for other choices.

The next lemma shows that for an optimizer V_L satisfying Equations 5 and 6, membership in the majority adverse zone forces a strict decrease of the minority loss along the trajectory. A proof is given in Section B.1.

Lemma A.4. *Let θ be a solution to Equation 4. Assume that V_L is a descent field, i.e. Equations 5 and 6 hold and $Z_{\mathrm{maj-adv}}(L_0) \neq \emptyset$. Then, for every $L_0 \in \mathcal{K}_{\min}$, for a.e. $t \geq 0$, $\theta(t) \in Z_{\mathrm{maj-adv}}(L_0) \implies \dfrac{d}{dt} L_0(\theta(t)) < 0.$*

The next theorem shows uniform majority dominance out of $Z_{\min}$.

Theorem A.5 (Uniform majority dominance: the multivalued case)**.** *Under Assumptions 2.1 and 3.1, for every $\theta \in \mathbb{R}^d$,*

$$\theta \notin Z_{\min} \iff \exists v_\theta \in V_L(\theta),\ \forall L_1 \in \mathcal{K}_{\mathrm{maj}}(L),\ dL_1(\theta)[-v_\theta] \leq 0.$$

Thus, outside $Z_{\min}$, there exists a common admissible update that is majority-favouring to first order for every admissible decomposition.

If V_L is single-valued, this is the actual update. In particular, let $\theta : I \to \mathbb{R}^d$ be an absolutely continuous solution of $\dot{\theta}(t) = -V_L(\theta(t))$ on an interval I, with $\theta(t) \notin Z_{\min}$ for every $t \in I$. Then, for every $L_1 \in \mathcal{K}_{\mathrm{maj}}(L)$,

$$\frac{d}{dt} L_1(\theta(t)) \leq 0 \qquad \text{for almost every } t \in I,$$

and $L_1 \circ \theta$ is nonincreasing on I.

Proof. Fix θ and write $V = V_L(\theta)$. Define

$$C_\theta := \overline{\{\nabla L_1(\theta) : L_1 \in \mathcal{K}_{\mathrm{maj}}(L)\}}.$$

The set $\mathcal{K}_{\min}$ is nonempty and convex, and so is $\mathcal{K}_{\mathrm{maj}}(L)$. Moreover, for every $L_1 \in \mathcal{K}_{\mathrm{maj}}(L)$,

$$\|\nabla L_1(\theta) - \nabla L(\theta)\|_2 \leq \psi(\|\theta\|_2 + D).$$

Hence C_θ is nonempty, convex, and compact. The set V has the same properties by Assumption 3.1.

Suppose first that $\theta \notin Z_{\min}$. By the definitions of $Z_{\min}$ and $Z_{\mathrm{maj-adv}}(L_0)$,

$$\max_{v \in V} \langle v, \nabla L_1(\theta) \rangle \geq 0 \qquad \text{for every } L_1 \in \mathcal{K}_{\mathrm{maj}}(L).$$

Since $g \mapsto \max_{v \in V} \langle v, g \rangle$ is continuous, this inequality extends to every $g \in C_\theta$. The bilinear minimax theorem therefore gives

$$0 \leq \min_{g \in C_\theta} \max_{v \in V} \langle v, g \rangle = \max_{v \in V} \min_{g \in C_\theta} \langle v, g \rangle.$$

The last maximum is attained. Any maximizer v_θ satisfies

$$\langle v_\theta, g \rangle \geq 0 \qquad \text{for every } g \in C_\theta.$$

In particular, $dL_1(\theta)[-v_\theta] \leq 0$ for every $L_1 \in \mathcal{K}_{\mathrm{maj}}(L)$.

Conversely, suppose such a common direction v_θ exists. For every $L_0 \in \mathcal{K}_{\min}$, with $L_1 = L - L_0$,

$$\sup_{v \in V_L(\theta)} \langle v, \nabla L_1(\theta) \rangle \geq \langle v_\theta, \nabla L_1(\theta) \rangle \geq 0.$$

Thus $\theta \notin Z_{\mathrm{maj-adv}}(L_0)$ for every $L_0 \in \mathcal{K}_{\min}$, and therefore $\theta \notin Z_{\min}$.

Finally, when V_L is single-valued, the common direction is necessarily $v_\theta = V_L(\theta)$. Along the trajectory in the statement, the chain rule yields

$$\frac{d}{dt} L_1(\theta(t)) = dL_1(\theta(t))[-V_L(\theta(t))] \leq 0 \qquad \text{for almost every } t \in I.$$

Integrating proves the claimed monotonicity and, in particular, Lemma 3.1. □

Lemma A.6 (Minority activation)**.** *Under Assumptions 2.1 and 3.1, suppose that V_L satisfies the descent conditions (5)–(6). For every $\bar{\theta} \in Z_{\min}$, there exist an admissible minority loss $L_0 \in \mathcal{K}_{\min}$ and a neighbourhood U of $\bar{\theta}$ such that, writing $L_1 = L - L_0$,*

$$dL_0(\theta)[-v] < 0 < dL_1(\theta)[-v], \qquad \forall \theta \in U, \quad \forall v \in V_L(\theta).$$

Thus, every admissible update in U strictly favors the same minority loss while increasing its corresponding majority loss, to first order.[3]

Proof. Since $\bar{\theta} \in Z_{\min}$, the union defining $Z_{\min}$ provides $L_0 \in \mathcal{K}_{\min}$ such that $\bar{\theta} \in Z_{\mathrm{maj-adv}}(L_0)$. With $L_1 = L - L_0$, this means

$$\max_{v \in V_L(\bar{\theta})} \langle v, \nabla L_1(\bar{\theta}) \rangle < 0.$$

This strict inequality persists in a neighbourhood U of $\bar{\theta}$. Indeed, otherwise there would exist $\theta_n \to \bar{\theta}$ and $v_n \in V_L(\theta_n)$ with $\langle v_n, \nabla L_1(\theta_n) \rangle \geq 0$. Local boundedness gives a convergent subsequence of (v_n), and the closed-graph property places its limit in $V_L(\bar{\theta})$, contradicting the displayed inequality by continuity of ∇L_1.

Consequently, for every $\theta \in U$ and $v \in V_L(\theta)$,

$$dL_1(\theta)[-v] = -\langle v, \nabla L_1(\theta) \rangle > 0.$$

The descent conditions give $dL(\theta)[-v] \leq 0$. Since $L_0 = L - L_1$, it follows that

$$dL_0(\theta)[-v] = dL(\theta)[-v] - dL_1(\theta)[-v] < 0.$$

Finally, continuity of the trajectory gives a time interval I around t_0, within its domain, such that $\theta(t) \in U$ for all $t \in I$. The chain rule yields

$$\frac{d}{dt} L_0(\theta(t)) < 0 \quad \text{and} \quad \frac{d}{dt} L_1(\theta(t)) > 0 \qquad \text{for almost every } t \in I.$$

Absolute continuity and integration establish the claimed strict monotonicity. □

[3] In particular, if an absolutely continuous solution of $\dot{\theta}(t) \in -V_L(\theta(t))$ satisfies $\theta(t_0) \in Z_{\min}$, then there is a time neighborhood of t_0, relative to the solution's domain, on which one fixed admissible minority loss strictly decreases and its corresponding majority loss strictly increases.

B PROOFS OF SECTIONS 3 AND 4 AND LOWER BOUNDS

B.1 PROOFS OF SECTION 3

Proof of Lemma A.4. Let $L_1 = L - L_0$. Since θ is absolutely continuous and L, L_0, L_1 are C^1, the chain rule holds for a.e. $t \geq 0$. Fix such a time and write

$$\dot{\theta}(t) = -v(t), \qquad v(t) \in V_L(\theta(t)).$$

Assume moreover that $\theta(t) \in Z_{\mathrm{maj-adv}}(L_0)$. By definition of $Z_{\mathrm{maj-adv}}(L_0)$,

$$\langle v(t), \nabla L_1(\theta(t)) \rangle < 0.$$

Using $L_0 = L - L_1$, $\dot{\theta}(t) = -v(t)$, and the descent property, Equations 5 and 6, we obtain

$$\begin{aligned} \frac{d}{dt} L_0(\theta(t)) &= \left\langle \nabla L_0(\theta(t)), \dot{\theta}(t) \right\rangle \\ &= -\langle v(t), \nabla L(\theta(t)) \rangle + \langle v(t), \nabla L_1(\theta(t)) \rangle . \end{aligned}$$

The first term is nonpositive by Equations 5 and 6, and the second term is strictly negative. Hence $\frac{d}{dt} L_0(\theta(t)) < 0$. □

Proof of Lemma 3.3. Fix $\theta \in \mathbb{R}^d$ and set $r := \|\theta\|_2$. The inclusion "$\subseteq$" is exactly the gradient bound in Equation 3. Conversely, fix $\xi \in \mathbb{R}^d$ with $\|\xi\|_2 \leq \psi(r + D)$. If $r + D = 0$, then $\xi = 0$ and the zero function realizes it. Assume now that $r + D > 0$ and split

$$\xi_r := \frac{r}{r + D} \xi, \qquad \xi_D := \frac{D}{r + D} \xi.$$

Then $\|\xi_r\|_2 \leq \psi r$ and $\|\xi_D\|_2 \leq \psi D$. Let $A : \mathbb{R}^d \to \mathbb{R}^d$ be linear. If $r = 0$ or $\xi_r = 0$, take $A = 0$. Assume now that $r > 0$ and $\xi_r \neq 0$, and set

$$u := \frac{\theta}{r}, \qquad s := \frac{\xi_r}{\|\xi_r\|_2}.$$

Then $\|u\|_2 = \|s\|_2 = 1$. If $s = u$, set $B := I$. Otherwise define

$$B := I - 2 \frac{(u - s)(u - s)^\top}{\|u - s\|_2^2}.$$

The matrix

$$P := \frac{(u - s)(u - s)^\top}{\|u - s\|_2^2}$$

is the orthogonal projector onto $\mathrm{span}\{u - s\}$, so $B = I - 2P$ is symmetric and satisfies $B^2 = I$. Hence $\|B\|_{\mathrm{sp}} = 1$. Moreover, using $\|u\|_2 = \|s\|_2 = 1$,

$$\|u - s\|_2^2 = 2(1 - \langle u, s \rangle),$$

therefore

$$Bu = u - 2(u - s) \frac{\langle u - s, u \rangle}{\|u - s\|_2^2} = u - (u - s) = s.$$

Set

$$A := \frac{\|\xi_r\|_2}{r} B.$$

Then A is symmetric, $\|A\|_{\mathrm{sp}} = \|\xi_r\|_2 / r \leq \psi$, and

$$A\theta = \frac{\|\xi_r\|_2}{r} Bru = \|\xi_r\|_2 s = \xi_r.$$

Now define

$$L_{0,\xi}(z) := \frac{1}{2} \langle Az, z \rangle + \langle \xi_D, z \rangle.$$

Then $L_{0,\xi}$ is C^2, $\nabla L_{0,\xi}(\theta) = \xi_r + \xi_D = \xi$, and, for every $z \in \mathbb{R}^d$,

$$\|\nabla L_{0,\xi}(z)\|_2 \leq \|A\|_{\mathrm{sp}} \|z\|_2 + \|\xi_D\|_2 \leq \psi(\|z\|_2 + D).$$

Thus $L_{0,\xi} \in \mathcal{K}_{\min}$, proving the reverse inclusion in Equation 14. □

Proof of Theorem 3.4. Fix $\theta \in \mathbb{R}^d$ and set $\rho(\theta) := \psi(\|\theta\|_2+D)$. By the definitions of $\mathcal{Q}$ and Q_{L_0},

$$\mathcal{Q}(\theta) = \inf_{L_0 \in \mathcal{K}_{\min}} \sup_{v \in V_L(\theta)} \langle v, \nabla(L - L_0)(\theta) \rangle .$$

At this fixed θ, the expression depends on L_0 only through the vector $\xi = \nabla L_0(\theta)$. The pointwise envelope identity Equation 14, given by Lemma 3.3, therefore gives $\mathcal{Q}(\theta) = \inf_{\|\xi\|_2 \le \rho(\theta)} \sup_{v \in V_L(\theta)} \langle v, \nabla L(\theta) - \xi \rangle$, which proves the following equation

$$\mathcal{Q}(\theta) = \inf_{\|\xi\|_2 \le \psi(\|\theta\|_2+D)} \sup_{v \in V_L(\theta)} \langle v, \nabla L(\theta) - \xi \rangle .$$

Let us observe that $C(\theta) := \{\xi \in \mathbb{R}^d, \quad \|\xi\|_2 \le \psi(\|\theta\|_2+D)\}$ is nonempty convex and compact. Consider the bilinear form

$$\phi_\theta : \begin{cases} C(\theta) \times V_L(\theta) & \to \mathbb{R} \\ (\xi, v) & \to \langle v, \nabla L(\theta) - \xi \rangle. \end{cases}$$

Using von Neumann's minimax Theorem (see also Sion (1958)), we deduce

$$\inf_{\xi \in C(\theta)} \sup_{v \in V_L(\theta)} \phi_\theta(\xi, v) = \sup_{v \in V_L(\theta)} \inf_{\xi \in C(\theta)} \phi_\theta(\xi, v).$$

Therefore, for $\theta \in \mathbb{R}^d$,

$$\begin{aligned} \mathcal{Q}(\theta) &= \sup_{v \in V_L(\theta)} \inf_{\|\xi\|_2 \le \psi(\|\theta\|_2+D)} \langle v, \nabla L(\theta) - \xi \rangle \\ &= \sup_{v \in V_L(\theta)} \Big(\langle v, \nabla L(\theta) \rangle + \inf_{\xi \in C(\theta)} (-\langle v, \xi \rangle) \Big) \\ &= \sup_{v \in V_L(\theta)} \Big(\langle v, \nabla L(\theta) \rangle - \psi(\|\theta\|_2+D)\|v\|_2 \Big). \end{aligned}$$

Hence Equation 15 holds. Lemma 3.2 yields the final equivalence in Theorem 3.4. □

Corollary B.1 (Inner minority aware Zones)**.** *Under Assumptions 2.1 and 3.1 and Equation 5, we know*

$$\{\theta \in \mathbb{R}^d, 0 < \|\nabla L(\theta)\|_2 < \psi(\|\theta\|_2+D)\} \subseteq Z_{\min}.$$

Proof. Let $\theta \in \mathbb{R}^d$ such that

$$0 < \|\nabla L(\theta)\|_2 < \psi(\|\theta\|_2+D).$$

Since $\nabla L(\theta) \neq 0$, we know $0 \notin V_L(\theta)$. Therefore, since $V_L(\theta)$ is compact, we can define

$$m(\theta) := \min_{v \in V_L(\theta)} \|v\|_2 > 0.$$

We know for all $v \in V_L(\theta)$,

$$\begin{aligned} \langle v, \nabla L(\theta) \rangle - \psi(\|\theta\|_2+D)\|v\|_2 &\le \|v\|_2 \|\nabla L(\theta)\|_2 - \psi(\|\theta\|_2+D)\|v\|_2 \\ &= \|v\|_2 \Big(\|\nabla L(\theta)\|_2 - \psi(\|\theta\|_2+D) \Big) \\ &\le m(\theta) \Big(\|\nabla L(\theta)\|_2 - \psi(\|\theta\|_2+D) \Big). \end{aligned}$$

Therefore, from Theorem 3.4, we know :

$$\mathcal{Q}(\theta) = \sup_{v \in V_L(\theta)} \langle v, \nabla L(\theta) \rangle - \psi(\|\theta\|_2+D)\|v\|_2 \le m(\theta) \Big(\|\nabla L(\theta)\|_2 - \psi(\|\theta\|_2+D) \Big) < 0,$$

which implies $\theta \in Z_{\min}$. □

Corollary B.2 (Robust zone for norm-gradient fields)**.** *Under the assumptions of Theorem 3.4, one has:*

- *If* $V_L(\theta) = \rho(\mathrm{grad}_{\|\cdot\|} L(\theta))$*, then*

$$\mathcal{Q}(\theta) = \|\mathrm{d}L(\theta)\|_* - \psi(\|\theta\|_2 + D) m_{\|\cdot\|}(\theta),$$

and therefore

$$\theta \in Z_{\min} \iff \|\mathrm{d}L(\theta)\|_* < \psi m_{\|\cdot\|}(\theta)(\|\theta\|_2 + D). \tag{25}$$

- *If* $V_L(\theta) = \mathrm{grad}_{\|\cdot\|} L(\theta)$*, then*

$$\mathcal{Q}(\theta) = \|\mathrm{d}L(\theta)\|_*^2 - \psi(\|\theta\|_2 + D) m_{\|\cdot\|}(\theta) \|\mathrm{d}L(\theta)\|_*,$$

and therefore

$$\theta \in Z_{\min} \iff \|\mathrm{d}L(\theta)\|_*^2 < \psi m_{\|\cdot\|}(\theta) \|\mathrm{d}L(\theta)\|_* (\|\theta\|_2 + D). \tag{26}$$

Note that in Equations 25 and 26, the zones do not change. Indeed the strict negativity in Equation 12 implies that the zones do not contain critical points, so that we may divide Equation 26 by $\|dL(\theta)\|_*$.

Proof. Fix $\theta \in \mathbb{R}^d$ and set $\rho := \psi(\|\theta\|_2 + D)$ and $N := \rho(\mathrm{grad}_{\|\cdot\|} L(\theta))$. For the normalized field in Equation 25, by Theorem 3.4, and from the fact that $\mathrm{d}L(\theta)[N] = \{\|\mathrm{d}L(\theta)\|_*\}$ (see Equation 8), one has

$$\begin{aligned}
\mathcal{Q}(\theta) &= \sup_{v \in N} \Big(\langle v, \nabla L(\theta) \rangle - \psi(\|\theta\|_2 + D) \|v\|_2 \Big) \\
&= \|\mathrm{d}L(\theta)\|_* - \psi(\|\theta\|_2 + D) \inf_{v \in N} \|v\|_2 \\
&= \|\mathrm{d}L(\theta)\|_* - \psi(\|\theta\|_2 + D) m_{\|\cdot\|}(\theta).
\end{aligned}$$

For the norm-gradient field in Equation 26, using Equation 8 and a change of variable argument,

$$\begin{aligned}
\mathcal{Q}(\theta) &= \sup_{v \in \|\mathrm{d}L(\theta)\|_* N} \Big(\langle v, \nabla L(\theta) \rangle - \psi(\|\theta\|_2 + D) \|v\|_2 \Big) \\
&= \|\mathrm{d}L(\theta)\|_*^2 - \psi(\|\theta\|_2 + D) \inf_{v \in \|\mathrm{d}L(\theta)\|_* N} \|v\|_2 \\
&= \|\mathrm{d}L(\theta)\|_*^2 - \psi(\|\theta\|_2 + D) m_{\|\cdot\|}(\theta) \|\mathrm{d}L(\theta)\|_*.
\end{aligned}$$

The two characterizations of $Z_{\min}$, Equations 25 and 26, follow from Lemma 3.2. □

Now we present different minority aware zones.

Gradient descent For gradient descent $V_L = \nabla L$, Theorem 3.4 gives

$$\mathcal{Q}_{\mathrm{GD}}(\theta) = \|\nabla L(\theta)\|_2^2 - \psi(\|\theta\|_2 + D) \|\nabla L(\theta)\|_2. \tag{27}$$

Therefore

$$Z_{\min}^{\mathrm{GD}} = \left\{ \theta \in \mathbb{R}^d : \|\nabla L(\theta)\|_2^2 < \psi \|\nabla L(\theta)\|_2 (\|\theta\|_2 + D) \right\}. \tag{28}$$

Gradient clipping Using the notation of Section 3.1, Theorem 3.4 gives

$$\mathcal{Q}_{\mathrm{clip}}(\theta) = \|\mathrm{Clip}_{\tau_{\mathrm{clip}}}(\nabla L(\theta))\|_2 (\|\nabla L(\theta)\|_2 - \psi(\|\theta\|_2 + D)). \tag{29}$$

Consequently,

$$Z_{\min}^{\mathrm{clip}} = Z_{\min}^{\mathrm{GD}}. \tag{30}$$

Thus clipping merely changes velocities but not the minority aware zone.

Sign descent For the supremum norm ($\|\cdot\| = \|\cdot\|_\infty$), letting $|\mathrm{supp}(v)|$ be the number of non-zero components of a vector v, we get

$$\mathcal{Q}_{\mathrm{SD}}(\theta) = \|\nabla L(\theta)\|_1 - \psi \sqrt{|\mathrm{supp}(\nabla L(\theta))|} (\|\theta\|_2 + D). \tag{31}$$

Thus

$$Z_{\min}^{\mathrm{SD}} = \left\{ \theta \in \mathbb{R}^d : \|\nabla L(\theta)\|_1 < \psi \sqrt{|\mathrm{supp}(\nabla L(\theta))|} (\|\theta\|_2 + D) \right\}. \tag{32}$$

Spectral descent Let $\nabla L(\Theta) = U\Sigma V^\top$ be a compact singular value decomposition, as in Section 3.1. Combining Equation 24 with Equation 15 gives

$$\mathcal{Q}_{\mathrm{Spec}}(\Theta) = \|\nabla L(\Theta)\|_{\mathrm{nuc}} - \psi(\|\Theta\|_F + D) \inf_{Z \in \rho\left(\mathrm{grad}_{\|\cdot\|_{\mathrm{sp}}} L(\Theta)\right)} \|Z\|_F \tag{33}$$

$$= \|\nabla L(\Theta)\|_{\mathrm{nuc}} - \psi(\|\Theta\|_F + D)\sqrt{\mathrm{rank}\, \nabla L(\Theta)}. \tag{34}$$

The full proof of Equation 34 is below. Thus

$$Z^{\mathrm{Spec}}_{\min} = \left\{\Theta \in \mathbb{R}^{a\times b} : \|\nabla L(\Theta)\|_{\mathrm{nuc}} < \psi\sqrt{\mathrm{rank}(\nabla L(\Theta))}(\|\Theta\|_F + D)\right\}. \tag{35}$$

Newton descent For Newton descent, assuming in addition Equation (10), we have, by Theorem 3.4,

$$\mathcal{Q}_{\mathrm{N}}(\theta) = \langle v_N(\theta), \nabla L(\theta)\rangle - \psi\|v_N(\theta)\|_2(\|\theta\|_2 + D), \tag{36}$$

with $v_N(\theta) := \nabla^2 L(\theta)^{-1}\nabla L(\theta)$ for all θ.

Proof of the expression of $\mathcal{Q}$ for Spectral Descent (Equation 34). To establish the formula it suffices to apply Corollary B.2, i.e. it suffices to compute

$$m_{\|\cdot\|}(\Theta) := \mathrm{dist}(0, \rho(\mathrm{grad}_{\|\cdot\|} L(\Theta))) = \mathrm{dist}(0, \partial\|\nabla L(\Theta)\|_{\mathrm{nuc}}).$$

Set $G = \nabla L(\Theta)$ and $r = \mathrm{rank}\, G$, then in view of the previous developments

$$\partial\|G\|_{\mathrm{nuc}} = \{UV^\top + W : U^\top W = 0,\ WV = 0,\ \|W\|_{\mathrm{sp}} \le 1\}.$$

Hence

$$\|UV^\top + W\|_F^2 = \|UV^\top\|_F^2 + \|W\|_F^2 = r + \|W\|_F^2,$$

where the cross term vanishes by the orthogonality constraints. Therefore

$$\mathrm{dist}(0, \partial\|G\|_{\mathrm{nuc}}) = \min_{Z \in \partial\|G\|_{\mathrm{nuc}}} \|Z\|_F = \|UV^\top\|_F = \sqrt{r}.$$

Consequently,

$$\Theta \in Z^{\mathrm{Spec}}_{\min} \iff \|\nabla L(\Theta)\|_{\mathrm{nuc}} < \psi\sqrt{\mathrm{rank}(\nabla L(\Theta))}(\|\Theta\|_F + D). \tag{37}$$

□

B.2 Proofs of Section 4

Table 4 indicates the references for the upper bounds and lower bounds for each mechanism studied in Section 4.

Dynamics	Upper bounds	Lower bounds
Gradient descent	Section B.2.5	Theorem B.8
Gradient clipping	Theorem B.5	×
Sign descent	Section B.2.5	Theorem B.10
Spectral descent	Section B.2.5	Theorem B.11
Newton descent	Theorem B.6	×

Table 4: References for the entry times in Section 4.

B.2.1 Preliminaries

Lemma B.3 (Entry-time for multivalued dynamics)**.** *Under Assumptions 2.1 and 3.1, let θ be an absolutely continuous solution of Equation 4 with V_L satisfying the descent conditions, Equations 5 and 6. For all t such that $\theta(t)$ is not in the minority aware zone, we have the following property, for all $L_0 \in \mathcal{K}_{\min}$,*

$$\sup_{v \in V_L(\theta(t))} \langle v, \nabla L_1(\theta(t))\rangle \ge 0.$$

For almost all t such that $\theta(t)$ is in the minority aware zone, there exists a decomposition $L = L_1 + L_0$, with $L_0 \in \mathcal{K}_{\min}$, $L_1 \in \mathcal{K}_{\mathrm{maj}}$, such that

$$\frac{d}{dt}L_1(\theta(t)) \ge -\sup_{v \in V_L(\theta(t))} \langle v, \nabla L_1(\theta(t))\rangle > 0, \quad \frac{d}{dt}L_0(\theta(t)) < 0.$$

Proof. For the first inequality, by definition,

$$\theta(t) \notin Z_{\min} \Leftrightarrow \inf_{L_0 \in \mathcal{K}_{\min}} Q_{L_0}(\theta(t)) = \inf_{L_0 \in \mathcal{K}_{\min}} \sup_{v \in V_L(\theta(t))} \langle v, \nabla L_1(\theta(t)) \rangle \geq 0.$$

For the second inequality, denote by $v(t) \in V_L(\theta(t))$ the selected velocity, i.e., $\dot{\theta}(t) = -v(t)$. Fix a time t like in the assumptions at which θ is differentiable with derivative coinciding with $-v(t)$. Since $\theta(t) \in Z_{\mathrm{maj-adv}}(L_0)$ for some L_0, then $\langle v(t), \nabla L_1(\theta(t)) \rangle < 0$. Thus

$$\frac{d}{dt} L_1(\theta(t)) = \left\langle \nabla L_1(\theta(t)), \dot{\theta}(t) \right\rangle = - \langle v(t), \nabla L_1(\theta(t)) \rangle \geq - \sup_{v \in V_L(\theta(t))} \langle v, \nabla L_1(\theta(t)) \rangle > 0.$$

Since V_L satisfies Equation 6, one has $\frac{d}{dt} L_0(\theta(t)) = - \langle v(t), \nabla L(\theta(t)) \rangle + \langle v(t), \nabla L_1(\theta(t)) \rangle < 0$. □

Lemma B.4 (Lower bound of $m_{\|\cdot\|}$)**.** *Consider an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$. We recall from Equation* (16) *that*

$$c_{\|\cdot\|} := \inf_{v \in \mathbb{R}^d / \{0\}} \frac{\|v\|_2}{\|v\|}.$$

Then we know that for all θ such that $\nabla L(\theta) \neq 0$,

$$m_{\|\cdot\|}(\theta) \geq c_{\|\cdot\|} > 0.$$

Proof. Let $\theta \in \mathbb{R}^d$ such that $\nabla L(\theta) \neq 0$. Recall that

$$m_{\|\cdot\|}(\theta) := \inf_{v \in \rho\left(\mathrm{grad}_{\|\cdot\|} L(\theta)\right)} \|v\|_2.$$

Since $\nabla L(\theta) \neq 0$, the linear function

$$\phi : \begin{cases} \{v, \|v\| \leq 1\} & \to \mathbb{R} \\ z & \mapsto \langle \nabla L(\theta), z \rangle \end{cases}$$

reaches its maximum over the boundary $\|v\| = 1$. For all $v \in \rho(\mathrm{grad}_{\|\cdot\|} L(\theta))$ such that $\|v\| = 1$, we know

$$\|v\|_2 \geq c_{\|\cdot\|} \|v\| = c_{\|\cdot\|}.$$

We conclude the proof by taking the $\inf$ over v. □

B.2.2 Proofs of Theorems 4.1 and 4.2

Common pre-entry estimates We use the following elementary estimates in the entry-time proofs.

- The constant $\lambda_{\max}$ is any constant satisfying Assumption 4.1; for a C^2 loss, the Lipschitz-gradient part is ensured by

  $$\lambda_{\max} \geq \sup_{\theta \in \mathbb{R}^d} \sigma_{\max}(\nabla^2 L(\theta)).$$

  The parameter ψ is the granularity of the admissible minority class $\mathcal{K}_{\min}$, and D is the offset scale in the bound $\|\nabla L_0(\theta)\|_2 \leq \psi(\|\theta\|_2 + D)$.
- We know from Corollary B.1 that $\|\nabla L(\theta)\| < \psi D \implies \theta \in Z_{\min}$. By the continuity of $\theta \to \|\nabla L(\theta)\|$, we deduce that if $\nabla L(\theta_{\mathrm{init}}) \neq 0$, there is no critical point before the entry time. As consequence, from Lemma B.4, $m_{\|\cdot\|}$ is greater than $c_{\|\cdot\|}$ before the entry time as well.
- Recall that

  $$m_{\|\cdot\|}(\theta) := \mathrm{dist}(0, \rho(\mathrm{grad}_{\|\cdot\|} L(\theta))) = \inf_{v \in \rho\left(\mathrm{grad}_{\|\cdot\|} L(\theta)\right)} \|v\|_2.$$

  For the normalized gradient $\rho(\mathrm{grad}_{\|\cdot\|} L(\theta))$ dynamics, the cost encoding the minority aware zone is

  $$\mathcal{Q}_{\|\cdot\|}(\theta) = \|\mathrm{d}L(\theta)\|_* - \psi m_{\|\cdot\|}(\theta)(\|\theta\|_2 + D).$$

For the normal gradient $\mathrm{grad}_{\|\cdot\|}\, L(\theta)$, the cost is

$$\mathcal{Q}(\theta) = \|\mathrm{d}L(\theta)\|_*^2 - \psi(\|\theta\|_2 + D) m_{\|\cdot\|}(\theta) \|\mathrm{d}L(\theta)\|_*.$$

Consequently, before entry,

$$\|\mathrm{d}L(\theta(t))\|_* \geq \psi\, m_{\|\cdot\|}(\theta(t))(\|\theta(t)\|_2 + D) \geq \psi D m_{\|\cdot\|}(\theta(t)). \tag{38}$$

We note that from Assumption 4.1, one has the following lower bound for all θ

$$\|\theta\|_2 + D \geq \sqrt{\frac{L(\theta) - L_\star}{\lambda_{\max}}}. \tag{39}$$

Consequently, before entry,

$$\|\mathrm{d}L(\theta(t))\|_* \geq \psi\, m_{\|\cdot\|}(\theta(t))(\|\theta(t)\|_2 + D) \geq \frac{\psi m_{\|\cdot\|}(\theta(t))}{\sqrt{\lambda_{\max}}} \sqrt{L(\theta(t)) - L_\star}. \tag{40}$$

Two differential inequalities

Let

$$G(t) := L(\theta(t)) - L_\star \quad (t \geq 0), \qquad C := \frac{L(\theta_{\mathrm{init}}) - L_\star}{\lambda_{\max} D^2}.$$

Recall from Equation 17 that

$$\bar{m}(t) := \frac{1}{t} \int_0^t m_{\|\cdot\|}(\theta(s))\, ds \quad (t > 0).$$

Since there are no critical points before the entry time, one has

$$G(t) > 0 \quad \text{for all } t < t_{\,\mathrm{entry}}(\theta).$$

Indeed, $G(t) = 0$ would imply that $\theta(t)$ is a global minimizer of L, and hence that $\nabla L(\theta(t)) = 0$.

a) Normalized norm-gradient.

Assume

$$\dot{\theta}(t) \in -\rho(\mathrm{grad}_{\|\cdot\|}\, L(\theta(t)))$$

and that

$$\nabla L(\theta_{\mathrm{init}}) \neq 0.$$

Set

$$T := t_{\,\mathrm{entry}}(\theta).$$

By the common pre-entry estimates above, there is no critical point before the entry time. Consequently, by Lemma B.4,

$$m_{\|\cdot\|}(\theta(t)) \geq c_{\|\cdot\|} > 0 \quad \text{for all } t < T.$$

For almost every $t < T$, the normalized norm-gradient dynamics gives

$$-G'(t) = \|\mathrm{d}L(\theta(t))\|_*.$$

Hence, by Equation 38,

$$-G'(t) \geq \psi D m_{\|\cdot\|}(\theta(t)).$$

Therefore, for every t such that $0 < t < T$,

$$\begin{aligned} G(0) - G(t) &\geq \psi D \int_0^t m_{\|\cdot\|}(\theta(s))\, ds \\ &= \psi D t \bar{m}(t). \end{aligned}$$

Since $G(t) \geq 0$, it follows that

$$t\bar{m}(t) \leq \frac{G(0)}{\psi D}. \tag{41}$$

Moreover, from Lemma B.4,

$$\bar{m}(t) = \frac{1}{t}\int_0^t m_{\|\cdot\|}(\theta(s))\,ds \geq c_{\|\cdot\|}.$$

The entry time T must therefore be finite. Indeed, if $T = +\infty$, then

$$tc_{\|\cdot\|} \leq \frac{G(0)}{\psi D}$$

would hold for every $t > 0$, which is impossible.

If $T = 0$, the theorem holds. Suppose now that $T > 0$. Letting t increase to T in the preceding inequality, Equation 41, gives

$$\int_0^T m_{\|\cdot\|}(\theta(s))\,ds \leq \frac{G(0)}{\psi D}.$$

By the definition of $\bar{M}_\theta$ in Equation 18,

$$T\bar{M}_\theta \leq \frac{G(0)}{\psi D}.$$

Since

$$G(0) = L(\theta_{\text{init}}) - L_\star,$$

we obtain

$$t_{\,\text{entry}}(\theta) \leq \frac{L(\theta_{\text{init}}) - L_\star}{\psi D\bar{M}_\theta}.$$

Now suppose that Assumption 4.1 also holds. By Equation 40, for almost every $t < T$,

$$-G'(t) = \|\mathrm{d}L(\theta(t))\|_* \geq \frac{\psi m_{\|\cdot\|}(\theta(t))}{\sqrt{\lambda_{\max}}}\sqrt{G(t)}.$$

Recall that

$$G(t) > 0 \quad \text{for all } t < T.$$

Thus, for almost every $t < T$,

$$\begin{aligned}\frac{d}{dt}\sqrt{G(t)} &= \frac{G'(t)}{2\sqrt{G(t)}} \\ &\leq -\frac{\psi}{2\sqrt{\lambda_{\max}}} m_{\|\cdot\|}(\theta(t)).\end{aligned}$$

Integrating over $[0, t]$, for every t such that $0 < t < T$, gives

$$\sqrt{G(t)} \leq \sqrt{G(0)} - \frac{\psi}{2\sqrt{\lambda_{\max}}}\int_0^t m_{\|\cdot\|}(\theta(s))\,ds.$$

Since $G(t) \geq 0$, we obtain

$$\int_0^t m_{\|\cdot\|}(\theta(s))\,ds \leq \frac{2\sqrt{\lambda_{\max}}}{\psi}\sqrt{G(0)}.$$

Letting t increase to T gives

$$T\bar{M}_\theta \leq \frac{2\sqrt{\lambda_{\max}}}{\psi}\sqrt{G(0)}.$$

Consequently,

$$t_{\,\text{entry}}(\theta) \leq \frac{2\sqrt{\lambda_{\max}}}{\psi\bar{M}_\theta}\sqrt{L(\theta_{\text{init}}) - L_\star}.$$

b) Norm-gradient without normalization. Assume

$$\dot{\theta}(t) \in -\operatorname{grad}_{\|\cdot\|} L(\theta(t))$$

and that $\nabla L(\theta_{\text{init}}) \neq 0$. We get for $t < t_{\text{entry}}(\theta)$

$$\|\mathrm{d}L(\theta(t))\|_*^2 \geq \psi^2 D^2 m_{\|\cdot\|}(\theta(t))^2.$$

For a.e. $t < t_{\text{entry}}(\theta)$, Equation (19) gives

$$-G'(t) = \|\mathrm{d}L(\theta(t))\|_*^2 \geq \psi^2 D^2 m_{\|\cdot\|}(\theta(t))^2.$$

By Jensen's inequality,

$$\int_0^t m_{\|\cdot\|}(\theta(s))^2 \, ds \geq t\bar{m}(t)^2 \geq t c_{\|\cdot\|}^2.$$

Since $G(t) \geq 0$, we obtain that t_{entry} is finite and

$$\bar{M}_\theta^2 t_{\text{entry}} \leq \frac{G(0)}{\psi^2 D^2}. \tag{42}$$

Now suppose that Assumption 4.1 holds. Set

$$T := t_{\text{entry}}(\theta) \qquad \text{and} \qquad C := \frac{G(0)}{\lambda_{\max} D^2} = \frac{L(\theta_{\text{init}}) - L_\star}{\lambda_{\max} D^2}.$$

Recall that $C > 0$, since $\nabla L(\theta_{\text{init}}) \neq 0$.

Let $\beta \in (0, 2]$. For almost every $t < T$, using Equations 38 and 40, we obtain

$$\begin{aligned}
-G'(t) &= \|\mathrm{d}L(\theta(t))\|_*^2 \\
&= \|\mathrm{d}L(\theta(t))\|_*^{2-\beta} \|\mathrm{d}L(\theta(t))\|_*^\beta \\
&\geq \left(\frac{\psi m_{\|\cdot\|}(\theta(t))}{\sqrt{\lambda_{\max}}} \sqrt{G(t)} \right)^{2-\beta} \left(\psi D m_{\|\cdot\|}(\theta(t)) \right)^\beta \\
&= \frac{\psi^2}{\lambda_{\max}} \left(\lambda_{\max} D^2 \right)^{\beta/2} m_{\|\cdot\|}(\theta(t))^2 G(t)^{1-\beta/2}.
\end{aligned}$$

Since $G(t) > 0$ for every $t < T$, it follows that, for almost every $t < T$,

$$\frac{d}{dt} G(t)^{\beta/2} = \frac{\beta}{2} G(t)^{\beta/2-1} G'(t) \leq -\frac{\beta \psi^2}{2\lambda_{\max}} \left(\lambda_{\max} D^2 \right)^{\beta/2} m_{\|\cdot\|}(\theta(t))^2.$$

Therefore, for every $t < T$,

$$\begin{aligned}
G(t)^{\beta/2} \leq\ & G(0)^{\beta/2} \\
& - \frac{\beta \psi^2}{2\lambda_{\max}} \left(\lambda_{\max} D^2 \right)^{\beta/2} \int_0^t m_{\|\cdot\|}(\theta(s))^2 \, ds.
\end{aligned}$$

Since $G(t) \geq 0$, this gives

$$\int_0^t m_{\|\cdot\|}(\theta(s))^2 \, ds \leq \frac{2\lambda_{\max}}{\beta \psi^2} C^{\beta/2}.$$

Letting t increase to T, and using Jensen's inequality, we obtain

$$\begin{aligned}
T \bar{M}_\theta^2 &= \frac{1}{T} \left(\int_0^T m_{\|\cdot\|}(\theta(s)) \, ds \right)^2 \\
&\leq \int_0^T m_{\|\cdot\|}(\theta(s))^2 \, ds \\
&\leq \frac{2\lambda_{\max}}{\beta \psi^2} C^{\beta/2}.
\end{aligned}$$

Consequently, for every $\beta \in (0, 2]$,

$$T \leq \frac{\lambda_{\max}}{\psi^2 \bar{M}_\theta^2} \frac{2C^{\beta/2}}{\beta}.$$

It remains to minimize the function

$$f_C(\beta) := \frac{2C^{\beta/2}}{\beta}, \qquad \beta \in (0, 2].$$

Its derivative is

$$f'_C(\beta) = \frac{2C^{\beta/2}}{\beta^2}\left(\frac{\beta}{2}\log C - 1\right).$$

If $0 < C \leq e$, the minimum of f_C over $(0, 2]$ is attained at $\beta = 2$, and

$$\inf_{\beta\in(0,2]} f_C(\beta) = f_C(2) = C.$$

If $C > e$, the minimum is attained at

$$\beta_\star = \frac{2}{\log C} \in (0, 2),$$

and

$$\inf_{\beta\in(0,2]} f_C(\beta) = f_C(\beta_\star) = e \log C.$$

We have thus proved the piecewise bound

$$t_{\,\mathrm{entry}}(\theta) \leq \frac{\lambda_{\max}}{\psi^2 \bar{M}_\theta^2} \begin{cases} C, & 0 < C \leq e, \\ e \log C, & C > e. \end{cases}$$

B.2.3 Clipped gradient descent

Theorem B.5 (Upper bound of entry time for clipped gradient descent)**.** *Suppose that Assumptions 2.1 and 3.1 hold and consider the clipped gradient flow*

$$\theta(0) = \theta_{\mathrm{init}} \in \mathbb{R}^d,$$
$$\forall t \geq 0, \quad \dot\theta(t) = -\mathrm{Clip}_{\tau_{\mathrm{clip}}}(\nabla L(\theta(t))),$$

with threshold $\tau_{\mathrm{clip}} > 0$. Assume that $\nabla L(\theta_{\mathrm{init}}) \neq 0$. Then,

$$t_{\,\mathrm{entry}}(\theta) \leq \frac{L(\theta_{\mathrm{init}}) - L_\star}{\min\{\psi^2 D^2, \tau_{\mathrm{clip}}\psi D\}}. \tag{43}$$

If Assumption 4.1 holds, we obtain that

$$t_{\,\mathrm{entry}}(\theta) \leq \frac{2\sqrt{\lambda_{\max}}}{\psi \min\{\tau_{\mathrm{clip}}, \psi D\}} \sqrt{L(\theta_{\mathrm{init}}) - L_\star}. \tag{44}$$

Proof. Set $E(t) := L(\theta(t)) - L_\star$. Before entry, Equation 30 gives

$$\|\nabla L(\theta(t))\|_2 \geq \psi(\|\theta(t)\|_2 + D) \geq \psi D.$$

As in the proof of Theorems 4.1 and 4.2, the chain rule gives, for a.e. pre-entry time,

$$E'(t) = -\min\{\|\nabla L(\theta(t))\|_2^2, \tau_{\mathrm{clip}}\|\nabla L(\theta(t))\|_2\} \leq -\min\{\psi^2 D^2, \tau_{\mathrm{clip}}\psi D\}.$$

Integrating up to any $s < t_{\,\mathrm{entry}}(\theta)$ and using $E(s) \geq 0$ gives the claim in Equation 43.

Now suppose that Assumption 4.1 holds. We have for $t < t_{\,\mathrm{entry}}$, from Equation 40 and with $m_{\|\cdot\|} \equiv 1$,

$$E'(t) \leq -\min\{\tau_{\mathrm{clip}}, \psi D\}\|\nabla L(\theta(t))\|_2 \leq -\min\{\tau_{\mathrm{clip}}, \psi D\}\frac{\psi}{\sqrt{\lambda_{\max}}}\sqrt{E(t)}.$$

Similarly to Theorem 4.1, we deduce that

$$2\frac{d}{dt}\sqrt{L(\theta(t)) - L_\star} \leq -\frac{\psi \min\{\tau_{\mathrm{clip}}, \psi D\}}{\sqrt{\lambda_{\max}}}$$

for a.e. pre-entry time. Integration gives the stated bound in Equation 44. □

B.2.4 NEWTON'S DESCENT

Theorem B.6 (Upper bound of entry time for Newton method)**.** *Consider a solution $\theta(t)$ to the following dynamical system,*

$$\theta(0) = \theta_{\text{init}} \in \mathbb{R}^d$$
$$\forall t \geq 0, \quad \dot{\theta}(t) = -\left(\nabla^2 L(\theta(t))\right)^{-1} \nabla L(\theta(t)),$$

such that $\nabla^2 L(\theta(t))$ is positive definite along the trajectory. Suppose that we have at the initialization,

$$\|\nabla L(\theta_{\text{init}})\|_2 > \psi(\|\theta_{\text{init}}\|_2 + D).$$

Then we know that

$$t_{\,\text{entry}}(\theta) \leq \log\left(\frac{\|\nabla L(\theta_{\text{init}})\|_2}{\psi D}\right).$$

Proof. Along the Newton flow,

$$\frac{d}{dt}\nabla L(\theta(t)) = \nabla^2 L(\theta(t))\dot{\theta}(t) = -\nabla L(\theta(t)),$$

hence $\|\nabla L(\theta(t))\|_2 = \|\nabla L(\theta_{\text{init}})\|_2 e^{-t}$. By Equation 36 and Cauchy–Schwarz,

$$\|\nabla L(\theta(t))\|_2 < \psi D \quad \Longrightarrow \quad \mathcal{Q}(\theta(t)) < 0.$$

Thus the first time at which $\|\nabla L(\theta_{\text{init}})\|_2 e^{-t} < \psi D$ is an upper bound for $t_{\,\text{entry}}(\theta)$, which gives the result. □

B.2.5 APPLICATIONS TO STANDARD DESCENTS

We apply Theorems 4.1 and 4.2.

Gradient descent. For $\|\cdot\| = \|\cdot\|_2$, before entry

$$\rho(\text{grad}_{\|\cdot\|_2} L(\theta)) = \left\{\frac{\nabla L(\theta)}{\|\nabla L(\theta)\|_2}\right\}, \qquad m_{\|\cdot\|_2}(\theta) = 1.$$

Thus $\bar{m}(s) = 1$, and Theorem 4.2 gives the stated bound.

Sign descent. For $\|\cdot\| = \|\cdot\|_\infty$,

$$m_{\|\cdot\|_\infty}(\theta) = \text{dist}(0, \text{sign}(\nabla L(\theta))) = \sqrt{|\text{supp}(\nabla L(\theta))|}. \tag{45}$$

Applying Theorem 4.1 gives the stated bound.

Spectral descent. Let $G = \nabla L(\Gamma)$ have rank r and compact SVD $G = U\Sigma V^\top$. For $\|\cdot\|_{\text{sp}}$,

$$\rho(\text{grad}_{\|\cdot\|_{\text{sp}}} L(\Gamma)) = \{UV^\top + W : \, U^\top W = 0, \, WV = 0, \, \|W\|_{\text{sp}} \leq 1\}, \tag{46}$$

so

$$m_{\|\cdot\|_{\text{sp}}}(\Gamma) = \text{dist}(0, \rho(\text{grad}_{\|\cdot\|_{\text{sp}}} G)) = \sqrt{r}. \tag{47}$$

Applying Theorem 4.1 gives the stated bound.

B.3 LOWER BOUNDS

We will build a one-dimensional function that will be useful for the lower bounds in this section. Consider $K \geq \psi$ where K is a parameter of the function, that quantifies its second derivative on the positive part of its input space. For all $\delta > 0$:

$$g_\delta(-\delta) := 0$$
$$g'_\delta(\theta) := \begin{cases} K\theta + \psi D & \text{for } \theta \geq 0 \\ P(\theta) & \text{for } \theta \in [-\delta, 0) \\ 0 & \text{for } \theta \leq -\delta \end{cases} \tag{48}$$

where P is an interpolation polynomial of degree at most 3 defined in the following way:

$$P(\theta) := A(\theta+\delta)^3 + B(\theta+\delta)^2$$
$$A := \frac{K\delta - 2\psi D}{\delta^3}$$
$$B := \frac{3\psi D - K\delta}{\delta^2}.$$

Lemma B.7. *For all $\delta > 0$, $g_\delta \in \mathcal{C}^2(\mathbb{R},\mathbb{R})$, $g_\delta(0) = \frac{\delta\psi D}{2} - \frac{K\delta^2}{12}$ and*

$$\forall 0 < x < \delta, \quad |g_\delta(0) - g_\delta(-x)| \leq K\delta^2 + 2\delta\psi D.$$

By looking at the sign of g'_δ, we know

$$g_\delta(0) - \min g_\delta \leq K\delta^2 + 2\delta\psi D.$$

Proof. The function g_δ is $\mathcal{C}^2$ in each segment separately. We verify that the derivative g'_δ is continuous at the borders:

$$P(0) = A\delta^3 + B\delta^2 = \psi D$$
$$P(-\delta) = 0,$$

and also g'_δ is continuously differentiable at the borders

$$P'(0) = 3A\delta^2 + 2B\delta = \frac{1}{\delta}\left(3K\delta - 6\psi D + 6\psi D - 2K\delta\right) = K$$
$$P'(-\delta) = 0.$$

Now we can calculate the value at 0:

$$\begin{aligned} g_\delta(0) = \int_{-\delta}^{0} P(\theta)d\theta &= \Big[A\frac{(\theta+\delta)^4}{4} + B\frac{(\theta+\delta)^3}{3}\Big]_{-\delta}^{0} \\ &= \frac{A\delta^4}{4} + \frac{B\delta^3}{3} \\ &= \frac{K\delta^2 - 2\delta\psi D}{4} + \frac{3\delta\psi D - K\delta^2}{3} \\ &= \frac{\delta\psi D}{2} - \frac{K\delta^2}{12}. \end{aligned}$$

Last but not least, for $0 < x < \delta$,

$$\begin{aligned} g_\delta(0) - g_\delta(-x) = \int_{-x}^{0} P(\theta)d\theta &= \Big[A\frac{(\theta+\delta)^4}{4} + B\frac{(\theta+\delta)^3}{3}\Big]_{-x}^{0} \\ &= \frac{\delta\psi D}{2} - \frac{K\delta^2}{12} - \frac{A(\delta-x)^4}{4} - \frac{B(\delta-x)^3}{3}. \end{aligned}$$

By taking the absolute values:

$$\begin{aligned} |g_\delta(0) - g_\delta(-x)| &\leq \frac{\delta}{4}|K\delta - 2\psi D| + \frac{\delta}{3}|3\psi D - K\delta| + \frac{\delta\psi D}{2} + \frac{K\delta^2}{12} \\ &\leq (\frac{1}{4} + \frac{1}{3} + \frac{1}{12})K\delta^2 + \delta\psi D(\frac{1}{2} + 1 + \frac{1}{2}). \end{aligned}$$

□

B.3.1 Lower bound for Gradient Descent

For gradient descent, we stay in dimension one with g_δ for the lower bound. The strategy is to lower-bound the entry time by the initial suboptimality gap divided by $3\psi^2 D^2$, up to an error that vanishes as $\delta \to 0$ (Remark B.1).

Theorem B.8. *Consider $0 < \theta_{\text{init}} \leq \frac{\psi}{K} D$. For any $\delta > 0$, we follow the Gradient descent dynamics with g_δ defined in Equation* (48)*. Then the entry time is lower bounded as*

$$t_{\text{entry}} \geq \frac{g_\delta(\theta_{\text{init}}) - \min g_\delta}{3\psi^2 D^2} - \frac{K\delta^2}{3\psi^2 D^2} - \frac{2\delta}{3\psi D}. \tag{49}$$

Proof. We recall from Table 1 that the entry time condition for Gradient descent in dimension one is

$$0 < |g'(\theta)| < \psi(|\theta| + D).$$

Note that we initialize at the positive part $\theta_{init} > 0$. We also notice that for $\theta > 0$,

$$|g'_\delta(\theta)| = K\theta + \psi D \geq \psi(\theta + D).$$

Therefore, the entry time is at least the travelling time T_0 from θ_{init} to 0. Now we know for $0 < t < T_0$:

$$\theta'(t) = -g'_\delta(\theta(t)) = -K\theta(t) - \psi D$$
$$\theta(t) = \text{Const} \exp(-Kt) - \frac{\psi D}{K}$$

with Const a constant that depends on the initialization $\theta(t = 0) = \theta_{\text{init}}$, that is $\text{Const} = \theta_{\text{init}} + \frac{\psi D}{K}$. Next, we have

$$\begin{aligned}
\theta(t) &= \left(\theta_{\text{init}} + \frac{\psi D}{K}\right) \exp(-Kt) - \frac{\psi D}{K} \\
\implies \quad 0 &= \left(\theta_{\text{init}} + \frac{\psi D}{K}\right) \exp(-KT_0) - \frac{\psi D}{K} \\
\implies \quad T_0 &= \frac{1}{K} \log\left(\frac{\theta_{\text{init}} + \frac{\psi D}{K}}{\frac{\psi D}{K}}\right) = \frac{1}{K} \log\left(1 + \frac{K\theta_{\text{init}}}{\psi D}\right).
\end{aligned}$$

The last step is to introduce the suboptimality gap:

$$\begin{aligned}
g_\delta(\theta_{\text{init}}) - g_\delta(0) &= \int_0^{\theta_{\text{init}}} (K\theta + \psi D)\, d\theta = \frac{K}{2}\theta_{init}^2 + \psi D \theta_{init} \\
\frac{\psi^2 D^2}{2K} + g_\delta(\theta_{\text{init}}) - g_\delta(0) &= \frac{1}{2}(\sqrt{K}\theta_{\text{init}} + \frac{\psi D}{\sqrt{K}})^2 \\
1 + \frac{2K(g_\delta(\theta_{\text{init}}) - g_\delta(0))}{\psi^2 D^2} &= \left(\frac{K\theta_{\text{init}}}{\psi D} + 1\right)^2.
\end{aligned}$$

Now we can rearrange the terms:

$$T_0 = \frac{1}{K} \log\left(1 + \frac{K\theta_{\text{init}}}{\psi D}\right) = \frac{1}{2K} \log\left(1 + \frac{2K(g_\delta(\theta_{\text{init}}) - g_\delta(0))}{\psi^2 D^2}\right).$$

Since we supposed that $\theta_{\text{init}} \leq \frac{\psi}{K} D$, we know that

$$2K(g_\delta(\theta_{\text{init}}) - g_\delta(0)) = K^2\theta_{\text{init}}^2 + 2K\psi\theta_{\text{init}} D \leq 3\psi^2 D^2.$$

Since $\forall x \in [0, 3], \ln(1 + x) \geq x/3$, we get

$$\begin{aligned}
t_{entry} \geq T_0 &\geq \frac{1}{2K} \frac{2K(g_\delta(\theta_{\text{init}}) - g_\delta(0))}{3\psi^2 D^2} \\
&= \frac{g_\delta(\theta_{\text{init}}) - \min g_\delta}{3\psi^2 D^2} + \frac{\min g_\delta - g_\delta(0)}{3\psi^2 D^2} \\
&\geq \frac{g_\delta(\theta_{\text{init}}) - \min g_\delta}{3\psi^2 D^2} - \frac{K\delta^2 + 2\delta\psi D}{3\psi^2 D^2} \\
&= \frac{g_\delta(\theta_{\text{init}}) - \min g_\delta}{3\psi^2 D^2} - \frac{K\delta^2}{3\psi^2 D^2} - \frac{2\delta}{3\psi D},
\end{aligned}$$

where we used Lemma B.7 in the last inequality. □

Remark B.1. *In Theorem B.8, for fixed θ_{init}, K, ψ and D, we can let $\delta > 0$ go to zero. In this case, in the right-hand side of Equation 49, the quantity $-\frac{K\delta^2}{3\psi^2 D^2} - \frac{2\delta}{3\psi D}$ vanishes, while the quantity $g_\delta(\theta_{\mathrm{init}}) - \min g_\delta$ converges to $K\theta_{\mathrm{init}}^2/2 + \psi D\theta_{\mathrm{init}} > 0$. Hence, for δ small enough, the right-hand side of Equation 49 is at least $\frac{g_\delta(\theta_{\mathrm{init}}) - \min g_\delta}{6\psi^2 D^2}$. Note also that the trajectory in Theorem B.8 encounters a global loss function that has second derivative K, which is arbitrarily larger than the parameter ψ defining the minority loss class in Equation 3. A similar remark also applies to Theorems B.10 and B.11.*

Remark B.2. *Consider the lower bound $t_{\mathrm{entry,LB}}$ as the right-hand side of Equation 49 in Theorem B.8, so that $t_{\mathrm{entry}} \geq t_{\mathrm{entry,LB}}$. Recall that t_{entry} is the entry time to the minority aware zone $Z_{\min} = \bigcup_{L_0 \in \mathcal{K}_{\min}} Z_{\mathrm{maj-adv}}(L_0)$. For any decomposition $L = L_0 + L_1$, consider the entry time to the majority adverse zone $Z_{\mathrm{maj-adv}}(L_0)$, denoted $t_{\mathrm{entry,maj}}(L_0)$. Then, because $Z_{\mathrm{maj-adv}}(L_0) \subseteq Z_{\min}$, we have $t_{\mathrm{entry,maj}}(L_0) \geq t_{\mathrm{entry}} \geq t_{\mathrm{entry,LB}}$. In words, $t_{\mathrm{entry,LB}}$ is a lower bound not only on the minority aware zone, that is oblivious to the choice of the decomposition $L = L_0 + L_1$, but also on the majority adverse zone, for any choice of the decomposition $L = L_0 + L_1$. This remark also applies to Theorems B.10 and B.11.*

B.3.2 Lower bound for Sign Descent

Now we will make a small modification to g_δ by substituting $D \leftarrow D/\sqrt{k}$ with $k \geq 1$. We define for all $\delta > 0$:

$$g_{\delta,k}(-\delta) := 0$$
$$g'_{\delta,k}(\theta) := \begin{cases} K\theta + \frac{\psi D}{\sqrt{k}} & \text{for } \theta \geq 0 \\ P(\theta) & \text{for } \theta \in [-\delta, 0) \\ 0 & \text{for } \theta \leq -\delta \end{cases} \tag{50}$$

where P is an interpolation polynomial of degree at most 3 defined in the following way:

$$P(\theta) := A(\theta + \delta)^3 + B(\theta + \delta)^2$$
$$A := \frac{K\delta - 2\psi(D/\sqrt{k})}{\delta^3}$$
$$B := \frac{3\psi(D/\sqrt{k}) - K\delta}{\delta^2}.$$

We know that the following lemma holds for the same reason as before.

Lemma B.9. *For all $\delta > 0$, $g_{\delta,k} \in \mathcal{C}^2(\mathbb{R}, \mathbb{R})$, $g_{\delta,k}(0) = \frac{\delta\psi D}{2\sqrt{k}} - \frac{K\delta^2}{12}$ and*

$$\forall 0 < x < \delta, \quad |g_{\delta,k}(0) - g_{\delta,k}(-x)| \leq K\delta^2 + \frac{2\delta\psi D}{\sqrt{k}}.$$

Now we consider a class of diagonal functions. Let $1 \leq k \leq d$, and let

$$\forall \theta \in \mathbb{R}^d, \quad L_{\delta,k}(\theta) := \sum_{i=1}^{k} g_{\delta,k}(\theta_i). \tag{51}$$

In the next theorem, $\mathbf{1}_k$ is the d-dimensional vector with first k components equal to one, and the remaining ones equal to zero.

Theorem B.10. *Consider $\delta > 0$, $\theta_{\mathrm{init}} = R\mathbf{1}_k$ with $R > 0$ and assume that $\|\theta_{\mathrm{init}}\|_2 \leq \frac{\psi D}{K}$ and that we follow the Sign Descent dynamics of $L_{\delta,k}$ defined in Equation* (51)*. Then, there exists an absolutely continuous solution of the sign-descent inclusion for which the entry time is lower bounded as*

$$t_{\mathrm{entry}} \geq \frac{2}{3\bar{M}_{\mathrm{SD}}} \frac{L_{\delta,k}(\theta_{\mathrm{init}}) - \min L_{\delta,k}}{\psi D} - \frac{K\bar{M}_{\mathrm{SD}}\delta^2}{\psi D} - 2\delta.$$

Proof. First, by the symmetry of the system, there exists a solution that always stays on the diagonal:

$$\forall t \geq 0, \quad \theta(t) = r(t)\mathbf{1}_k$$

where $r : \mathbb{R}_+ \to \mathbb{R}$. We notice that for $1 \leq i \leq k$, if $\theta_i(t) = r(t) \geq 0$,

$$\nabla_i L_{\delta,k}(\theta(t)) = Kr(t) + \frac{\psi D}{\sqrt{k}}.$$

Therefore for $t \geq 0$

$$r(t) > 0 \implies r'(t) = -\operatorname{sign}(Kr(t) + \frac{\psi D}{\sqrt{k}}) = -1.$$

Therefore, we can define the travel time T_0 for $r(t)$ from R to 0 and we have,

$$\forall t \in [0, T_0), \quad r(t) = R - t, \quad |\operatorname{supp}(\nabla L_{\delta,k}(\theta(t)))| = k.$$

Now we calculate the entry time. We recall from Table 1 that the entry time condition for Sign Descent is

$$\|\nabla L(\theta)\|_1 < \psi\sqrt{|\operatorname{supp}(\nabla L(\theta))|}(\|\theta\|_2 + D).$$

In our diagonal setting, for $0 \leq t < T_0$,

$$\begin{aligned} \|\nabla L(\theta(t))\|_1 &= k|Kr(t) + \frac{\psi D}{\sqrt{k}}| \\ &= |kKr(t) + \sqrt{k}\psi D| \\ \psi\sqrt{|\operatorname{supp}(\nabla L(\theta(t)))|}(\|\theta(t)\|_2 + D) &= \psi\sqrt{k}(\sqrt{k}|r(t)| + D) \\ &= \psi k|r(t)| + \psi\sqrt{k}D. \end{aligned}$$

We notice that

$$r(t) \geq 0 \implies \|\nabla L(\theta(t))\|_1 \geq \psi\sqrt{|\operatorname{supp}(\nabla L(\theta(t)))|}(\|\theta(t)\|_2 + D).$$

Therefore, we deduce that $R = T_0 \leq t_{\text{entry}}$.

We now show that $t_{\text{entry}} = R$. Since $g'_{\delta,k}(0) = \psi D/\sqrt{k} > 0$ and $g''_{\delta,k}(0) = K > 0$, there exists $\varepsilon \in (0, \delta)$ such that

$$0 < g'_{\delta,k}(r) < \frac{\psi D}{\sqrt{k}} \qquad \text{for every } r \in (-\varepsilon, 0).$$

Consequently, the chosen diagonal solution continues to satisfy $r(t) = R - t$ for $R < t < R + \varepsilon$. For these times, $|\operatorname{supp}(\nabla L_{\delta,k}(\theta(t)))| = k$ and

$$\begin{aligned} \|\nabla L_{\delta,k}(\theta(t))\|_1 &= k\, g'_{\delta,k}(r(t)) \\ &< \psi D\sqrt{k} \\ &\leq \psi\sqrt{k}(\|\theta(t)\|_2 + D). \end{aligned}$$

Thus the trajectory lies in the minority aware zone for every $t \in (R, R + \varepsilon)$, which gives $t_{\text{entry}} \leq R$. Together with $t_{\text{entry}} \geq R$, this proves $t_{\text{entry}} = R$.

Now, we introduce the suboptimality gap.

$$\begin{aligned} L_{\delta,k}(\theta_{\text{init}}) - L_{\delta,k}(0) &= k\,(g_{\delta,k}(R) - g_{\delta,k}(0)) \\ &= k\int_0^R g'_{\delta,k}(s)ds \\ &= k\Big(\frac{KR^2}{2} + \frac{\psi DR}{\sqrt{k}}\Big) \leq \frac{3\psi DR\sqrt{k}}{2} \end{aligned}$$

since $\|\theta_{\text{init}}\|_2 = R\sqrt{k} \leq \frac{\psi D}{K}$. Therefore

$$t_{\text{entry}} \geq T_0 = R \geq \frac{2}{3\sqrt{k}} \frac{L_{\delta,k}(\theta_{\text{init}}) - L_{\delta,k}(0)}{\psi D}.$$

By construction, we also know that $L_{\delta,k}(0) - \min L_{\delta,k} \leq kK\delta^2 + 2\delta\sqrt{k}\psi D$. We deduce that

$$t_{\text{entry}} \geq \frac{2}{3\sqrt{k}} \frac{L_{\delta,k}(\theta_{\text{init}}) - \min L_{\delta,k}}{\psi D} - \frac{kK\delta^2 + 2\delta\sqrt{k}\psi D}{\sqrt{k}\psi D}.$$

After noting that

$$\bar{M}_{\mathrm{SD}} = \frac{1}{t_{\mathrm{entry}}} \int_0^{t_{\mathrm{entry}}} \sqrt{|\mathrm{supp}(\nabla L_{\delta,k}(\theta(t)))|}\, dt$$
$$= \frac{1}{R} \int_0^R \sqrt{k}\, dt = \sqrt{k}$$

from Equation 45, we know

$$t_{\mathrm{entry}} \geq \frac{2}{3\bar{M}_{\mathrm{SD}}} \frac{L_{\delta,k}(\theta_{\mathrm{init}}) - \min L_{\delta,k}}{\psi D} - \frac{K\bar{M}_{\mathrm{SD}}\delta^2 + 2\psi\delta D}{\psi D}.$$

□

B.3.3 Lower bound for Spectral Descent

We consider a very similar diagonalizing process for the Spectral Descent lower bound. Let $I_k \in \mathbb{R}^{a\times b}$ be the partial identity matrix with $1 \leq k \leq \min(a,b)$. Note that $\|I_k\|_F = \sqrt{k}$. We define the following objective function, for $k \geq 1$:

$$\forall \Theta \in \mathbb{R}^{a\times b}, \quad L_{\delta,k}(\Theta) := \sum_{i=1}^{k} g_{\delta,k}(\Theta_{ii}). \tag{52}$$

Theorem B.11. *Consider $\Theta_{\mathrm{init}} = RI_k$ with $R > 0$ and $\|\Theta_{\mathrm{init}}\|_F \leq \frac{\psi D}{K}$ and that we follow the Spectral Descent dynamics of $L_{\delta,k}$ defined in Equation* (52)*. Then there exists an absolutely continuous solution of the spectral-descent inclusion for which the entry time is lower bounded as*

$$t_{\mathrm{entry}} \geq \frac{2}{3m_{\|\cdot\|_{sp}}} \frac{L_{\delta,k}(\Theta_{\mathrm{init}}) - \min L_{\delta,k}}{\psi D} - \frac{Km_{\|\cdot\|_{sp}}\delta^2}{\psi D} - 2\delta.$$

Here $m_{\|\cdot\|_{sp}} = m_{\|\cdot\|_{sp}}(\Theta(t)) = \sqrt{k}$ ($0 \leq t < t_{\mathrm{entry}}$).

Proof. First, by symmetry, there exists a trajectory that is restricted to $\Theta(t) = r(t)I_k$ with $r : \mathbb{R}_+ \to \mathbb{R}$. For $t \geq 0$, the matrix gradient is :

$$\nabla L_{\delta,k}(r(t)I_k) = g'_{\delta,k}(r(t))I_k.$$

Similar to the Sign Descent case, we note T_0 the travel time for $r(t)$ from R to 0. Whenever $g'_{\delta,k}(r(t)) > 0$, we have

$$I_k \in \underset{\|V\|_{\mathrm{sp}}\leq 1}{\arg\max}\, \mathrm{tr}((g'_{\delta,k}(r(t))I_k)^\top V) = \rho(\mathrm{grad}_{\|\cdot\|_{\mathrm{sp}}} L_{\delta,k}(\Theta(t))).$$

We select the velocity $-I_k$ until $r(t)$ reaches the first zero of $g'_{\delta,k}$, and keep the trajectory constant thereafter. This zero lies in $[-\delta, 0)$, and zero velocity is admissible there. Hence, we know for $t \in [0, T_0)$,

$$r(t) = R - t.$$

Therefore, for the entry time we know for $0 \leq t \leq R$

$$\|\nabla L_{\delta,k}(\Theta(t))\|_{nuc} = k|g'_{\delta,k}(r(t))| = |Kkr(t) + \psi D\sqrt{k}|$$
$$\psi\sqrt{\mathrm{rank}(\nabla L_{\delta,k}(\Theta(t)))}(\|\Theta(t)\|_F + D) = \psi\sqrt{k}(\sqrt{k}|r(t)| + D).$$

Therefore $t_{\mathrm{entry}} \geq T_0 = R$ as in the Sign Descent case, from Table 1.

We now show that $t_{\mathrm{entry}} = R$. Since $g'_{\delta,k}(0) = \psi D/\sqrt{k} > 0$ and $g''_{\delta,k}(0) = K > 0$, there exists $\varepsilon \in (0, \delta)$ such that

$$0 < g'_{\delta,k}(r) < \frac{\psi D}{\sqrt{k}} \qquad \text{for every } r \in (-\varepsilon, 0).$$

For $R < t < R + \varepsilon$, the selected trajectory therefore still satisfies $r(t) = R - t$, its gradient has rank k, and

$$\begin{aligned}\|\nabla L_{\delta,k}(\Theta(t))\|_{\text{nuc}} &= k\, g'_{\delta,k}(r(t)) \\ &< \psi D\sqrt{k} \\ &\le \psi\sqrt{k}(\|\Theta(t)\|_F + D).\end{aligned}$$

Thus the trajectory lies in the minority aware zone for every $t \in (R, R+\varepsilon)$, giving $t_{\text{entry}} \le R$. Together with $t_{\text{entry}} \ge R$, this proves $t_{\text{entry}} = R$. Moreover, $m_{\|\cdot\|_{sp}}(\Theta(t)) = \sqrt{k}$ for $0 \le t < R$, and its time average satisfies

$$\begin{aligned}\frac{1}{t_{\text{entry}}}\int_0^{t_{\text{entry}}} m_{\|\cdot\|_{sp}}(\Theta(t))\,dt &= \frac{1}{R}\int_0^R \sqrt{\text{rank}(\nabla L_{\delta,k}(\Theta(t)))}\,dt \\ &= \sqrt{k}.\end{aligned}$$

We introduce the suboptimality gap and obtain a similar conclusion as for the Sign Descent case, using that $m_{\|\cdot\|_{sp}} = \sqrt{k}$ from Equation 47:

$$t_{\text{entry}} \ge \frac{2}{3m_{\|\cdot\|_{sp}}}\frac{L_{\delta,k}(\Theta_{\text{init}}) - \min L_{\delta,k}}{\psi D} - \frac{K m_{\|\cdot\|_{sp}}\delta^2 + 2\psi\delta D}{\psi D}.$$

□

C Details for numerical experiments

First, we present the material and computational time required to run our experiments in Table 5.

Experiment	Dataset	GPU	Computational Time
Linear Regression	Superconductor	None	<10min
Linear Regression	YearPredictionMSD	None	<10min
Classification	Covertype	Nvidia Spark	12H
Classification	Cifar10	Nvidia Spark	24H
LLM	WikiText-103	Nvidia Spark	48H

Table 5: Computational resources and time across different models.

C.1 Linear regression

To generate a family of problems with consistent properties across varying dimensions, we apply Principal Component Analysis (PCA). We will create a series of subproblems with only $\alpha\%$ of the dimension of the original problem.

The minority loss function is defined on a subset of $k \in \{3, \dots, 20\}$ samples that exhibit the largest second moments in the feature matrix X. We then numerically compute the entry times for both optimizers alongside the mass of the minority group, ψ. By deriving empirical scaling laws that relate these quantities to the dimension d, we can directly compare the experimental results against our theoretical predictions.

The experimental protocol is defined as follows.

- Choose a diameter $D \in \{0.1, 0.3, 1\}$.
- Choose a shared learning rate $\eta \in \left\{\frac{1}{100\lambda_{\max}}, \frac{1}{300\lambda_{\max}}, \frac{1}{1000\lambda_{\max}}\right\}$ for both GD and SD. Here $\lambda_{\max}$ is the largest eigenvalue of the global loss Hessian.
- For each configuration (α, k), execute both GD and SD.
- Estimate ψ as the maximum eigenvalue of the Hessian of the minority $\psi \leftarrow \lambda_{\max}(\nabla^2(L_0))$.

• Compute $t_{\text{entry}}(\text{GD})$ and $t_{\text{entry}}(\text{SD})$ using their formal definitions.

• Compute $\frac{t_{\text{entry}}(\text{GD})}{t_{\text{entry}}(\text{SD})}$ for each configuration (α, k) and make a log log plot to fit the model $\frac{t_{\text{entry}}(\text{GD})}{t_{\text{entry}}(\text{SD})} \propto d^{c_1}\psi^{c_2}$.

We will apply the protocol to the two following dataset.

Superconductor. The Superconductor dataset comprises $21,000$ samples and 80 features, with the material's critical temperature serving as the target variable.

YearPredictionMSD. The YearPredictionMSD dataset contains $513,000$ samples and 90 features, with the release year serving as the prediction target.

The results are plotted in Figure 3 and the regression results can be found in Tables 6 and 7.

The estimated log log regression coefficients are presented in Tables 6 and 7.

	$\eta = \frac{1}{100\lambda_{\max}}$	$\eta = \frac{1}{300\lambda_{\max}}$	$\eta = \frac{1}{1000\lambda_{\max}}$
$D = 0.1$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.51}}{\psi^{0.34}}$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.50}}{\psi^{0.35}}$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.50}}{\psi^{0.35}}$
$D = 0.3$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.52}}{\psi^{0.39}}$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.51}}{\psi^{0.38}}$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.51}}{\psi^{0.38}}$
$D = 1.0$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.49}}{\psi^{0.29}}$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.50}}{\psi^{0.26}}$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.50}}{\psi^{0.24}}$

Table 6: Estimated log log regression coefficients for different choices of (D, η) on the YearPredictionMSD dataset.

	$\eta = \frac{1}{100\lambda_{\max}}$	$\eta = \frac{1}{300\lambda_{\max}}$	$\eta = \frac{1}{1000\lambda_{\max}}$
$D = 0.1$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.71}}{\psi^{0.80}}$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.71}}{\psi^{0.80}}$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.64}}{\psi^{0.68}}$
$D = 0.3$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.65}}{\psi^{0.91}}$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.65}}{\psi^{0.91}}$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.64}}{\psi^{0.88}}$
$D = 1.0$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.55}}{\psi^{0.69}}$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.55}}{\psi^{0.68}}$	$\frac{\mathcal{T}_{GD}}{\mathcal{T}_{SD}} \propto \frac{d^{0.56}}{\psi^{0.68}}$

Table 7: Estimated log log regression coefficients for different choices of (D, η) on the Superconductivity Dataset.

C.2 CLASSIFICATION

To closely align with our theoretical framework, we evaluate optimizers without momentum for the Covertype and CIFAR-10 classification tasks. Across all experiments, the batch size is fixed at 128. Full hyperparameter configurations are detailed in Table 8. We repeat all experiments across 3 fixed

Optimizers	Parameters
AdamW	$lr = 10^{-3}, \beta = (0.0, 0.0)$, weight_decay $= 10^{-3}$
SGD	$lr = 0.05$, momentum $= 0.0$, weight_decay $= 10^{-3}$
Muon	$lr = 0.02$, momentum $= 0.0$, adamw_lr $= 10^{-3}$, adamw_wd $= 10^{-2}$

Table 8: Hyperparameters of optimizers for classification

random seeds and report the average accuracy at each iteration. The results for CIFAR-10 are shown in Figure 4. Those from Covertype are shown in Table 9.

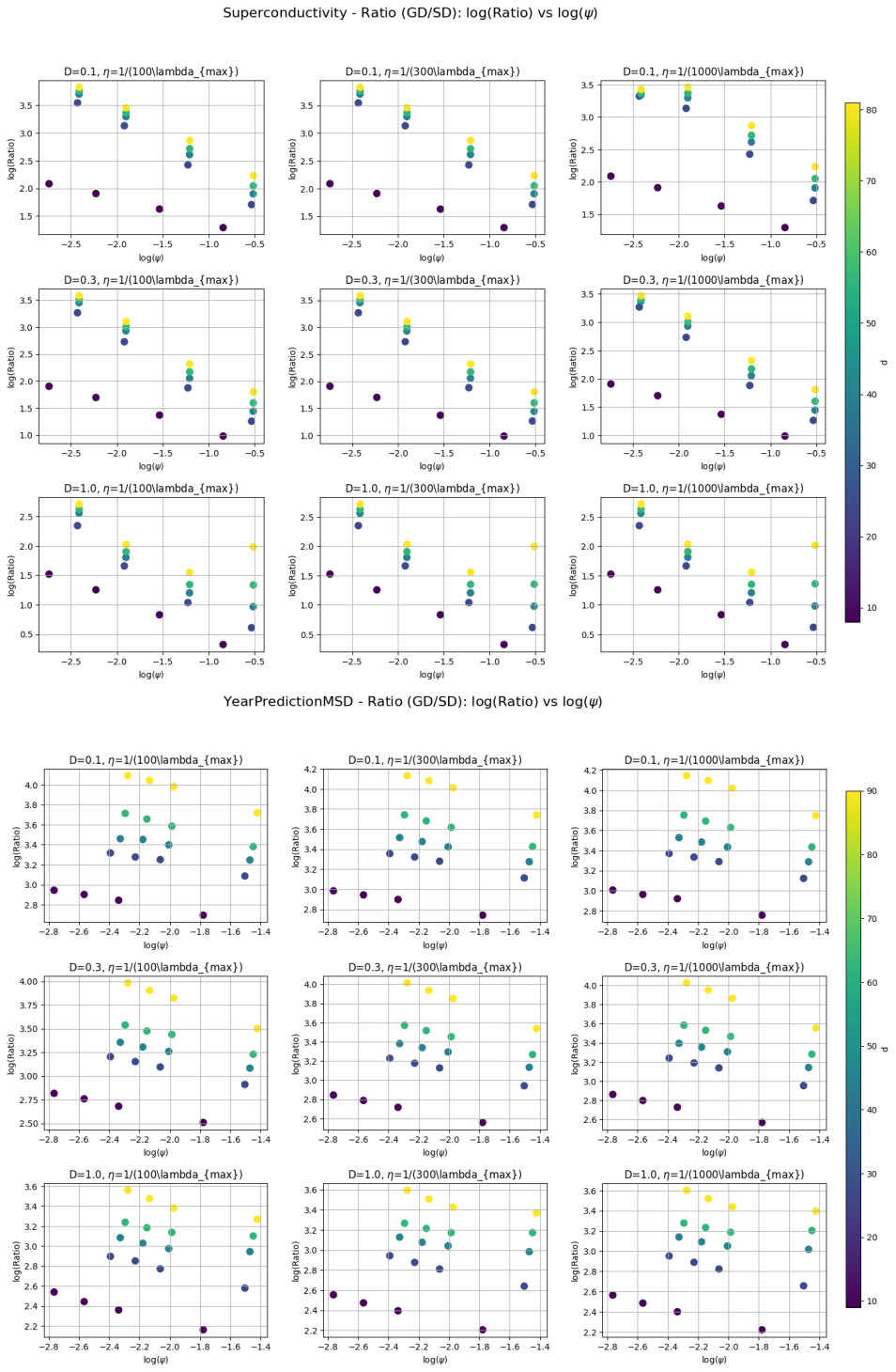


Figure 3: Ratio of Entry times for Gradient Descent and Sign Descent in the linear regression setting. The top panel corresponds to the Superconductor dataset, and the bottom panel corresponds to the YearPredictionMSD dataset.

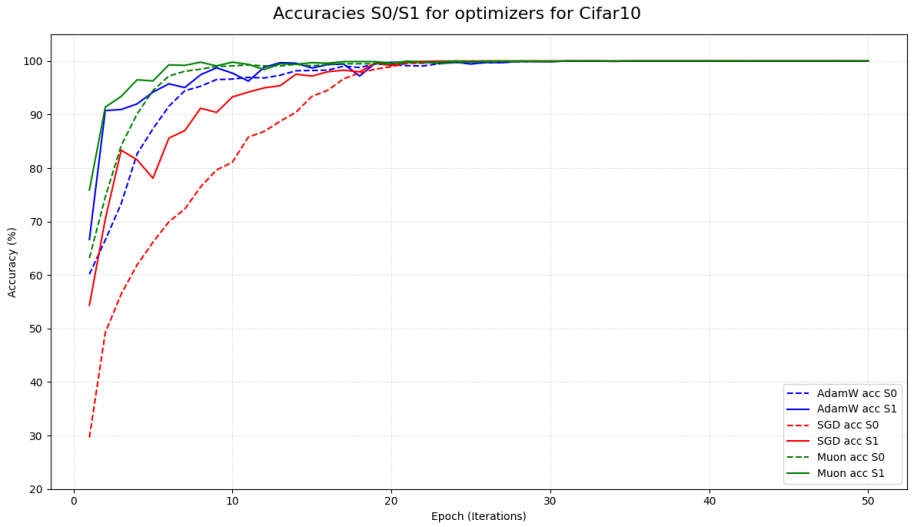


Figure 4: Evolution of S_0/S_1 accuracy during training phase for Cifar10. Here S_0 denotes the minority group and S_1 the majority group.

Across both the tabular and image classification experiments, we observe a consistent trend: although AdamW, SGD, and Muon achieve comparable final test performance on the majority population, their efficiency on the hidden minority diverges significantly. Notably, normalized gradient methods—such as AdamW, and particularly Muon—demonstrate substantially stronger learning capabilities on the minority group.

Metric	AdamW	SGD
Average S_0 accuracy	63.93%	51.40%
Average S_1 accuracy	90.99%	88.25%

Table 9: Evaluation of minority/majority accuracy, averaged during training, on the Covertype dataset with AdamW and SGD.

C.3 Language Modelling

We train a SmallGPT2 model on WikiText-103 from scratch with the BPE Tokenizer. The details of the experiment can be found in Table 10. We consider the normalized losses

$$\overline{L}_1(\theta) = \frac{\sum_{i=1}^N L_i(\theta) \cdot \mathbb{I}_{\text{major}}(y_i)}{\sum_{i=1}^N \mathbb{I}_{\text{major}}(y_i)}$$
$$\overline{L}_0(\theta) = \frac{\sum_{i=1}^N L_i \cdot \mathbb{I}_{\text{minor}}(y_i)}{\sum_{i=1}^N \mathbb{I}_{\text{minor}}(y_i)},$$

where $\mathbb{I}_{\text{major}}(y_i)$ indicates that the token y_i belongs to the majority group, and $\mathbb{I}_{\text{minor}}(y_i)$ indicates that the token y_i belongs to the minority group.

To account for the distinct minimum values of these objectives, we normalize the loss trajectories using a fully pre-trained SmallGPT-2 model with parameters $\theta_\star$. By evaluating the normalized majority loss $\overline{L}_1 - \overline{L}_1(\theta_\star)$ and minority loss $\overline{L}_0 - \overline{L}_0(\theta_\star)$, we plot the shifted objective trajectories, to compare the optimizer dynamics.

The results are presented in Figure 5. The optimization of the majority loss L_1 aligns with the established hierarchy of convergence rates, where Muon outpaces AdamW, and AdamW outperforms SGD. However, the optimization dynamics of SGD reveal a fundamental divergence: it neglects to minimize the minority loss L_0, prioritizing majority tokens at the direct expense of rare tokens.

Interestingly, by step 2000, the Muon-optimized parameters actually surpass the pre-trained baseline's performance on rare tokens, albeit at the cost of a degraded majority loss. This behavior suggests a potential multi-stage learning process, prompting the hypothesis that a later training phase may reverse this trade-off to recover majority performance.

	Settings
Model architecture	$L = H = 12, d_{\text{model}} = 768, \sigma_{\text{init}} = 0.02$
Common Training	Sequence Length $= 1024$, Accumulation steps = 16, Batch Size $= 512$
AdamW	$lr = 5 \times 10^{-4}, \beta = (0.9, 0.999), \epsilon = 10^{-8}$, weight_decay $= 10^{-2}$
SGD	$lr = 0.02$, momentum $= 0.9$
Muon	$lr = 0.02$, momentum $= 0.95$

Table 10: Hyperparameters of optimizers for LLM

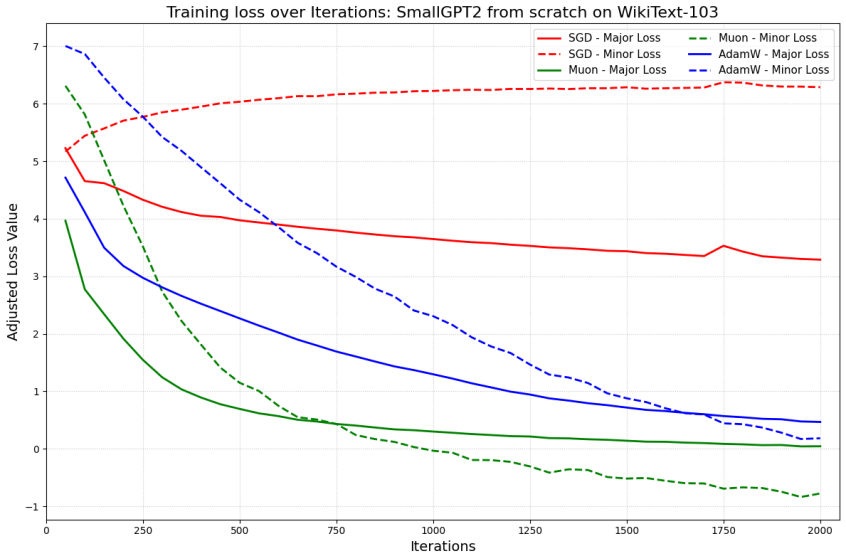


Figure 5: Evolution of training loss of SmallGPT2 from scratch on WikiText-103 over 2000 iterations. We plot the relative loss $\overline{L}_1 - \overline{L}_1(\theta_\star)$ and $\overline{L}_0 - \overline{L}_0(\theta_\star)$ for comparison.

D TABLE OF COMPARISON OF κ

In Tables 11, 12 and 13, we now examine the values of κ reported in the literature across three task families: image classification, autoregressive language modelling and tabular learning. Among the configurations reported below, peak learning rates range from 10^{-3} to 2×10^{-2} for image classification, from 10^{-4} to 5×10^{-2} for tabular learning, and from 1.5×10^{-4} to 10^{-1} for language modelling. These ranges yield $\kappa \approx 20$, $\kappa \approx 500$, and $\kappa \approx 700$, respectively.

From Section 1, recall that $\kappa = \eta_{\max}/\eta_{\min}$ denotes the ratio between the largest and smallest learning-rate magnitudes before scheduling. Applying a shared multiplicative schedule, such as cosine annealing, preserves these learning-rate ratios whenever the scheduling factor is positive.

E FURTHER GEOMETRIC CONSIDERATIONS UNDER STRONG CONVEXITY

In this section, we examine how strong convexity of the loss function impacts our theoretical results.

Lemma E.1. *Suppose that L is μ-strongly-convex, $L_0 \in \mathcal{K}_{\min}$ and $\hat{\theta} = \arg\min L$. Then*

$$\forall \theta \in \mathbb{R}^d, \quad |L_0(\theta) - L_0(\hat{\theta})| \leq \frac{\psi}{2\mu^2}\|\nabla L(\theta)\|_2^2 + \frac{\psi}{\mu}\|\nabla L(\theta)\|_2(\|\hat{\theta}\|_2 + D).$$

Citation	Model (Size)	Dataset	Optimizer	Peak LR	Schedule
Krizhevsky et al. (2012)	AlexNet (60M)	ImageNet	SGD	0.01	Step
Loshchilov & Hutter (2019)	AlexNet (60M)	ImageNet	AdamW	1e-3	Cosine
Dosovitskiy et al. (2021)	ViT-Base (86M)	ImageNet	AdamW	1e-3	Cosine
Southworth et al. (2026)	ViT-Base (86M)	ImageNet	Muon	0.02	Cosine

Table 11: Image Classification, $\kappa \approx 20$.

Citation	Model (Size)	Dataset	Optimizer	Peak LR	Schedule
Touvron et al. (2023)	LLaMA (7B)	1T tokens	AdamW	3e-4	Cosine
Touvron et al. (2023)	LLaMA (65B)	1.4T tokens	AdamW	1.5e-4	Cosine
Karpathy (2023)	NanoGPT (124M)	OpenWebText	AdamW	6e-4	Cosine
Hu et al. (2026)	NanoGPT (124M)	OpenWebText	SGD (Momentum)	$\sim$0.1	Cosine
Jordan et al. (2024)	NanoGPT (124M)	OpenWebText	Muon	0.02	Varies
Hu et al. (2026)	Modded-NanoGPT	Fineweb	Muon	0.025	Trapezoidal

Table 12: Autoregressive Language Modeling, $\kappa \approx 700$.

Proof. Fix $\theta \in \mathbb{R}^d$ and consider $g(t) := L_0((1-t)\hat{\theta} + t\theta)$. Since L_0 is C^2, we deduce that

$$\begin{aligned} g(1) - g(0) &= \int_0^1 g'(s)ds \\ |L_0(\theta) - L_0(\hat{\theta})| &= \Big| \int_0^1 \langle \nabla L_0((1-s)\hat{\theta} + s\theta), \theta - \hat{\theta} \rangle ds \Big| \\ &\leq \int_0^1 \|\nabla L_0((1-s)\hat{\theta} + s\theta)\|_2 \|\theta - \hat{\theta}\|_2 ds \\ &\leq \psi \int_0^1 \Big(\|((1-s)\hat{\theta} + s\theta)\|_2 + D \Big) \|\theta - \hat{\theta}\|_2 ds \\ &\leq \psi \int_0^1 \Big(s\|\hat{\theta} - \theta\|_2 + \|\hat{\theta}\|_2 + D \Big) \|\theta - \hat{\theta}\|_2 ds. \end{aligned}$$

We deduce that

$$\begin{aligned} |L_0(\theta) - L_0(\hat{\theta})| &\leq \psi \|\theta - \hat{\theta}\|_2 \Big(\frac{\|\theta - \hat{\theta}\|_2}{2} + \|\hat{\theta}\|_2 + D \Big) \\ &\leq \frac{\psi}{2\mu^2} \|\nabla L(\theta)\|_2^2 + \frac{\psi}{\mu} \|\nabla L(\theta)\|_2 (\|\hat{\theta}\|_2 + D) \end{aligned}$$

since $\|\nabla L(\theta)\|_2 \geq \mu \|\theta - \hat{\theta}\|_2$ due to the strong convexity of L. □

Building on the preceding lemma, strong convexity allows us to directly bound the minority loss gap, $|L_0(\theta) - L_0(\hat{\theta})| < \epsilon$ in terms of ψ and the gradient norm $\|\nabla L\|_2$.

Lemma E.2. *Under Assumptions 2.1 and 3.1, suppose that L is μ-strongly-convex. Then for any fixed L and θ_{init} such that $\nabla L(\theta_{\text{init}}) \neq 0$,*

$$\begin{aligned} \forall \psi, D > 0, \quad & t_{\text{entry}}^{\text{SD}} \leq \sqrt{\frac{2(L(\theta_{\text{init}}) - L_\star)}{\mu}} \\ \forall D > 0, \quad & t_{\text{entry}}^{\text{GD}} \xrightarrow[\psi \to 0]{} +\infty. \end{aligned}$$

Proof. We note that by strong-convexity :

$$\forall \theta, \quad \frac{1}{2} \|\nabla L(\theta)\|_2^2 \geq \mu \Big(L(\theta) - L(\hat{\theta}) \Big).$$

Citation	Model / Architecture	Dataset	Optimizer	Peak LR	Schedule
Gorishniy et al. (2021)	ResNet (Tabular)	Adult, Higgs, CovType	AdamW	1e-3	Constant / Step
Arik & Pfister (2021)	TabNet	Forest Cover, KDD	Adam	1e-3	Step Decay
Arik & Pfister (2021)	TabNet	Rossmann, Adult	Adam	2e-2	Step Decay
Wang et al. (2024)	TabKANet	General Tabular Reg.	SGD	1e-4	Step
Qu et al. (2026)	TabICLv2	Synthetic	Muon	1e-4 - 2e-5	Cosine / Trap

Table 13: Tabular Data, $\kappa \approx 1000$.

For Sign Descent dynamism, by the chain rule, we obtain for $t \geq 0$

$$\begin{aligned}
&\frac{d}{dt}\Big(L(\theta(t)) - L_\star\Big) = -\|\nabla L(\theta(t))\|_1 \leq -\|\nabla L(\theta(t))\|_2 \\
\implies &\frac{d}{dt}\Big(\sqrt{L(\theta(t)) - L_\star}\Big) = \frac{-\|\nabla L(\theta(t))\|_1}{2\sqrt{L(\theta(t)) - L_\star}} \leq -\frac{1}{2}\sqrt{2\mu} \\
\implies &-\sqrt{L(\theta_{\text{init}}) - L_\star} \leq -\sqrt{\frac{\mu}{2}} t^{\text{SD}}_{\text{entry}} \\
\implies &t^{\text{SD}}_{\text{entry}} \leq \sqrt{\frac{2(L(\theta_{\text{init}}) - L_\star)}{\mu}}.
\end{aligned}$$

For Gradient Descent, we start by defining

$$\mathcal{S}(\theta_{\text{init}}) := \{\theta \in \mathbb{R}^d, L(\theta) \leq L(\theta_{\text{init}})\}.$$

By the strong convexity of L, we know that for all θ

$$L(\theta) - L(\theta_{\text{init}}) \geq \langle \nabla L(\theta_{\text{init}}), \theta - \theta_{\text{init}} \rangle + \frac{\mu}{2}\|\theta - \theta_{\text{init}}\|_2^2.$$

Therefore

$$L(\theta) \leq L(\theta_{\text{init}}) \implies \|\nabla L(\theta_{\text{init}})\|_2 \|\theta - \theta_{\text{init}}\|_2 \geq \frac{\mu}{2}\|\theta - \theta_{\text{init}}\|_2^2.$$

If $\theta \neq \theta_{\text{init}}$, we have $\|\nabla L(\theta_{\text{init}})\|_2 \geq \frac{\mu}{2}\|\theta - \theta_{\text{init}}\|_2$. Therefore

$$\theta \in \mathcal{S}(\theta_{\text{init}}) \implies \|\theta - \theta_{\text{init}}\|_2 \leq \frac{2}{\mu}\|\nabla L(\theta_{\text{init}})\|_2.$$

Then $\mathcal{S}(\theta_{\text{init}})$ is compact, we note

$$\lambda_{\max} := \max_{\theta \in \mathcal{S}(\theta_{\text{init}})} \sigma_{\max}(\nabla^2 L(\theta)).$$

By the chain rule, we have for $t \geq 0$

$$\begin{aligned}
\frac{d}{dt}\|\nabla L(\theta(t))\|_2^2 &= -2\langle \nabla L(\theta(t)), \nabla^2 L(\theta(t)) \nabla L(\theta(t)) \rangle \\
&\geq -2\lambda_{\max}\|\nabla L(\theta(t))\|_2^2 \\
\implies \|\nabla L(\theta(t))\|_2^2 &\geq \|\nabla L(\theta_{\text{init}})\|_2^2 e^{-2\lambda_{\max} t}.
\end{aligned}$$

Since, from Table 1,

$$\|\nabla L(\theta(t^{\text{GD}}_{\text{entry}}))\|_2 \leq \psi(\|\theta(t^{\text{GD}}_{\text{entry}})\|_2 + D) \leq \psi(\|\theta_{\text{init}}\|_2 + \frac{2}{\mu}\|\nabla L(\theta_{\text{init}})\|_2 + D),$$

we obtain

$$\begin{aligned}
t^{\text{GD}}_{\text{entry}} &\geq \frac{1}{\lambda_{\max}} \log\left(\frac{\|\nabla L(\theta_{\text{init}})\|_2}{\|\nabla L(\theta(t^{\text{GD}}_{\text{entry}}))\|_2}\right) \\
&\geq \frac{1}{\lambda_{\max}} \log\left(\frac{\|\nabla L(\theta_{\text{init}})\|_2}{\psi(\|\theta_{\text{init}}\|_2 + \frac{2}{\mu}\|\nabla L(\theta_{\text{init}})\|_2 + D)}\right) \xrightarrow[\psi \to 0]{} +\infty.
\end{aligned}$$

□

The previous lemma demonstrates that for strongly convex objective functions, normalized methods such as Sign Descent exhibit fundamentally different entry time dynamics compared to standard Gradient Descent.